Genetic analysis of inherited X-linked retinitis pigmentosa:
development of a transcriptional map of the RP2 region

Manimekalai Nayudu

A thesis submitted for the degree of Doctor of Philosophy

University College London
Department of Molecular Genetics
Institute of Ophthalmology
London

1999
DECLARATION

I declare that
a) This thesis has been composed by myself
b) The work in this thesis is my own

Manimekalai Nayudu BSc.Hons., MSc.
Abstract

X-linked retinitis pigmentosa is genetically heterogeneous with at least two loci on proximal Xp, known as RP2 and RP3. The RP2 locus is not well defined due to a lack of deletion patients; at the onset of this study the RP2 locus had been mapped to a 13cM region between markers DXS7 and DXS255 (Xp11.3-Xp11.23). The primary objective of this thesis was to further progress positional cloning and identification of the RP2 disease-causative gene.

Initially, thirteen new XIRP families were analysed using twenty eight microsatellite markers spanning Xp22.13 to Xp11.22 with the aim of refining the genetic interval within which the RP2 gene lies. From this haplotype analysis, three families were found to be linked to the RP3 locus, one to the Choroideremia locus, seven could not be unambiguously assigned and two families were found to be linked to the RP2 locus with crossovers observed below the marker MAOA, further localising the RP2 gene proximal to MAOA and confirming previous linkage data. The TIMP-1 gene was considered to be a candidate for RP2 since mutations in the TIMP-3 gene were found to cause Sorsby's fundus dystrophy, a macular degeneration. Direct sequencing of the six exons and exon/intron boundaries of TIMP-1 in affected male RP2 patients did not reveal mutations associated with the disease phenotype, hence TIMP-1 is unlikely to be the RP2 disease-causative gene.

Heterogeneity analysis suggests that the RP2 gene probably lies 2cM distal to DXS426. Well characterised YACs around DXS426 were utilised in the development of a transcriptional map of this region. The selection for enriched cDNAs using the magnetic bead capture technique was applied to four overlapping YACs which cover a region of 500kb. Two adult retinal cDNA libraries were used in the selection procedure. cDNA/YAC DNA hybrids were captured using streptavidin coated magnetic beads and two rounds of selection were performed. The A1220 and 33CA11 YACs contain the TIMP-1 and properdin (PFC) genes respectively, and these genes were also present in the starting retinal cDNA libraries with the TIMP-1 gene shown to be especially rare. The YACs chosen from the region of interest therefore contain these internal positive control genes. A rhodopsin YAC was included for use as an external positive control for the technique.

YAC specific sub-libraries created for enriched cDNAs were screened for contaminating sequences such as ribosomal genes, yeast DNA and β-actin cDNAs and showed depletion for such unwanted transcripts. The successful establishment of the selection procedure for the efficient enrichment of low abundance retinal
transcripts was demonstrated by a 1,000 fold enrichment for the rare positive control gene TIMP-1. A selection of enriched cDNA clones (40-50 per YAC) were sequenced and compared to each other in order to identify overlapping clones, and to database entries using the Gapped BLAST search program. Twelve interesting clones were used in mapping back studies to the region investigated.
Acknowledgements

I would like to express my gratitude to the Wellcome Trust for provision of the Wellcome Prize studentship and support for this research. I am also thankful to Prof. Shomi Bhattacharya for his advice and guidance and Dr. Alison Hardcastle for her suggestions and help.

My special thanks to Dr. Emma Tarttelin for her friendship, support and constructive advice and comments during the reading of this thesis.

I would especially like to convey my appreciation and thanks to my friends Dr. Ho Meng Fatt (Mac), Mrs. Ho Ai Peng, Dr. Shahla Abadeh, Dr. Lai Poh San, Dr. B. Ninan, Dr. Elisha Orr, Miss Naheed Kanuga and last but not least, Mr. Neil Ebeneezer for their care, concern, encouragement, and many happy memories.

I am indebted to my parents, Naina and Amma for their trust, faith, encouragement and continual support throughout my studies and especially for their unfailing love. I extend my sincere appreciation and thanks to both my sister Miss Carnegie Naidu and my brother in Christ Mr. Vivek Verma for their constant encouragement, inspiration and prayers. Lastly, my heartfelt gratitude to my husband, Mr. William Alwyn Orr, for his unwavering support, prayers and his steadfast love.
**Habakkuk 3:17-19**

"Though the fig tree does not bud and there are no grapes on the vines, though the olive crop fails and the fields produce no food, though there are no sheep in the pen and no cattle in the stalls, yet I will rejoice in the Lord, I will be joyful in God my Saviour. The Sovereign Lord is my strength; he makes my feet like the feet of a deer, he enables me to go on the heights."

**2 Samuel 22:2,3**

"The Lord is my rock, my fortress and my deliverer...... He is my stronghold, my refuge and my Saviour."

**Philippians 4:13**

"I can do all things through Christ who gives me strength"

**Psalm 28:7**

"The Lord is my strength and my shield; my heart trusts in him and I am helped. My heart leaps for joy and I will give thanks to him in song."

**Dedication**

To my loving parents

Naina and Amma

and my husband Alwyn
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>1</td>
</tr>
<tr>
<td>Declaration</td>
<td>2</td>
</tr>
<tr>
<td>Abstract</td>
<td>3</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>5</td>
</tr>
<tr>
<td>Dedication</td>
<td>6</td>
</tr>
<tr>
<td>Contents</td>
<td>7</td>
</tr>
<tr>
<td>List of Figures</td>
<td>19</td>
</tr>
<tr>
<td>List of Tables</td>
<td>23</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>25</td>
</tr>
</tbody>
</table>

### 1. INTRODUCTION

#### 1.1 Principles and strategies in identifying disease genes

- 1.1.1 Functional cloning
- 1.1.2 Positional cloning
- 1.1.3 Candidate gene approaches
  - 1.1.3.1 Position-independent candidate gene approaches
  - 1.1.3.2 Positional candidate gene approaches

#### 1.2 The Human Genome Project

- 1.2.1 Human genetic maps
- 1.2.2 Physical maps of the human genome
- 1.2.3 Transcriptional maps of the human genome
- 1.2.4 Genome maps of model organisms
- 1.2.5 Sequencing the human genome
- 1.2.6 Data storage and access in the Human Genome Project

#### 1.3 Genetic Mapping

- 1.3.1 Meiotic recombination
- 1.3.2 Principles of linkage analysis
- 1.3.3 Calculation of lod scores
- 1.3.4 Genetic map distance
- 1.3.5 Genetic markers
- 1.3.6 Cross-over (haplotype) analysis
1.4 Physical Mapping

1.4.1 Low resolution physical mapping
1.4.1.1 Somatic cell hybrids
1.4.1.2 In-situ hybridisation

1.4.2 High resolution physical mapping
1.4.2.1 Pulse field gel electrophoresis (PFGE)
1.4.2.2 Assembly of clone contigs

1.5 Identification of gene sequences from a defined genomic interval

1.5.1 Hybridisation to RNA/cDNA
1.5.2 Zoo-blot hybridisation
1.5.3 CpG island identification
1.5.4 Direct screening of cDNA libraries
1.5.5 Exon trapping/exon amplification technique
1.5.6 cDNA selection/capture
1.5.7 Subtractive selection cloning and differential hybridisation methods
1.5.8 Computer-based DNA sequence analysis
1.5.9 Summary, biases and limitations of gene identification methods
1.5.10 Similarity searches using sequence databases

1.6 Inherited retinal degeneration

1.6.1 Structure of the eye
1.6.1.1 The globe
1.6.1.2 The neural retina
1.6.2 The photoreceptors
1.6.3 The rod phototransduction pathway
1.6.4 Calcium feedback and the recovery of the dark state
1.6.5 Apoptosis - a cause for photoreceptor cell death
1.6.5.1 Pathogenesis
1.6.5.2 Animal models of retinal degeneration
1.6.6 Proteins associated with various retinal degenerative diseases
1.6.6.1 Rhodopsin
1.6.6.2 Rod cGMP phosphodiesterase α and β subunits (PDEA and PDEB)
1.6.6.3 cGMP-gated channel
1.6.6.4 Peripherin/RDS
1.6.6.5 Rod outer membrane 1 (ROM-1)
1.6.6.6 Retinal guanylate cyclase-1 (RETGC -1)
1.6.6.7 Guanylate cyclase activating protein 1 (GCAP 1) 88
1.6.6.8 Retinal rod-specific ATP binding cassette transporter (ABCR/ABCR/RIM) 88
1.6.6.9 Cone-rod homeobox (CRX) 89
1.6.6.10 Cellular retinaldehyde-binding protein (CRALBP) 89
1.6.6.11 RPE65 90
1.6.6.12 Bestrophin 91

1.7 The human X chromosome 91
1.7.1 Mapping X-linked diseases 92
1.7.2 Mutations in X-linked diseases 93
1.7.3 The genetic, physical and transcriptional map of the human X chromosome 93
1.7.4 Other aspects of the X chromosome 99
  1.7.4.1 X-inactivation 99
  1.7.4.2 Human and mouse X-syntenic regions 100

1.8 Eye diseases on the X chromosome 102
1.8.1 Norrie disease 102
1.8.2 X-linked congenital stationary night blindness 105
1.8.3 Aland Island Eye Disease (AIED) 107
1.8.4 X-linked progressive cone dystrophy (XLPCD) 108
1.8.5 X-linked optic atrophy 109
1.8.6 Choroideremia 110
1.8.7 X-linked retinitis pigmentosa 112
  1.8.7.1 Genetic heterogeneity in XLRP 114
  1.8.7.2 Phenotype/Genotype correlations in XLRP 116
  1.8.7.3 Recent progress in positional cloning of the RP3 and RP2 genes 117
    1.8.7.3.1 RP3 - Physical mapping studies and isolation of candidate genes 117
    1.8.7.3.2 RP2 - further genetic mapping studies to confirm and refine the RP2 locus and recent isolation of a candidate gene 120

1.9 Aims of this thesis 121
Summary of Introduction

1.1 Principles and strategies for identifying disease genes

1.1.1 The Human Genome Project
   1.1.1.1 Human genetic maps
   1.1.1.2 Physical maps of the human genome
   1.1.1.3 Transcriptional maps of the human genome
   1.1.1.4 Sequencing the human genome
   1.1.1.5 The genetic, physical and transcriptional map of the human X-chromosome

1.1.2 Genetic mapping
   1.1.2.1 Principles of linkage analysis
   1.1.2.2 Calculation of lod scores

1.1.3 Physical mapping

1.1.4 Identification of gene sequences from a defined genomic interval
   1.1.4.1 Cross species sequence homology and selection of CpG islands
   1.1.4.2 Direct YAC screening of cDNA libraries
   1.1.4.3 Exon amplification
   1.1.4.4 cDNA selection
   1.1.4.5 Sequence based methods

1.2 Inherited retinal degeneration

1.2.1 Structure of the retina
1.2.2 Phototransduction
1.2.3 X-linked eye diseases
   1.2.3.1 X-linked retinitis pigmentosa
      1.2.3.1.1 Retinitis pigmentosa 3
      1.2.3.1.2 Retinitis pigmentosa 2
   1.2.3.2 Choroideremia
   1.2.3.3 Other eye diseases on proximal Xp
2. MATERIALS AND METHODS

2.1 Genomic DNA Preparation and Analysis

2.1.1 DNA extraction
  2.1.1.1 Standard protocol
  2.1.1.2 Rapid protocol

2.1.2 Restriction Endonuclease Digestions

2.1.3 Agarose gel electrophoresis

2.2 Blotting Procedures

2.2.1 Southern (Alkali) Blotting

2.2.2 Bidirectional Southern Blotting

2.3 Hybridisation Procedures

2.3.1 Probe preparation and purification of inserts

2.3.2 Radioisotopic oligolabelling of probes

2.3.3 Removal of repetitive sequences from labelled DNA

2.3.4 Estimation of incorporation of radioisotopes into the synthesised probe

2.3.5 Nucleic acid hybridisation to Southern blots

2.3.6 Membrane stripping protocols

2.3.7 Autoradiography

2.4 Manipulation of bacterial and bacteriophage cloned DNA

2.4.1 Vectors and bacterial strains

2.4.1.1 Storage and recovery of bacterial cultures in glycerol

2.4.2 Transformation

2.4.2.1 Preparation of competent cells using calcium chloride

2.4.2.2 Transformation of bacterial strains with clone of interest

2.4.2.3 Growth media and selection of recombinants

2.4.3 Preparation of plasmid DNA

2.4.3.1 Midi-preparation of plasmid DNA

2.4.3.2 Mini-preparation of plasmid DNA

2.4.4 Manipulation of the λZAPII and λgt10 human adult retinal cDNA libraries

2.4.4.1 Preparation of host cells

2.4.4.2 Titration of λZAPII and λgt10 phage stock DNA

2.4.4.3 Amplification and preparation of plate lysate stocks of the bacteriophage λgt10 library

2.4.5 Immobilization of bacteriophage λ plaques onto Hybond-N+
2.5 Yeast Artificial Chromosome (YAC) Methods

2.5.1 Yeast strain and Vector 137
2.5.2 Preparation of YAC DNA in plugs 137
2.5.3 Pulsed Field Gel Electrophoresis (PFGE) 138
   2.5.3.1 Casting and loading the gel 139
   2.5.3.2 Gel electrophoresis 139
   2.5.3.3 Gel staining 139
2.5.4 Isolation of YAC clones from Pulse Field gels 139

2.6 The Polymerase Chain Reaction (PCR)

2.6.1 General considerations 140
2.6.2 Standard parameters for a typical PCR 140
   2.6.2.1 PCR from a bacteriophage plaque 141
   2.6.2.2 PCR from a bacterial colony 141
   2.6.2.3 PCR from YAC plug DNA 141

2.7 Analysis of microsatellites 141

2.7.1 5'-end-labelling of primers 141
2.7.2 Polyacrylamide denaturing gel electrophoresis 142

2.8 Sequencing double stranded DNA 143

2.8.1 Overview of dideoxy chain termination sequencing 143
2.8.2 Dideoxy sequencing from a cloned product 143
   2.8.2.1 DNA Purification and annealing of primer to double stranded template 143
   2.8.2.2 Sequencing using T7 (Pharmacia) kits with α35 incorporation labelling of primer 144
2.8.3 Direct genomic sequencing of PCR products 144
   2.8.3.1 DNA purification 144
   2.8.3.2 γ32 P 5'-end labelling of primer 144
   2.8.3.3 Sequencing using T7 (Pharmacia) kits 145
2.8.4 Database usage and computational analysis of sequence information 145

2.9 Reagents and buffers used in this research 146

2.9.1 Stock reagents in general use 146
2.9.2 Preparation of DNA 147
2.9.3 Gel electrophoresis 148
2.9.4 Southern blot reagents 148
2.9.5 Transformation of cloned DNA using competent E.coli host cells 149
3. CHARACTERISATION OF XLRP PEDIGREES USING HAPLOTYPE ANALYSIS

3.1 Introduction

3.1.1 Clinical and pathological characteristics of XLRP
3.1.2 Clinical heterogeneity of XLRP
3.1.3 Obligate carriers and variable expressivity
3.1.4 Definition and the use of haplotype analysis in XLRP pedigrees

3.2 Methodology

3.2.1 Subjects and Samples
3.2.2 Polymorphic markers on Xp used in this study
3.2.3 Detection of Microsatellite Polymorphisms

3.3 Results of clinical and haplotype analysis

3.3.1 Family RP 85
3.3.1.1 Clinical diagnosis of affected males and carrier females
3.3.1.2 Haplotype analysis of XLRP 85

3.3.2 Family RP 87
3.3.2.1 Clinical diagnosis of affected males and carrier females
3.3.2.2 Haplotype analysis of XLRP 87

3.3.3 Family RP 88
3.3.3.1 Clinical diagnosis of affected males and carrier females
3.3.3.2 Haplotype analysis of XLRP 88

3.3.4 Family RP 89
3.3.4.1 Clinical diagnosis of affected males and carrier females
3.3.4.2 Haplotype analysis of XLRP 89

3.3.5 Family RP 90
3.3.5.1 Clinical diagnosis of affected males and carrier females
3.3.5.2 Haplotype analysis of XLRP 90

3.3.6 Family RP 91
3.3.6.1 Clinical diagnosis of affected males and carrier females
3.3.6.2 Haplotype analysis of XLRP 91

3.3.7 Family RP 92
3.3.7.1 Clinical diagnosis of affected males and carrier females
3.3.7.2 Haplotype analysis of XLRP 92
3.3.8 Family RP 93
   3.3.8.1 Clinical diagnosis of affected males and carrier females 178
   3.3.8.2 Haplotype analysis of XLRP 93 178
3.3.9 Family RP 94
   3.3.9.1 Clinical diagnosis of affected males and carrier females 180
   3.3.9.2 Haplotype analysis of XLRP 94 180
3.3.10 Family RP 1120
   3.3.10.1 Clinical diagnosis of affected males and carrier females 183
   3.3.10.2 Haplotype analysis of XLRP 1120 183
3.3.11 Family RP 88/10
   3.3.11.1 Clinical diagnosis of affected males and carrier females 185
   3.3.11.2 Haplotype analysis of XLRP 88/10 185
3.3.12 Family RP 88/75
   3.3.12.1 Clinical diagnosis of affected males and carrier females 188
   3.3.12.2 Haplotype analysis of XLRP 88/75 188
3.3.13 Choroideremia family 22 (CHM 22)
   (Spanish family misdiagnosed as XLRP) 191
   3.3.13.1 Clinical diagnosis of affected males and carrier females 191
   3.3.13.2 Haplotype analysis of CHM 22 using markers in the
   Xp21.1-Xp11.23 region 194
   3.3.13.3 Haplotype analysis of CHM 22 using markers in the
   Xq21.1-Xq21.2 region of the Choroideremia
disease interval 195
3.3.14 Summary of Families 197

3.3A Further genetic analysis of the 12 XLRP families 197a
3.3A.1 Linkage analysis of eight European XLRP families 197a
3.3A.2 Founder effect analysis of eight European XLRP families 197e

3.4 Discussion 198

3.5 Conclusion 202
4. cDNA SELECTION FOR ENRICHED TRANSCRIPTS FROM GENOMIC CLONES

4.1 Introduction

4.1.1 General description of the cDNA direct selection method
4.1.2 Key aspects of the methodology
4.1.3 Technical considerations
4.1.4 Potential limitations and possible solutions
4.1.5 Application of the cDNA selection technique to the RP2 region
4.1.6 Elements of the YAC contig used
4.1.6.1 Positive controls used in selection procedure
4.1.7 cDNA libraries used
4.1.8 Pre-blocking agents used

4.2 Establishment and application of the cDNA selection method to YACs mapping in the critical region of the RP2 locus

4.2.1 Preparation of cDNA
4.2.1.1 Amplification & titration of adult retinal cDNA libraries
4.2.1.2 Testing for known retinal genes in cDNA libraries
4.2.1.3 PCR amplification and preparation of cDNA inserts using vector primers of λZAPII and λgt10
4.2.1.4 Wizard PCR Preps, DNA purification and precipitation of cDNA inserts

4.2.2 Preparation of genomic DNA
4.2.2.1 Electrophoretic separation of YACs by pulsed field gel electrophoresis.
4.2.2.2 Digestion and ligation of YAC genomic DNA to linkers
4.2.2.3 5'-Biotinylated PCR of linked genomic DNA
4.2.2.4 Assessment of quality of linked genomic DNA using Southern hybridisation
4.2.2.5 Precipitation and resuspension of 5'-biotinylated PCR amplified YAC genomic DNA
4.2.3 Preparation of Pre-blocking quenching agents (Yeast DNA, Ribosomal DNAs, Human Placental DNA and LINE1 DNA) 226
4.2.3.1 Isolation of insert DNA from plasmid clones 227
4.2.3.2 Recovery and purification of insert DNA from Low-melting point agarose gels 229
4.2.3.2.1 Glasswool method 229
4.2.3.2.2 β-agarase method 229
4.2.3.3 Sonication of quenching agents 230
4.2.4 First round of direct cDNA selection 234
4.2.4.1 Pre-blocking of genomic DNA with quenching agents 234
4.2.4.2 Solution hybridisation of preblocked genomic DNA to cDNA inserts 235
4.2.4.3 Capture, wash and elution of specific primary selected cDNAs 235
4.2.4.4 PCR amplification of primary selected cDNAs 236
4.2.4.5 Purification & precipitation of primary selected cDNAs 238
4.2.5 Second round of direct cDNA selection 239
4.2.5.1 Pre-blocking of genomic DNA with quenching agents 239
4.2.5.2 Solution hybridisation of genomic DNA to primary selected cDNA inserts 239
4.2.5.3 Capture, wash and elution of specific secondary selected cDNAs 239
4.2.5.4 PCR amplification of secondary selected cDNAs 240
4.2.5.5 Purification and precipitation of secondary selected cDNAs 241
4.2.6 Qualitative evaluation for enrichment & depletion of cDNA transcripts by screening starting, primary and secondary selected cDNAs 241
4.2.6.1 Bio-dot preparation 241
4.2.6.2 Southern blot preparation 242
4.2.6.3 Hybridisation of dot-bLOTS and Southern blots with β-actin cDNA, ribosomal DNA, yeast DNA and a known retinal gene in the region 242
4.2.6.4 Evaluation of enrichment of selected transcripts and depletion of non-selected transcripts 249

4.3 Discussion 251
4.4 Conclusion 253
5. ANALYSIS OF SELECTED SUB-LIBRARY CLONES

5.1 Introduction

5.2 Pre-screening λZAPII and λgt10 starting cDNA libraries with internal control genes rhodopsin and TIMP-1

5.3 Construction of YAC specific cDNA sub-libraries from secondary selected cDNAs

5.3.1 Cloning of secondary selected cDNAs

5.3.1.1 PCR primer design for PCR-Direct cloning and amplification of cDNA inserts

5.3.1.2 Purification of PCR products using CHROMA SPIN Columns

5.3.1.3 Treatment of PCR products with T4 DNA polymerase

5.3.1.4 Hybridisation of PCR insert to pDIRECT vector

5.3.2 Transformation of cloned cDNAs and blue/white screening for recombinant plasmids

5.3.3 Transfer of recombinant clones onto master plates, glycerol stocks and preparation of sub-library filters

5.4 Sub-library filter hybridisation screening for quantitative evaluation of secondary selected libraries

5.4.1 Assessing for contaminating sequences

5.4.1.1 Probing with human placental DNA and LINE 1 DNA

5.4.1.2 Probing with ribosomal DNA

5.4.1.3 Probing with yeast DNA

5.4.1.4 Probing with negative control gene, β-actin cDNA

5.4.2 Assessing for enrichment of internal positive control genes

5.4.2.1 Probing with TIMP-1 cDNA fragment

5.4.2.2 Probing with rhodopsin cDNA fragment

5.4.2.3 Probing with the properdin cDNA fragment

5.5 Further analysis of sub-libraries

5.5.1 Colony PCR of selected cDNA clones

5.5.2 Sequence analysis using Gapped BLAST

5.5.3 Further contaminants in clones: vector sequences

5.5.4 Clones containing novel cDNAs

5.5.5 Clones with a poly-A tail
5.6 Description of selected cDNA clones

Description of twelve sequenced clones used in mapping back studies

5.6.1 Clone 4HG2-76
5.6.2 Clone 4HG2-108
5.6.3 Clone 4HG2-110
5.6.4 Clone 4HG2-125
5.6.5 Clone 4HG2-162
5.6.6 Clone 4HG2-164
5.6.7 Clone 4HG2-218
5.6.8 Clone 33CA11-239
5.6.9 Clone 33CA11-276
5.6.10 Clone 33CA11-283
5.6.11 Clone A1220-164
5.6.12 Clone A1220-238

Description of six sequenced clones not used in mapping back studies

5.6.13 Clone 4HG2-49
5.6.14 Clone 4HG2-51
5.6.15 Clone 4HG2-140
5.6.16 Clone 4HG2-201
5.6.17 Clone 4HG2-215
5.6.18 Clone A1220-169

5.7 Mapping studies of twelve selected cDNA clones

5.7.1 Choice, design and testing of internal primers of twelve selected cDNAs
5.7.2 Mapping studies of twelve selected cDNAs to the YAC genomic contig

5.8 Discussion

5.9 Conclusion
6. INVESTIGATION OF THE HUMAN TIMP-1 GENE FOR A CAUSATIVE ROLE IN THE PATHOGENESIS OF X-LINKED RETINITIS PIGMENTOSA-2

6.1 Introduction
6.2 Expression of human TIMP-1
6.3 Determination of the genomic structure of human TIMP-1
6.4 Mutation screening in patients with Retinitis Pigmentosa 2
6.5 Discussion
6.6 Conclusion

7. GENERAL DISCUSSION

7.1 Overview of the work described in this study
7.2 Apoptosis and potential therapeutic roles for RP disease
7.3 Potential gene therapy for the treatment of RP
7.4 Recent discovery of the RP2 gene and its implications
7.5 Future work

7A. SUMMARISED GENERAL DISCUSSION

7.1A Overview of the work described in this study
7.2A Future work

REFERENCES

PUBLICATIONS
# List of Figures

1.1 Trends in methods for cloning human disease genes, 1980-2010 36  
1.2 Schematic representation of the positional candidate approach 36  
1.3 A biotin-streptavidin capture system for direct cDNA selection 60  
1.4 Cloned and/or mapped genes causing retinal degeneration in humans and retina-specific genes 71  
1.5 Transverse section of the eye 73  
1.6 Schematic diagram of cell types and histologic layers of the human retina 75  
1.7 Illustrations of a typical rod and cone 77  
1.8 Key components of the phototransduction cascade in vertebrate rod photoreceptor cells 79  
1.9 The proximal short arm of the human X chromosome consensus map 98  
1.10 Genetic localisations of inherited eye disorders on proximal Xp 103  
1.11 Fundus photograph of a retina in a choroideremia carrier 111  
1.12 Fundus photograph of a retina in the late stages of XLRP 113  
3.1 Schematic diagram showing the order and approximate location and distance (cM) of the 28 genetic markers from Xp22.13 to Xp11.22 used in the haplotype analysis study 156  
3.2 Pedigree of XLRP 85 showing haplotypes constructed with the markers listed 159  
3.3 Pedigree of XLRP 87 showing haplotypes constructed with the markers listed 162  
3.4 Pedigree of XLRP 88 showing haplotypes constructed with the 9 markers listed 165  
3.5 Pedigree of XLRP 89 showing haplotypes constructed with the 11 markers listed 168  
3.6 Pedigree of XLRP 90 shows haplotypes constructed with the 9 markers listed 170  
3.7 Pedigree of XLRP 91 showing haplotypes constructed with 9 of the markers listed 173
3.8 Pedigree of XLRP 92 showing haplotypes constructed with the 11 markers listed 176
3.9 Pedigree of XLRP 93 showing haplotypes constructed with the 8 markers listed 179
3.10 Pedigree of XLRP 94 shows haplotypes constructed with 10 of the markers listed 181
3.11 Pedigree of XLRP 1120 showing haplotypes constructed with the 11 markers listed 184
3.12 Pedigree of XLRP 88/10 showing haplotypes constructed with the 10 markers listed 186
3.13 Pedigree of XLRP 88/75 showing haplotypes constructed with the 10 markers listed 189
3.14 Pedigree of CHM 22 shows haplotypes constructed with the 10 markers listed in the Xp21.1-Xp11.23 region 192
3.15 Pedigree of CHM 22 showing haplotypes constructed with the 6 markers listed in the Xq21.1-Xq21.2 region 196
4.1 General scheme describing the direct cDNA selection method for magnetic bead capture of expressed cDNAs from YACs 206
4.2 Section of YAC contig on Xp11.23 showing the four overlapping YACs used for cDNA selection 210
4.3 Testing for known retinal genes in the λgt10 cDNA library 215
4.4 Unpurified cDNAs inserts from PCR amplified starting cDNA libraries using λZAPII and λgt10 vector primers 217
4.5 Purified cDNA inserts from PCR amplified λZAPII and λgt10 starting cDNAs 219
4.6 Electrophoretic separation of YAC genomic DNAs using pulse field gels 221
4.7 Assessment of quality of linkered YAC genomic DNA 224
4.8 Isolation of insert DNA from ribosomal RNA plasmid clones 228
4.9 Purified insert DNAs isolated from LMP agarose using β-agarase 1 231
4.10 Qualitative and quantitative analysis of sonication products of quenching agents 232
4.11 Comparison of PCR amplified products from first round and second round cDNA selection 237
4.12 Qualitative evaluation for enrichment and depletion of cDNA transcripts 244
4.13 Monitoring enrichment of the internal control TIMP-1 cDNA using Southern blot hybridisation analysis 247

4.14 Monitoring selection of the external control rhodopsin cDNA using Southern blot hybridisation analysis 248

4.15 Monitoring depletion of the negative control β-actin cDNAs using Southern blot hybridisation analysis 250

5.1 Prescreening the λgt10 starting adult retinal cDNA library 257

5.2A Schematic map of the pDIRECT vector 259

5.2B Schematic diagram of the PCR-Direct cloning system 260

5.3 Screening the YAC specific λgt10 secondary selected sub-libraries
5.3A Hybridisation of duplicate filters of the 4HG2 YAC λgt10 secondary selected sublibrary with total human DNA 266

5.3B Hybridisation of duplicate filters of the A1220 YAC λgt10 secondary selected sub-library with the TIMP-1 cDNA 269

5.3C Hybridisation of duplicate filters of the 33CA11 YAC λgt10 secondary selected sub-library with properdin cDNA 272

5.4 PCR analysis of recombinant clones isolated from YAC 4HG2 secondary selected sub-library 274

5.5 Nucleotide sequence data of a contaminant vector clone 277

5.6.1A Clone 4HG2-76 T7 end sequence data and sequence alignments obtained after Gapped BLAST analysis 280

5.6.1B Clone 4HG2-76 T3 end sequence data and sequence alignments obtained after Gapped BLAST analysis 281

5.6.2 Clone 4HG2-108 T3 end sequence data and sequence alignments obtained after Gapped BLAST analysis 283

5.6.3 Clone 4HG2-110 T3 end sequence data and sequence alignments obtained after Gapped BLAST analysis 285

5.6.4 Clone 4HG2-125 T3 end sequence data and sequence alignments obtained after Gapped BLAST analysis 286

5.6.5 Clone 4HG2-162 T7 end sequence data and sequence alignments obtained after Gapped BLAST analysis 286

5.6.6 Clone 4HG2-164 T3 end sequence data and sequence alignments obtained after Gapped BLAST analysis 288

5.6.7A Clone 4HG2-218 T7 end sequence data and sequence alignments obtained after Gapped BLAST analysis 290

5.6.7B Clone 4HG2-218 T3 end sequence data and sequence alignments obtained after Gapped BLAST analysis 292
5.6.8 Clone 33CA11-239 T3 end sequence data and sequence alignments obtained after Gapped BLAST analysis

5.6.9 Clone 33CA11-276 T3 end sequence data and sequence alignments obtained after Gapped BLAST analysis

5.6.10 Clone 33CA11-283 T3 end sequence data and sequence alignments obtained after Gapped BLAST analysis

5.6.11 Clone A1220-164 T7 end sequence data and sequence alignments obtained after Gapped BLAST analysis

5.6.12 Clone A1220-238 T3 end sequence data and sequence alignments obtained after Gapped BLAST analysis

5.6.13 Clone 4HG2-49 T7 end sequence data and sequence alignments obtained after Gapped BLAST analysis

5.6.14 Clone 4HG2-51 T3 and T7 end sequence data and sequence alignments obtained after Gapped BLAST analysis

5.6.15 Clone 4HG2-140 T3 end sequence data and sequence alignments obtained after Gapped BLAST analysis

5.6.16 Clone 4HG2-201 T7 end sequence data and sequence alignments obtained after Gapped BLAST analysis

5.6.17 Clone 4HG2-215 T7 end sequence data and sequence alignments obtained after Gapped BLAST analysis

5.6.18 Clone A1220-169 T3 and T7 end sequence data and sequence alignments obtained after Gapped BLAST analysis

5.7 Mapping studies of cDNA clone 4HG2-218 using PCR analysis

6.1 Identification of a common polymorphism in TIMP-1 gene in an unrelated male patient with X-linked retinitis pigmentosa 2.
List of Tables

1.1 Some successes of disease gene isolation using positional candidate gene approaches 35
1.2 Summary of some advantages and disadvantages of direct cDNA selection 61
1.3 Experimental steps necessary to analyse a YAC for transcribed sequences 66
1.4 Biases and problems of different gene identification methods 67
1.5 Genetic maps for the X chromosome 95
3.1 Microsatellite markers used for genetic characterisation of XLRP families 157
3.2 Haplotype data for 18 microsatellite markers used in the study of XLRP 85 160
3.3 Haplotype data for 18 microsatellite markers used in the study of XLRP 87 163
3.4 Haplotype data for 17 microsatellite markers used in the study of XLRP 88 166
3.5 Haplotype data for 19 microsatellite markers used to refine the XLRP2 interval in RP 89 169
3.6 Haplotype data for 17 microsatellite markers used in the study of XLRP 90 171
3.7 Haplotype data for 12 microsatellite markers used to establish critical crossover events in XLRP 91 174
3.8 Haplotype data for 18 microsatellite markers used in the study of XLRP 92 177
3.9 Haplotype data for 14 microsatellite markers used in the study of XLRP 94 182
3.10 Haplotype data for 14 microsatellite markers used in the study of XLRP 88/10 187
3.11 Haplotype data for 15 microsatellite markers used in the study of XLRP 88/75 190
3.12 Haplotype data for 15 microsatellite markers in the Xp21.1-Xp11.23 region used in the study of CHM 22 family 193
3.13 Summary of families 197
3.14 Mutations identified in the RPGR gene in four XLRP families 197a
3.15 Two-point lod scores between Xp21.1-Xp11.22 markers and XLRP disease phenotype in 8 European XLRP families 197b
3.16 X-linked retinitis pigmentosa-associated haplotypes in eight European pedigrees 197e
5.1 Combined percentages of the contaminating and enriched sequences of the YAC secondary selected cDNA sub-libraries 264
5.2 Summary of sequence analysis of 18 selected cDNA clones 319
6.1 Intron/exon structure and splice junctions for the human TIMP-1 gene 336
6.2 Primers and conditions for PCR amplification of the human TIMP-1 gene 337
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Å</td>
<td>ångstrom unit = (10^{-10}) m (0.1 nm)</td>
</tr>
<tr>
<td>ABCR</td>
<td>retinal rod-specific ATP binding cassette transporter</td>
</tr>
<tr>
<td>ADRP</td>
<td>autosomal dominant retinitis pigmentosa</td>
</tr>
<tr>
<td>AIED</td>
<td>Aland island eye disease</td>
</tr>
<tr>
<td>AR</td>
<td>androgen receptor</td>
</tr>
<tr>
<td>ARAF-1</td>
<td>proto-oncogene A-raf-1</td>
</tr>
<tr>
<td>ARRP</td>
<td>autosomal recessive retinitis pigmentosa</td>
</tr>
<tr>
<td>ATP</td>
<td>adenosine 5'-triphosphate</td>
</tr>
<tr>
<td>BACs</td>
<td>bacterial artificial chromosomes</td>
</tr>
<tr>
<td>BCM</td>
<td>blue cone monochromatism</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic Local Alignment Search Tool</td>
</tr>
<tr>
<td>bp</td>
<td>base pair</td>
</tr>
<tr>
<td>BRCA1</td>
<td>gene for hereditary breast and ovarian cancer</td>
</tr>
<tr>
<td>BSA</td>
<td>bovine serum albumin</td>
</tr>
<tr>
<td>cDNA</td>
<td>complementary DNA</td>
</tr>
<tr>
<td>cen</td>
<td>centromere</td>
</tr>
<tr>
<td>CEPH</td>
<td>Centre d’Etude du Polymorphisme Humaine</td>
</tr>
<tr>
<td>CF</td>
<td>cystic fibrosis</td>
</tr>
<tr>
<td>CGD</td>
<td>chronic granulomatous disease</td>
</tr>
<tr>
<td>cGMP</td>
<td>cyclic GMP</td>
</tr>
<tr>
<td>CHEF</td>
<td>clamped homogeneous electric fields</td>
</tr>
<tr>
<td>CHLC</td>
<td>Co-operative Human Linkage Centre</td>
</tr>
<tr>
<td>CHM</td>
<td>Choroideremia</td>
</tr>
<tr>
<td>cM</td>
<td>centimorgan</td>
</tr>
<tr>
<td>CMC</td>
<td>chemical mismatch cleavage</td>
</tr>
<tr>
<td>CMGT</td>
<td>chromosome-mediated gene transfer</td>
</tr>
<tr>
<td>CNCG</td>
<td>cGMP-gated ion channel</td>
</tr>
<tr>
<td>COD</td>
<td>Cone-rod dystrophy</td>
</tr>
<tr>
<td>CRALBP</td>
<td>cellular retinaldehyde-binding protein</td>
</tr>
<tr>
<td>CRX</td>
<td>cone-rod homeobox gene</td>
</tr>
<tr>
<td>CSNB</td>
<td>Congenital stationary night blindness</td>
</tr>
<tr>
<td>dATP</td>
<td>deoxyadenosine triphosphate</td>
</tr>
<tr>
<td>dCTP</td>
<td>deoxyctydine triphosphate</td>
</tr>
<tr>
<td>dGTP</td>
<td>deoxyguanosine triphosphate</td>
</tr>
<tr>
<td>dTTP</td>
<td>deoxythymidine triphosphate</td>
</tr>
</tbody>
</table>
DDBJ  DNA Database of Japan
DEAE  diethylaminoethyl
DGGE  denaturing gradient gel electrophoresis
DMD  Duchenne muscular dystrophy
DNA  deoxyribonucleic acid
DNAse  deoxyribonuclease
DS  direct sequencing
EBI  European Bioinformatics Institute
ECM  extracellular matrix
EDA  ectodermal dysplasia locus
EDTA  ethylene diaminetetraacetic acid
ELK  ets-related gene
EPA  erythroid-potentiating activity
ERG  electroretinography
EST  expression sequence tag
FBN  fibrillin
FISH  fluorescent in-situ hybridisation
g  grams
$g$  acceleration due to gravity
G6PD  glucose-6-phosphate dehydrogenase
GCAP  guanylate cyclase activating protein
GDB  Genome database
GDP  guanine diphosphate
GMP  guanine monophosphate
GTP  guanine triphosphate
h  hour
HA  heteroduplex analysis
HD  Huntingdon disease
HGMP  Human Genome Mapping Project
HGPRT  hypoxanthine-guanine phosphoribosyl transferase
HNPPC  non-polyposis colon cancer
HSP  high-scoring segment pair
HTLV-2  human T-cell leukemia virus type-2
IMAGE  integrated molecular analysis of gene expression
IPTG  isopropylthio-β-D-galactoside
kb  kilobase
kbp  kilobase pair
L1  LINE 1
LB  luria broth
LCR  locus control region
LINEs  long interspersed nuclear elements
LMP  low melting point
mg  milligram
ml  millilitre
min  minute
mM  millimolar
mRNA  messenger mRNA
M  molar
MAO  monoamine oxidase
Mb  megabase
MGD  mouse genome database
MHC  major histocompatibility complex
MMP  matrix metalloproteinases
MNK  Menkes disease
MPC  magnetic particle concentrator
MSP  maximal-scoring segment pair
M.W.M.  molecular weight marker
ng  nanogram
NCBI  National Centre for Biotechnology Information
ND  Norrie disease
NDP  Norrie disease protein
NTP  nucleotide triphosphate
OA1  Ocular albinism type-1
OAT  ornithine aminotransferase
OMIM  On-line mendelian inheritance in man
ORF  open reading frame
OTC  ornithine transcarbamylase
PACs  PI-derived artificial chromosomes
PAGE  polyacrylamide gel electrophoresis
PCR  polymerase chain reaction
PDE  cGMP phosphodiesterase
PDEA  cGMP phosphodiesterase α subunit
PDEB  cGMP phosphodiesterase β subunit
PFC  properdin
PEG  polyethylene glycol
PFGE  pulsed-field gel electrophoresis
p.f.u.  plaque forming units
PGK  phosphoglycerate kinase
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>PKD</td>
<td>Polycystic kidney disease</td>
</tr>
<tr>
<td>PMSF</td>
<td>phenyl-methyl-sulphonyl-fluoride</td>
</tr>
<tr>
<td>Rab</td>
<td>Rab geranyl geranyl transferase</td>
</tr>
<tr>
<td>RCC-1</td>
<td>regulator of chromosome condensation type-1</td>
</tr>
<tr>
<td>RD</td>
<td>retinal degeneration</td>
</tr>
<tr>
<td>RDS</td>
<td>retinal degeneration slow</td>
</tr>
<tr>
<td>REP</td>
<td>Rab escort protein transferase</td>
</tr>
<tr>
<td>RETGC</td>
<td>retinal guanylate cyclase</td>
</tr>
<tr>
<td>RFLP</td>
<td>restriction fragment length polymorphism</td>
</tr>
<tr>
<td>RH</td>
<td>radiation hybrid</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>RNAse</td>
<td>ribonuclease</td>
</tr>
<tr>
<td>ROM1</td>
<td>rod outer membrane protein</td>
</tr>
<tr>
<td>RPE</td>
<td>retinal pigment epithelium</td>
</tr>
<tr>
<td>RPE65</td>
<td>a 65-kD protein specific to the RPE</td>
</tr>
<tr>
<td>RPGR</td>
<td>retinitis pigmentosa GTPase regulator</td>
</tr>
<tr>
<td>RT-PCR</td>
<td>reverse transcription PCR</td>
</tr>
<tr>
<td>S</td>
<td>specific activity of dCTP</td>
</tr>
<tr>
<td>SA</td>
<td>specific activity</td>
</tr>
<tr>
<td>sec</td>
<td>second</td>
</tr>
<tr>
<td>SETS</td>
<td>specific exon trapping system</td>
</tr>
<tr>
<td>SDS</td>
<td>sodium dodecyl sulphate</td>
</tr>
<tr>
<td>SDW</td>
<td>sterile distilled water</td>
</tr>
<tr>
<td>SFD</td>
<td>Sorsby fundus dystrophy</td>
</tr>
<tr>
<td>SHAC</td>
<td>selection of hybrids by affinity capture</td>
</tr>
<tr>
<td>SINE</td>
<td>short interspersed nuclear element</td>
</tr>
<tr>
<td>SMA</td>
<td>spino-bulbar muscular atrophy</td>
</tr>
<tr>
<td>SSCP</td>
<td>single stranded conformational polymorphism</td>
</tr>
<tr>
<td>SSCP</td>
<td>single stranded conformational polymorphism</td>
</tr>
<tr>
<td>SSLP</td>
<td>simple sequence length polymorphism</td>
</tr>
<tr>
<td>SSR</td>
<td>short sequence repeat</td>
</tr>
<tr>
<td>STRP</td>
<td>short tandem-repeat polymorphism</td>
</tr>
<tr>
<td>STS</td>
<td>sequence tagged site</td>
</tr>
<tr>
<td>SYN1</td>
<td>synapsin 1</td>
</tr>
<tr>
<td>SYP</td>
<td>synaptophysin</td>
</tr>
<tr>
<td>TAE</td>
<td>Tris-acetate EDTA buffer</td>
</tr>
<tr>
<td>TBE</td>
<td>Tris-borate EDTA buffer</td>
</tr>
<tr>
<td>TCA</td>
<td>trichloroacetic acid</td>
</tr>
<tr>
<td>TE</td>
<td>Tris-EDTA buffer</td>
</tr>
<tr>
<td>Tel</td>
<td>telomere</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>TEMED</td>
<td>N, N, N', N'-tetramethyl 1,1,2-diaminoethane</td>
</tr>
<tr>
<td>TIMP</td>
<td>tissue inhibitor of metalloproteinase</td>
</tr>
<tr>
<td>Tm</td>
<td>melting temperature (primer)</td>
</tr>
<tr>
<td>tRNA</td>
<td>transfer RNA</td>
</tr>
<tr>
<td>μg</td>
<td>microgram</td>
</tr>
<tr>
<td>μl</td>
<td>microlitre</td>
</tr>
<tr>
<td>μM</td>
<td>micromolar</td>
</tr>
<tr>
<td>UBE</td>
<td>ubiquitin activating enzyme</td>
</tr>
<tr>
<td>UTR</td>
<td>untranslated region</td>
</tr>
<tr>
<td>UV</td>
<td>ultraviolet</td>
</tr>
<tr>
<td>v/v</td>
<td>volume/volume</td>
</tr>
<tr>
<td>VNTRs</td>
<td>variable number of tandem repeats</td>
</tr>
<tr>
<td>w/v</td>
<td>weight/volume</td>
</tr>
<tr>
<td>w/w</td>
<td>weight/weight</td>
</tr>
<tr>
<td>WAS</td>
<td>Wiscott-Aldrich syndrome</td>
</tr>
<tr>
<td>WASP</td>
<td>Wiscott-Aldrich syndrome protein</td>
</tr>
<tr>
<td>WWW</td>
<td>World Wide Web</td>
</tr>
<tr>
<td>X-gal</td>
<td>5-bromo-4-chloro-3-indolyl-β-D-galactoside</td>
</tr>
<tr>
<td>XIC</td>
<td>X inactivation centre</td>
</tr>
<tr>
<td>XIST</td>
<td>X-inactive specific transcript</td>
</tr>
<tr>
<td>XK</td>
<td>McLeod syndrome</td>
</tr>
<tr>
<td>XLA</td>
<td>X-linked agammaglobulinaemia</td>
</tr>
<tr>
<td>XLPCD</td>
<td>X-linked progressive cone dystrophy</td>
</tr>
<tr>
<td>XLRP</td>
<td>X-linked retinitis pigmentosa</td>
</tr>
<tr>
<td>XLRSl</td>
<td>X-linked juvenile retinoschisis 1 gene</td>
</tr>
<tr>
<td>YACs</td>
<td>yeast artificial chromosomes</td>
</tr>
<tr>
<td>ZNF</td>
<td>zinc finger protein</td>
</tr>
</tbody>
</table>
CHAPTER ONE

INTRODUCTION
CHAPTER ONE
INTRODUCTION

1.1 Principles and strategies in identifying disease genes

Genetic disease is a major cause of human suffering. Before 1980, only a few human disease-causative genes had been identified. Early successes were largely the result of exceptional characteristics: the biochemical basis of the disease had previously been established and purification of the gene product could be achieved without too much difficulty. Such advantages do not apply, however, to the great majority of diseases resulting from mutation in human genes. In the 1980's, the application of recombinant DNA technology offered new approaches to mapping and identification of the genes underlying inherited single gene disorders and somatic cancers, and the number of disease genes identified started to increase rapidly. With the subsequent advent of rapid PCR-based linkage studies and PCR mutation screening technologies, the identification of novel disease genes became commonplace.

It is important to note that not all 65 000-80 000 human genes will be identified as disease genes. Some genes are indispensable to embryonic function so that deleterious mutations result in embryonic lethality and go unrecorded in humans. In other cases, abolition of gene function may have no effect on the phenotype because of genetic redundancy, that is other non-allelic genes also supply the same function. Mendelian Inheritance in Man, the catalogue of inherited human disorders, currently lists about 5000 mendelian traits. There is not a one-to-one correspondence between genes and clinical syndromes. This is because different mutations in one gene cause different phenotypes, and frequently the same disease phenotype can be caused by mutations in different genes.

The more common inherited disorders are the ones that are the most difficult to study by molecular genetics: a combination of different genes is often involved (oligogenic or polygenic disorders) as well as different environmental triggers. Similarly, the more common cancers involve cellular events in which multiple genes are involved. Not unexpectedly, therefore, the human disease genes that have been isolated to date are very largely those responsible for inherited single gene disorders, or somatic cancers where there is a single major susceptibility gene. The strategies used to identify genes involved in such single gene disorders are considered below.
The choice of strategy for identifying a disease gene depends on what resources (animal models, chromosomal abnormalities, clone libraries, etc.) are available, and on how much is known about the pathogenesis of the disease. Several of the possible strategies aim initially to identify a number of candidate genes, which then have to be tested individually for evidence that implicates them as the disease-causative gene. Several different methods have been used to identify candidate genes such as functional cloning, positional cloning, position-independent candidate gene approaches and positional candidate gene approaches, all of which are discussed below.

1.1.1 Functional Cloning

Functional cloning uses information about the function of an unidentified disease gene in order to identify the gene. If the gene product is known, partial purification of the product can permit various strategies for identifying the underlying gene, such as partial amino acid sequence determination and then synthesis of gene-specific oligonucleotides to screen cDNA libraries. For example, the identification of the factor VIII gene, the locus for hemophilia A, was achieved by product-directed oligonucleotide screening of DNA (Gitschier et al., 1984). In other cases antibodies against the protein product were utilized. Phenylketonuria was known to be caused by a lack of the enzyme phenylalanine hydroxylase. The enzyme was purified from liver, where it was known to be expressed. Specific antibodies were raised and employed to immunoprecipitate polysomes containing phenylalanine hydroxylase mRNA (Robson et al., 1982). The purified mRNA was then converted to cDNA and a specific cDNA clone was isolated.

An alternative method, a functional complementation assay, is possible even when the gene product is unknown. For example genes involved in human DNA repair have been identified by functional complementation cloning. A variety of mammalian cell lines have been generated which are deficient in DNA repair. They show abnormal responses following exposure to UV irradiation or chemical mutagens. These mutant cells, or alternatively cells derived from patients with a DNA repair deficiency, can be transformed with fragments of normal human DNA or human chromosomes in order to produce a repair-competent phenotype. In this way cDNA clones for the Fanconi's anemia group C (FACC) gene were first obtained (Strathdee et al., 1992). While these functional cloning strategies have been successful in a few cases, they are limited to disorders for which biological information about the basic genetic defect exists. Unfortunately, for the vast majority of single gene disorders no such detailed functional insight exists. Hence mapping the disease to a specific sub-chromosomal localisation is generally the most fruitful initial step.
1.1.2 Positional Cloning

Positional cloning involves the cloning of a gene knowing nothing except its sub-chromosomal location (Collins, 1992, 1995). The general approach is to try to construct physical and genetic maps of the region, refine the sub-chromosomal localisation, and then identify genes within this region to investigate as disease gene candidates. Often the initial localisation is to a large candidate region of 10Mb or more. These localisations may come from various sources such as linkage analysis and chromosomal abnormalities.

Linkage analysis involves the study of pedigrees in which a particular disease phenotype is segregating. Analysis of very large families is highly beneficial if the disease is caused by mutations at more than one locus. The availability of any gross rearrangements in the gene, even if present in only a small subset of patients, speeds up the process of gene identification. Many positional cloning projects have relied upon the discovery of rearrangements detectable at least at Southern blot level. In several instances the process has been dramatically accelerated by the availability of gross visible cytogenetic rearrangements which interrupt or delete the disease locus. For example, translocations and deletions identified Xp21 as the location of the Duchenne muscular dystrophy (DMD) gene (Kunkel et al., 1985). A special case of the value of rearrangement is provided by the discovery of repeated sequences located within the genes for X-linked spinal and bulbar muscular atrophy (La Spada et al., 1991), fragile X syndrome (Verkerk et al., 1991) and myotonic dystrophy (DM) (Harley et al., 1992). In each of these instances, a trinucleotide repeat sequence is expanded to hundreds (or thousands) of copies in affected individuals, resulting in a visible abnormality on Southern analysis. The length of the expanded repeat appears to correlate with severity of disease (especially in DMD). The discovery of three such disorders in the space of one year suggests that this may be a general phenomenon responsible for a significant percentage of human genetic disease.

A first requirement for narrowing down the candidate region is a series of clones and markers from within this region. Polymorphic markers are needed to generate a high resolution genetic map of the region, and clones are required both as a source of such markers and to start physical mapping. With the progress of the Human Genome Project, many of the resources required have already been generated on a genome-wide scale. In many chromosomal regions, existing high resolution genetic maps allow a disease gene to be placed between markers only 1cM apart, if sufficient family material is available. If no further information is available, the effort then reduces to a search through the candidate region for all transcripts, followed by analysis of each of these for alterations in individuals with the disease. Other information, such as the pattern of
tissue expression of the various candidate genes, can be useful in investigating those of
greater priority during the search, but in general in a true positional cloning effort such
information will only be circumstantial and the ultimate proof will reside with sequence
analysis and the demonstration of mutations associated with the disease phenotype.

Positional cloning can be laborious, and is usually not seriously attempted until
genetic or physical mapping has located the disease gene to within 1-2 Mb of DNA. For
some disorders, such fine mapping is not practicable. If a disorder is rare, it may be very
difficult to identify crossovers that narrow the location of the disease gene to a small
interval, and the chance of identifying a patient with a suitable disease-associated
cytogenetic abnormality which could refine the map location, may be virtually zero. In
such cases candidate gene approaches are the only way forward.

1.1.3 Candidate gene approaches

An important conceptual difference between functional and positional cloning
and the candidate gene approach is that the latter does not require the isolation of new
genes but relies on the availability of information from previously isolated genes. All
approaches used to identify disease genes generate candidate genes, which then have to
be tested individually to see if there is compelling evidence that they are associated with
the disease in question. A particular gene is hypothesized to be the locus for the disease,
and the hypothesis is then tested by checking for evidence that the candidate gene is
associated with the disease.

1.1.3.1 Position-independent candidate gene approaches

A candidate gene for a human disorder may be suggested without any
knowledge of the chromosomal location. This can happen if the phenotype resembles
another phenotype in animals or humans for which the gene is known, or if the
molecular pathogenesis suggests that the gene may be a member of a known gene
family.

If an animal phenotype shows a striking similarity to a human disorder, then it
might result from mutations in the animal ortholog of the human disease gene. Then, if
the human gene is unknown but the animal gene is known, knowledge of the animal
ortholog can be used to help identify and characterize the human gene. This can then be
tested for its involvement in the disease.

Mutations at more than one locus may produce the same clinical result (locus
heterogeneity). If a gene is identified as the locus for one such disease, then genes closely
related to it in sequence or function may be candidates for similar diseases. For example,
the identified disease gene may be a member of a multigene family, other members of
which would be candidates for similar diseases. Candidate genes may also be suggested on the basis of a close functional relationship to a gene known to be the locus for a disease with a similar phenotype. The genes could be related by encoding a receptor and its ligand, or other interacting components in the same metabolic or developmental pathway.

For some disorders, observations on the pathogenesis may immediately suggest candidate genes with an appropriate expression pattern or function. For example, neural tube defects are likely to involve genes that are expressed shortly before or during neurulation in the developing embryo (the 3rd-4th weeks of human embryonic development). A gene which shows expression in the neural tube at these stages is therefore a candidate. This can be demonstrated by in situ hybridisation against mRNA in sections of embryos, most conveniently by using mouse embryonic sections at the equivalent developmental stages.

1.1.3.2 Positional candidate gene approaches

Confidence in a particular candidate disease gene is increased substantially if it can be shown to map to the same sub-chromosomal region as the disease gene. Such positional candidate gene approaches have already been very valuable (Table 1.1 lists some successes), but in the next few years they will completely dominate disease gene identification (Fig. 1.1), simply because of the rate at which sub-chromosomal locations are being established for both disease genes and human genes in general. With more and more human genes being assigned to sub-chromosomal locations, there is a high chance that database searches will reveal one or more possible candidate genes in the appropriate location.

Table 1.1 Some successes of disease gene isolation using positional candidate gene approaches

<table>
<thead>
<tr>
<th>Disease</th>
<th>Affected protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alzheimer's disease</td>
<td>β-Amyloid protein precursor, apoE</td>
</tr>
<tr>
<td>Charcot-Marie-Tooth disease type 1A</td>
<td>Myelin protein zero (P&lt;sub&gt;0&lt;/sub&gt;)</td>
</tr>
<tr>
<td>Charcot-Marie-Tooth disease type 1B</td>
<td>Peripheral myelin protein 22</td>
</tr>
<tr>
<td>Familial melanoma</td>
<td>p16</td>
</tr>
<tr>
<td>Hereditary nonpolyposis colon cancer</td>
<td>hMSH2, hMLH1, hPMS1, hPMS2</td>
</tr>
<tr>
<td>Malignant hypothermia</td>
<td>Ryanodine receptor</td>
</tr>
<tr>
<td>Marfan syndrome</td>
<td>Fibrillin</td>
</tr>
<tr>
<td>Multiple endocrine neoplasia type 2A</td>
<td>Receptor tyrosine kinase RET</td>
</tr>
<tr>
<td>Retinitis pigmentosa</td>
<td>Peripherin, Rhodopsin</td>
</tr>
<tr>
<td>Waardenburg syndrome type 1</td>
<td>Paired box gene PAX3</td>
</tr>
</tbody>
</table>
1. Introduction

Figure 1.1 Trends in methods for cloning human disease genes, 1980-2010. Exact quantitation is not intended; trends after 1995 are highly speculative. Collins, F.S. (1995)

Figure 1.2 Schematic representation of the positional candidate approach. When a new disease locus is assigned to a chromosomal region (shown in middle) candidate genes from the same region are analysed. The features of these candidate genes are compared with the features of that particular disease. Several types of features are listed in a specific order so that they can be matched with the disease features. Similarly, when a new gene is assigned to a chromosomal region candidate diseases are analysed by an analogous process. Ballabio, A. (1993)
When a new disease gene is assigned to a specific map position it may be possible to "query" that region of the chromosome in a database and have access to a list of genes assigned to the same region. These genes will be considered candidates for that particular disease. The features of the genes will then be compared to the features of the disease in order to find the gene which would be the strongest candidate for that particular disease. Subsequently, intragenic abnormalities of that gene will be screened for in affected individuals in order to determine if the candidate gene is the disease gene. This process could go both directions: from diseases to genes and from genes to diseases (Fig. 1.2). When a new gene is mapped to a region one will have access to a list of all disease loci mapped to that region. Similarly, the features of candidate diseases can be compared to the features of the newly isolated gene. The type of features to be compared obviously depends on the type of information available for both diseases and genes mapping to the same chromosomal region.

Some examples of positional candidate gene successes are the isolation of the X-linked agammaglobulinaemia (XLA) gene and the genes for hereditary non-polyposis colon cancer (HNPCC). The XLA gene was isolated using a strategy based entirely on positional cloning (Vetrie et al., 1993). The same gene was independently found by Tsukada et al., (1993), but this time it was isolated during a search for genes regulating B cell development. For HNPCC, linkage information pointing to chromosome 2p and phenotypic information about DNA instability in colon cancers from affected patients came together in a stunningly productive positional candidate confluence. Since then four genes involved in DNA mismatch repair have been implicated in HNPCC (Leach et al., 1993 and Nicolaides et al., 1994).

Sometimes one positional assignment can lead to another, as in the case of Marfan syndrome (MFS). The phenotype of MFS suggested some abnormality in a connective tissue component. Linkage analysis mapped the MFS gene to 15q, and subsequently the gene for the connective tissue protein fibrillin was localised to 15q21.1 by in situ hybridisation. Fibrillin was then a positional candidate, with additional support from the pathology. A patient-specific missense mutation was soon demonstrated in the fibrillin (FBN1) gene (Dietz et al., 1991). A second fibrillin gene was shown to map to 5q, which therefore became a candidate location for other Marfan-like phenotypes. A related condition, congenital contractual arachnodactyly was mapped to 5q and shown to be caused by mutations in the FBN2 gene (Putnam et al., 1995). Similarly, multiple endocrine neoplasia type 2 was mapped to chromosome 10 by linkage analysis. Perusal of the local map identified the gene for the RET proto-oncogene lying in the correct area; eventually leading to the identification of point mutations in affected patients, (Mulligan et al., 1993) and the surprising finding that other mutations in the same gene are capable of causing a completely different phenotype—Hirschsprung's disease (Edery et al., 1994)
Finally, if a tentative mouse model is available for a specific disease, comparison of map position between the murine locus and the candidate genes could be an important clue for the identification of the disease gene. The human-mouse synteny maps have been exploited to clone the gene responsible for Waardenburg syndrome, a hereditary disease responsible for over 2% of adult deafness. The tissues affected by this disease are related by their embryonic origin in the neural crest. One subtype of this syndrome, WS1, was mapped to a region of the long arm of chromosome 2 that is homologous to the proximal part of mouse chromosome 1. The mouse mutation *Splotch* maps to this region of synteny and the mutant mice are characterised by dysgenesis of neural crest cell-derived tissues. As the mutation in *Splotch* is the result of a deletion in *Pax-3* (Epstein *et al.*, 1991), the human homologue *HuP2* (Burri *et al.*, 1989) was investigated and mutations in this gene were identified in patients with Waardenburg syndrome (Tassabehji *et al.*, 1992).

The future expanding success of the positional candidate approach is predicted on an increasingly dense transcript map. Large databases of cDNA sequences may assist the pure candidate gene approach, but without associated mapping information their usefulness is greatly blunted. Since positional cloning efforts usually result in candidate intervals of 0.5-5 megabases (Mb), ideally the transcript map should also have this degree of resolution. Mapping of cDNAs to traditional somatic cell hybrids or by fluorescent in-situ hybridisation (FISH) will not usually achieve this; utilizing large insert clone libraries or radiation hybrids is more appropriate.

### 1.2 The Human Genome Project

The Human Genome Project is an international effort with the ultimate aim of obtaining a complete description of the human genome by DNA sequencing. The rationale of the Human Genome Project is to acquire fundamental information concerning our genetic make-up which will further the basic scientific understanding of human genetics and the role of various genes in health and disease. The major scientific thrust concerns constructing high resolution genetic and physical maps as a prelude to the ultimate physical map, the complete sequence of the human genome. Much of this work is being carried out in major genome mapping centres. There is also extensive interaction with research focusing on mapping disease genes, which is currently conducted in numerous laboratories throughout the world. The major 15-year goals also include a commitment to map and sequence the genomes of a variety of model organisms and to develop ancillary technologies including data analysis. Significant funding is also being provided for research on ethical, legal and social considerations, and support for technology transfer to the medical community.
1.2.1 Human Genetic Maps

The first human genetic markers were protein polymorphisms, notably blood group and serum protein markers, which are not very informative. The identification of abundant DNA-based polymorphisms was required to enable the construction of a complete linkage map of the human genome, and was first proposed using restriction fragment length polymorphisms (RFLP's), (Botstein et al., 1980). This was developed by using random single-copy DNA probes capable of detecting DNA sequence polymorphisms when hybridised to restriction enzyme digests of an individual's DNA. In addition to providing a framework for studying the nature of recombination in humans, a complete linkage map would benefit several important areas such as: gene localisation, any gene for which a polymorphism could be typed could immediately be placed on the genetic map, and its chromosomal localisation obtained. This has great use medically since linkage analyses could be employed in families segregating a disease gene, enabling the genes for many inherited conditions to be mapped for the first time, since with such a map, multilocus analysis would decrease the number of meiosis required to detect linkage (Lander and Botstein, 1986). The mapping of heterogeneous genetic disorders (Lander and Botstein, 1986) and rare recessive diseases (Lander and Botstein, 1987) would also become possible. Gene localisation would then provide starting points for efforts to clone genes. Also pre-natal or pre-symptomatic diagnosis of inherited disease genes by linkage analysis would be facilitated enormously by having many DNA markers in the vicinity of the disease locus.

In 1987 the first such map was published comprising 403 polymorphic loci, including 393 RFLP markers (Donis-Keller et al., 1987). The results provided 23 linkage groups (linear arrays of linked markers) corresponding to the 22 autosomes and the X chromosome. Important though this achievement was, there remained some serious drawbacks such as the average spacing between the markers which was >10 cM and the reliance on RFLP markers which are not very informative markers and are difficult to type. The next breakthrough in constructing a human genetic map came as a result of the use of microsatellite markers. Microsatellite markers (also described as short tandem repeat polymorphisms, or STRPs) have the advantage of being abundant, widely dispersed throughout the genome, highly informative and easy to type. The sequences of primers corresponding to each marker are electronically available and it is possible to carry out large numbers of assays rapidly in an automated fashion. The NIH/CEPH collaborative mapping group (1992) produced a genetic linkage map consisting of 1416 loci, of which 339 were microsatellite markers. They constituted more than 50% of reported markers with heterozygosity of at least 70%. Although 80% of the markers in the map reported were RFLPs, the rapid integration of microsatellites predicted that they
would become the dominant markers in the map. By focusing on this type of markers, a second-generation linkage map of the human genome (Weissenbach et al., 1992) was provided. This involved selecting suitably polymorphic CA/TG repeats, mapping them to specific chromosomes by typing panels of human-rodent somatic cell hybrids and performing statistical linkage analyses on markers from individual chromosomes. By 1994, the same group developed and mapped 1,267 new microsatellite markers and integrated them into their previous map. This third generation genetic linkage map comprised 2,066 (AC)n short tandem repeats, 60% of which have a heterozygosity of 0.7 (Gyapay et al., 1994) and covered a total distance of 3,690 cM. Subsequently, maps of increasing resolution have been produced by different genome centres (Murray et al., 1994 and Dib et al., 1996). The latest genetic map to be published (Dib et al., 1996) represents the last version of the Genethon map and is composed entirely of microsatellite markers. 5,264 markers were located to 2,335 positions using CEPH family data, so on average there is a marker every 0.7 cM or about every 700,000 base pairs. The markers on these maps provide a resource for ready transference to physical maps and hence the major effort has recently switched to the construction of high resolution physical maps of the human genome.

1.2.2 Physical maps of the human genome

Physical maps consist of ordered, overlapping cloned fragments of genomic DNA covering each chromosome. The different genetic maps of the human genome that have been assembled so far all represent the same concept: sets of linked polymorphic markers (linkage groups) corresponding to different chromosomes. However a variety of different types of physical map are possible. In a sense, the first physical map of the human genome was obtained when cytogenetic banding techniques were used not only to distinguish the different chromosomes, but also to enable discrimination of different sub-chromosomal regions (Bernardi, 1993). Although the resolution of this map is coarse (an average sized chromosome band in a 550-band preparation contains ~6 Mb of DNA), it has been very useful as a framework for ordering the locations of human DNA sequences by chromosome in situ hybridisation techniques (Trask, 1991; van Ommen et al., 1995).

Other maps have been obtained by mapping natural chromosome breakpoints, using panels of somatic cell hybrids containing fragments derived from translocation and deletion chromosomes (Abrams et al., 1995), or by mapping artificial breakpoints using radiation hybrids (RH) (Walter and Goodfellow, 1993). However, the resolution achieved by such hybrid cell panels, that is the average distance between neighbouring breakpoints can be limited for large parts of the genome. As a result, higher resolution physical maps are desirable. Clearly, the physical map which will provide the highest
possible resolution, that of single base pairs, is the ultimate map: the complete nucleotide sequence of the genome. As this will not be achieved for some time, attention has been focused on constructing physical maps of intermediate resolution. Comprehensive RH maps and rare-cutter restriction maps have been achieved thus far for only a few human chromosomes. One example is chromosome 21 where a NotI restriction map has been published for the entire long arm (Ichikawa et al., 1993). In addition, much of the current mapping effort is aimed at mapping of coding DNA sequences, thereby producing comprehensive transcription maps.

A major intermediate goal of the Human Genome Project is to construct a complete contig map of the DNA of each of the 24 different types of human chromosome. This means relating different DNA clones to define a series of partially overlapping DNA molecules covering the entire length of a chromosome. Identification of overlaps between the DNA segments of different clones can be achieved by a variety of different procedures such as: 1) chromosome walking which establishes clone contigs from fixed starting points (Anand et al., 1991; Little et al., 1992), 2) repetitive DNA fingerprinting, involving characterising each clone in terms of the pattern of restriction fragments detected by two human repetitive sequence probes. For example, a whole genome clone fingerprinting approach has been applied to mapping the human genome largely on the basis of repetitive DNA fingerprinting of YACs (Bellanne-Chantelot et al., 1992). Thirdly, sequence-tagged site (STS) content mapping (Green and Olson, 1990; Cole et al., 1992) has been used, which involves PCR-based screening with genetically mapped microsatellite markers; YACs identified as containing such markers were referred to as 'genetically anchored YACs'. In this multilevel mapping approach physical map construction is directly based on integration with the genetic map.

Given the large size of mammalian genomes, physical mapping of the entire human genome requires the use of clones with large DNA inserts, of the order of 1 megabase (Mb). Yeast artificial chromosomes (YACs) are currently the only cloning system capable of propagating such large DNA fragments and hence have been particularly useful in generating such contigs. However YACs suffer from high rates of chimerism and rearrangement (Larionov et al., 1994) and thus are unsuitable for genomic sequencing (Chumakov et al., 1995). STS-based maps sidestep this problem by having a sufficiently high density of landmarks so that one can rapidly regenerate physical coverage of any region by PCR-based screening of clones appropriate for sequencing - such as cosmids, bacteriophage P1 clones, bacterial artificial chromosomes (BACs) and P1-derived artificial chromosomes (PACs) (reviewed in Monaco and Larin, 1994). Ultimately, high resolution maps based on cosmid contigs will provide a suitable framework for sequencing whole chromosomes. Significant contig maps for individual human chromosomes were first obtained for chromosome 21 (Chumakov et al., 1992) and the Y chromosome (Foote et al., 1992). Reasonably comprehensive YAC contig maps have
been published for chromosomes 3 and 22 (Gemmill et al., 1995; Collins et al., 1995) and also integrated maps of chromosomes 16 and 19 which include high resolution cosmid contigs (Little, 1995).

The first generation physical map of the human genome was constructed by exhaustive screening of the CEPH YAC library which contains 33,000 YACs with an average insert size of 0.9 Mb, representing 10 haploid genome equivalents (Cohen et al., 1993). Overlaps between YAC clones were identified using three methods: repetitive DNA fingerprinting, STS content mapping and Alu-PCR probe hybridisation (Nelson et al., 1989). Whilst this physical map was far from complete, with poor coverage of some chromosomes, it provided a framework for the scientific community to build upon, in order to produce maps of all the chromosomes. This detailed mapping information which was made widely available by electronic access through the Internet, was used by various researchers to relate to specific chromosomes, or often sub-chromosomal regions that were of interest. An updated YAC contig map has since been published, covering about 75% of the human genome and consisting of 225 contigs with an average size of 10 Mb (Chumakov et al., 1995).

By providing a common language for physical mapping projects, the use of STSs allowed incorporation of any type of physical mapping data into the evolving map. A physical map of the human genome has now been constructed based on 15,086 sequence-tagged sites (STSs), with an average spacing of 199 kilobases (Hudson et al., 1995). This involved assembly of a radiation hybrid map of the human genome comprising 6,193 loci and incorporated a genetic linkage map of the human genome comprising 5,264 loci. This combined with the results of STS-content screening of 10,850 loci against a yeast artificial chromosome library produced an integrated map, anchored by the radiation hybrid and genetic maps. This map represented an early step in an international project to generate a transcript map of the human genome, with more than 3,235 expressed sequences localised. Recently a map of 30,181 human gene-based markers was assembled and integrated with the current genetic map (Schuler et al., 1996) by radiation hybrid mapping. This new gene map (Deloukas et al., 1998) consisted of data from 41,664 STS's based on 3' untranslated regions of cDNAs. It contained nearly twice as many genes as the previous release including most genes that encode proteins of known function.
1. Introduction

1.2.3 Transcriptional maps of the human genome

The recent focus of the Human Genome Project has undoubtedly shifted towards characterization of the coding DNA component and the development of a human gene map. Large scale cDNA sequencing can be carried out to collect fragmentary sequences from which mapping markers can be derived, and for pinpointing the location of the gene in the genomic sequence. Such fragmentary sequences are known as expressed sequence tags or ESTs; an STS which is specific for a coding DNA (Adams et al., 1991). The development of ESTs from the 3' untranslated regions (UTR) of cDNAs supplies both a marker and a gene for the map. Using the 3' UTR has two advantages: firstly, 3' UTRs rarely contain introns so the PCR product size should be the same in cDNA and genomic DNA, and secondly, the 3' UTR is more gene and species specific than the 5' sequence and allows for members of conserved gene families to be distinguished. By extending this approach and employing a strategy that enabled regional gene mapping on a genome-wide scale James Sikela and coworkers reported (Berry et al., 1995), the mapping of 318 cDNAs using the CEPH mega-YAC library. Recently 13,600 ESTs have been placed on the human physical map (Hillier et al., 1996). Shortly, most human genes will be sequence-tagged and placed on various physical maps. Such a 'transcript map' of the genome will be an important part of the sequencing infrastructure, as well as a critical resource for the positional candidate approach to disease gene cloning.

The screening and usage of cDNA libraries has been facilitated by using gridded arrays of clones in much the same way as for genomic clones. The IMAGE consortium (integrated molecular analysis of gene expression) has particularly fostered this development, and master arrays of clones representing unique genes have been prepared. Not all genes will be represented by ESTs, however, because some genes will be very poorly expressed or not expressed at all in the cDNA libraries analysed, including genes which are expressed at certain developmental stages. It has been expected that the EST program (Boguski and Schuler, 1995) should be able to identify sequences from about 80% of all genes. The remainder will be identified as coding DNA sequences in genomic clones. Most ESTs are not polymorphic but can be assigned to chromosomal and sub-chromosomal locations by physical mapping techniques. One approach is to map ESTs by PCR assays of panels of whole genome radiation hybrids (Boguski and Schuler, 1995). In addition to cDNA/EST mapping, the construction of comprehensive transcription maps of human chromosomes may also involve mapping of exons trapped from cosmid and YAC clones (Yaspo et al., 1995).

Finally, the transcript map will be an important resource for genomic sequencing. To sequence the genome of the nematode, C. elegans (Wilson et al., 1994), the strategy was to elect the gene-dense regions (defined genetically) as initial targets for sequencing;
the availability of a human transcript map will provide data on gene density and make it possible to use a similar strategy for sequencing human chromosomes. Another lesson drawn from the nematode project is that even with a high gene density (one gene per 5 kb), and the presence of many fewer dispersed repeats compared with mammals, ESTs still proved invaluable for estimating the total number of genes and for finding the exons by alignment with genomic sequences. Even before the human 'sequence map' is densely populated, the transcript map will shed new light on global aspects of gene organization, evolution and expression.

1.2.4 Genome maps of model organisms

At the onset of the Human Genome Project, it was recognised that comprehensive maps of certain model organisms would be highly desirable. Such organisms include a variety of species, some of which have been particularly amenable to genetic analysis. Research involving such models will continue to provide a basis for analysing normal gene regulation, genetic diseases and evolutionary processes. Genome sequencing of *E. coli* (Blattner *et al.*, 1997, Koonin, 1997) and *S. cerevisiae* (Johnston, 1996), two of the five model organisms targeted in the initial 5-year program has been completed. Availability of these sequences has led to the discovery of many new genes and other functional elements of the genome. Completion of the DNA sequence of the remaining model organisms, *C. elegans* (Wilson *et al.*, 1994, Berks *et al.*, 1995), *D. melanogaster* (Rubin, 1996) and mouse (Dietrich, *et al.*, 1995), continues to be a high priority.

1.2.5 Sequencing the human genome

Physical mapping has yielded a low resolution map of the complete human genome (Chumakov *et al.*, 1995). There are regions where the density of sequence tagged sites (STSs) and ordered cosmid clones are sufficient to approximate a 'sequence ready' map. When combined with the abundant partial cDNA sequences that have been generated by the enormously successful EST sequencing programmes (Adams *et al.*, 1995), and the fact that some of these cDNAs are now mapped, the foundation is in place for the final, difficult phase of genomic sequencing.

No revolutionary sequencing techniques have come to the fore in the past five years. Instead substantial amounts of DNA sequence have been acquired by a range of conventional methods. These data have addressed many questions concerning the practicality of large-scale sequencing strategies and have identified important factors related to sequence accuracy, the mechanics of manipulating large numbers of DNA clones, and the resource requirements necessary for handling extensive sequence information.
1.2.6 Data storage and access in the Human Genome Project

To be useful, the enormous amount of information being generated by the genome projects must be readily accessible and in a form that is easily analysed. Electronic databases are needed to serve this function. Currently, the rate of sequencing is such that the number of entries in the major databases is doubling in size every 18 months. As the amount of data rises exponentially, so databases are having to be modified to maintain simple and efficient data. In addition to data input, sequence databases also offer suites of computer programs for users to analyse sequences.

The central data resource for the human gene mapping effort, is the Genome Data Base (GDB, http://www.gdb.org). GDB is regularly updated (Fasman et al., 1996, 1997; Letovsky et al., 1998). It collects, organises, stores and distributes human genome mapping information, and also serves as a repository for genetic disease information. It also offers interaction with other databases, such as On-line Mendelian Inheritance in Man (OMIM). There has recently been a rapid rise in the data accumulation in the genome projects. Some of the major laboratories such as the Co-operative Human Linkage Center (CHLC) and the Genethon/CEPH laboratories, which generate mapping data have established their own file transfer protocol (ftp) servers, in order to provide the scientific community with ready access to recently produced data. Recently a distributed and integrated Human Genome Map database called HuGeMap (http://www.infobiogen.fr/services/Hugemap) has been established (Barillot et al., 1998). The HuGeMap database stores the major genetic and physical maps of the human genome. It is also interconnected with the gene radiation hybrid mapping database RHdb (http://www.ebi.ac.uk/RHdb). A variety of other databases exist for the model organisms too, such as the Mouse Genome Database (MGD, http://www.informatics.jax.org/). MGD stores data on mouse loci and probes and is integrated with the Encyclopedia of the Mouse Genome, an application that generates a graphical display of mouse genetic linkage maps using MGD locus and homology information. Such databases are useful to human gene mappers, as they provide human-mouse comparative mapping data. An important application of comparative mapping is the ability to predict the existence and location of unknown human homologs of known, mapped mouse genes.
1.3 Genetic Mapping

1.3.1 Meiotic recombination

The genetic material in humans is packaged into 23 pairs of chromosomes. The process of gametogenesis renders the cell to a haploid state, when the number of chromosomes is halved and each germ cell receives one member of each pair of homologous chromosomes. During the early stages of this meiotic cell division, when the chromosomes are aligned, an exchange of genetic material occurs between the chromatids of the homologous chromosomes. This process is known as recombination or crossing over. It follows that recombination events are more likely to occur between loci which are situated far apart on an individual chromosome. Two different loci which segregate together, thus deviating from Mendel's law of independent assortment, must lie in close physical proximity on a chromosome, and are said to be genetically linked.

1.3.2 Principles of linkage analysis

The aim of linkage analysis is to determine the recombination frequency between two loci, which may be either two genetic markers or a genetic marker and a disease locus. The extent of linkage between loci is measured as the recombination fraction or \( \theta \), which represents the number of recombination events observed as a fraction of the total number of opportunities for recombination. The maximum recombination fraction between any two loci is 0.5 or 50% which corresponds to unlinked loci which segregate independently. Where two markers lie close together on a chromosome, the number of crossovers will be minimal, and the \( \theta \) value will be less than 0.5, and approaches 0 for tightly linked loci. Linkage analysis in human pedigrees utilises polymorphic genetic markers to ascertain whether a given marker segregates with the disease phenotype in a family. Linkage is established by determining the likelihood of linkage between disease locus and marker locus. This requires elaborate methods because families are small and pedigree structures rarely ideal. The method most commonly used to calculate this likelihood is the method of lod scores (Morton, 1955).

1.3.3 Calculation of lod scores

The lod score method is based on the likelihood (L) of an observation of recombination at a given recombination fraction or \( \theta \). The maximum likelihood estimate between two loci is obtained by determining the value of \( \theta \) which gives the maximum probability for the observed data. The likelihood of linkage is expressed quantitatively as
the odds ratio, such as $L(\theta=0.5)/L(\theta=0.5)$, which denotes the probability of linkage versus no linkage. This ratio is calculated at a series of $\theta$ values, and is expressed as the logarithm of the odds: the decimal logarithm of the odds ratio is termed the lod score ($Z$). The lod score at a given recombination fraction is therefore calculated:

$$Z = \log_{10} \left[ \frac{L(\theta)}{L(0.5)} \right]$$

The $\theta$ value which gives the highest lod score obtained ($Z_{\text{max}}$), provides the best estimate of linkage between two loci. A lod score of +3 is considered to be strong evidence for linkage, as the odds ratio is 1000:1 in favour of linkage with 95% confidence. A lod score of +2, corresponding to a 100:1 chance of linkage is considered suggestive of linkage for autosomal loci, but is acceptable evidence for linkage on the X chromosome, where there is prior probability that two loci will show linkage if they are already known to be on the X chromosome. A lod score of -2 or less is taken as evidence for exclusion of linkage within an interval equal to the corresponding $\theta$ value from both sides of the marker locus. The use of lod scores to determine linkage between two loci is known as two point analysis.

Once a gene or disease has been roughly localised, it becomes important to fix its exact position relative to a framework or markers from the vicinity. Multipoint linkage analysis allows for calculation of linkage at more than two genetic loci at a time, utilising the computer program LESTK AGE (Lathrop et al., 1984). The overall likelihood of the data is analysed for a series of possible positions of an unknown locus relative to a fixed framework of marker loci. The likelihood of the data with the unknown locus at a particular position is compared to the likelihood with the unknown locus away from any of the markers. The logarithm of this likelihood ratio generates a location score, which can then be converted into a conventional lod score. Multipoint linkage is particularly useful where some of the markers for a given region are partially uninformative in a pedigree, as it provides a cumulative result for all markers analysed.

Many disorders are caused by defects at different loci, a situation known as locus heterogeneity. Disease locus heterogeneity often can be detected by linkage studies. On inspection of lod scores for individual families, if a mixture of linked and non-linked families is found, it is evidence for heterogeneity and formal tests of heterogeneity should then be applied. The HOMOG software package (Ott, 1992), which employs admixture analysis is commonly used for this purpose.

### 1.3.4 Genetic map distance

Distances between genetic loci are measured in map units or centimorgans (cM). Two loci which displays 1% recombination, that is recombine once in every 100 meioses, are defined as being genetically separated by 1 cM. Over short chromosomal distances
θ is directly proportional to genetic map distance so this relationship holds true. However, over longer chromosomal distances the linear relationship breaks down as multiple recombination events can occur between the two loci. There is also the effect of interference in crossing over, whereby one crossover in a given region will reduce the chance of a second recombination event in the immediate vicinity (where θ<0.10). Other factors, such as the unequal recombination frequencies observed in males and females, with higher recombination rates in females, and the presence of recombination hot-spots along the length of individual chromosomes together mean that the recombination fraction is not additive over larger distances. To overcome this problem several mapping functions have been devised to transform θ into additive map distance. The Kosambi map function is preferentially used as it allows for the strongest level of interference, and therefore produces more realistic map distances.

1.3.5 Genetic markers

Genetic markers can originate from either expressed sequences or non-coding DNA, but to be useful the marker must be polymorphic, that is exist in two or more forms in the general population. A measure of genetic variation is the amount of heterozygosity at a locus in a population, which is given by the total frequency of heterozygotes at a locus. The more polymorphic a marker locus is, the higher will be the frequency of heterozygotes at the locus, and therefore, the more useful the locus will be for linkage analysis and for genetic screening. Most human genetic mapping uses markers. There are two general approaches, firstly disease-marker mapping is used for locating disease genes, and secondly marker-marker mapping is used to construct framework maps. The first generation of DNA markers were restriction fragment length polymorphisms (RFLPs) (Botstein et al., 1980), owing their existence to restriction enzyme site polymorphisms, which have only two alleles: the site being present or absent. Their informativeness as markers are hence limited. RFLPs were initially typed by hybridising Southern blots of restriction digests with radiolabelled probes and more recently by PCR. Much higher levels of heterozygosity are associated with polymorphism that occurs in repeated sequences known as minisatelitites (Jeffreys et al., 1985). These polymorphisms also known as VNTRs or Variable Number of Tandem Repeats (Nakamura et al., 1987). These hypervariable loci consist of repeated sequences that are usually 10-60 bp long; the number of repeats varies from less than ten to several dozen. Minisatellites are difficult to analyse by standard PCR protocols because large alleles may fail to amplify, and they also tend to be found clustered in subtelomeric regions of chromosomes. This limits the usefulness of minisatellites as genetic markers.

There is another class of tandem repeats known as microsatellites. Microsatellites consist of around 10-50 copies of a 1 to 6 bp motif that can occur in perfect tandem
repetition, as imperfect repeats or together with another repeat type (Weber, 1990). Dinucleotide repeats are often polymorphic (Weber, 1990), with the number of alleles varying. The most common dinucleotide repeat unit is the (CA)$_n$ repeat. There are about 50,000-100,000 (CA)$_n$ sequences in the human genome or approximately 1 every 30-60 kb indicating the wide distribution of these repeats (Stallings et al., 1991). A direct relationship between the length of microsatellite motifs and the degree of polymorphism has been demonstrated (Weber, 1990). Many of these microsatellites are therefore highly informative for linkage analysis making high precision mapping possible and since they are abundant, provide an important source of marker loci. They can be readily detected by PCR (Weber and May, 1989). However, dinucleotide repeat sequences are prone to replication slippage during PCR amplification (Levinson and Gutman, 1987), so that each allele gives a ladder of 'stutter bands' on a gel (Hauge and Litt, 1993). In addition to the dinucleotide repeats, highly polymorphic tri- and tetra-nucleotide repeats have been isolated and characterised (Edwards et al., 1991). These repeats give clearer results with a single band from each allele and are gradually replacing dinucleotide repeats as the markers of choice.

1.3.6 Cross-over (haplotype) analysis

When a disease-associated linkage has been confirmed in a pedigree, it is important to examine individual cross-overs (recombinant individuals) in the family data in order to define the closest distal and proximal marker that flank the disease locus. Haplotypes of the chromosome marker loci are constructed for individuals based on established map order of markers. A genetic haplotype therefore refers to the combination of alleles observed at a linked locus. Such data are used to determine the minimal region of interest. As new markers become available they are mapped to the region of interest with respect to existing markers. The critical region of a disease gene can then potentially be further refined.

1.4 Physical Mapping

A wide variety of physical mapping strategies have been used to analyse the DNA of complex eukaryotic genomes. These can be classed into low resolution physical mapping, where the smallest map unit that can be resolved is typically one to several megabases of DNA and high resolution physical mapping, where the resolution is typically very high, from hundred of kilobases to a single nucleotide. The advent of somatic cell hybridisation gave a major impetus to human gene mapping and the more recent molecular techniques using restriction enzymes, cloned DNA probes, and high-resolution in situ hybridisation have greatly accelerated the process.
1.4.1 Low Resolution Physical Mapping

1.4.1.1 Somatic cell hybrids

Somatic cell hybrid panels can permit chromosomal localisation of any human DNA sequence. Under certain experimental conditions, cultured cells from different species can be induced to fuse together, thereby generating somatic cell hybrids. In human genetic mapping, hybrid cells are typically constructed by fusing human cells and rodent (usually mouse or hamster) cells. The initial fusion products described as heterokaryons, proceed to mitosis, resulting in dissolution of the two nuclear envelopes. The human and rodent nucleus are then brought together in a single nucleus. The hybrid cells are unstable initially; for reasons that remain unknown, most human chromosomes fail to replicate in subsequent rounds of cell division and are lost. This gives rise eventually to a variety of more or less stable hybrid cell lines, each with a full set of rodent chromosomes plus a few human chromosomes. The loss of the human chromosomes occurs essentially at random but can be controlled by selection.

The human chromosomes in somatic cell hybrids can be identified by PCR screening with sets of chromosome-specific primers (Abbott and Povey, 1991). By collecting hybrid cell lines with different sets of human chromosomes, it is possible to generate a hybrid cell panel that can be used to map any human DNA sequence to a specific chromosome. Localisation of a human DNA sequence to a single chromosome can be inferred by deduction: the chromosome must be common to all cell lines which are positive for the test and absent from all cell lines which are negative for the test, that is all hybrids should be concordant (Faraco et al., 1995). A disadvantage of traditional somatic cell hybrids is that the hybrid cells generally contain several human chromosomes rather than just a single human chromosome. In order to limit the amount of human genetic material transferred to a recipient rodent cell, the technique of microcell fusion (Fournier and Ruddle, 1977) can be applied. Microcell hybrids produced in this manner contain a few donor chromosomes, but the simplest contain a single donor chromosome (monochromosomal hybrids) (Warburton et al., 1990). Chromosomal localisation of human DNA clones can be established directly and rapidly using such monochromosomal hybrid panels which provide unambiguous evidence for the presence or absence of a given marker on a specified chromosome.

In order to obtain sub-chromosomal localisations, specialised hybrids are required which contain only part of a given human chromosome. The sub-chromosomal fragments may result from spontaneous chromosome breakage as a result of translocations or deletions, or they may be artificially induced. Cell lines are made from hybrids which contain the abnormal chromosome but which lack the normal homologue.
of the chromosome of interest (translocation hybrids and deletion hybrids) (Abrams et al., 1995). For this approach to be useful for a given chromosome, several different natural breakpoints must be available for that chromosome, a condition which may not always be met. One technique which enables genome-wide mapping involves the artificial breaking of human chromosomes and transfer of the sub-chromosomal fragments into rodent cells. Chromosome-mediated gene transfer (CMGT), is a procedure in which fragments of purified chromosomes are transferred into recipient cells in the presence of calcium phosphate (Porteous, 1987). Hybrids established by this method retain sub-chromosomal segments of human DNA (transgenomes) of a size that is useful for mapping (usually in the range of 1-50 Mb).

Irradiation fusion gene transfer is another method by which chromosome fragments are generated artificially, in this case by lethal irradiation of donor cells which are rescued by fusion with suitable recipient cells (Walter and Goodfellow, 1993). This method has been popular for use on somatic cell hybrids containing a single human chromosome. When DNA samples from a panel of such radiation hybrids are screened by hybridisation against a series of DNA clones, or by corresponding PCR assays, the patterns of cross-reactivity can be interpreted statistically to produce a linear map order for the DNA clones (Cox et al., 1990). A variant form of radiation hybrid mapping involves the use of human fibroblast cells as the starting donor cells instead of a monochromosomal somatic cell hybrid. This has the attraction of enabling construction of reasonably high resolution maps of the entire genome with a single panel of 100-200 hybrids (Walter et al., 1994).

1.4.1.2 In situ hybridisation

This is the most direct method of mapping. A cloned DNA fragment is hybridised directly to a spread of metaphase chromosomes, and the map location worked out by examining the result under the microscope (Gerhard et al., 1981). The sensitivity and resolution of in situ hybridisation has been increased significantly by the development of fluorescence in situ hybridisation (FISH) (Trask, 1991; van Ommen et al., 1995). In this technique, a DNA probe is labelled by incorporation of modified nucleotides, obtained by covalent binding of a reporter molecule such as biotin or digoxigenin, which can be detected by specific binding to another molecule such as modified nucleotides. FISH has the advantage of providing rapid results which can be scored conveniently by eye using a fluorescence microscope.

A special application of FISH has been the use of DNA probes where the starting DNA is composed of a large collection of different DNA fragments from a single type of chromosome. Such probes can be prepared by combining all human DNA inserts in a chromosome-specific DNA library (Davies et al., 1981). The resulting hybridisation signal
1. Introduction

represents the combined contributions of many loci spanning a whole chromosome and causes the whole chromosome to fluoresce (chromosome painting). A few differently coloured fluorescence labels (chromosome paints) can be used in different ratios to provide numerous different colours for labelling chromosomes, thereby providing a molecular karyotype (Dauwerse et al., 1992). The technique of chromosome painting has been extended recently by the ability to paint sub-chromosomal regions, using a mixed DNA probe corresponding to a particular sub-chromosomal region, as obtained from chromosome microdissection DNA libraries (Ludecke et al., 1989). Chromosome painting has found increasing applications in defining de novo rearrangements and marker chromosomes in clinical and cancer cytogenetics.

1.4.2 High Resolution Physical Mapping

1.4.2.1 Pulsed field gel electrophoresis (PFGE)

Restriction mapping permits molecular mapping with resolutions which depend on the frequency of the enzyme recognition site. Most restriction enzymes which recognise a 4- or 6-bp sequence typically cut vertebrate DNA once every few hundred or few thousand base pairs. The recognition sequences for rare-cutter restriction nucleases are typically 6-8 bp long and contain one or more CpG dinucleotides which are rare in vertebrate DNA (Bird, 1986). As a result, they generate fragments that are typically several kilobases in length. Whilst small restriction fragments can be size fractionated conveniently by agarose gel electrophoresis, large DNA fragments will remain unresolved. In order to separate such large restriction fragments, pulsed-field gel electrophoresis (PFGE) (Schwartz and Cantor, 1984), a modified form of agarose gel electrophoresis, is used. Large DNA fragments are separated in PFGE because the DNA molecules are subjected alternately to two approximately perpendicular fields. The presence of a discontinuous electric field means that the DNA molecules are intermittently forced to change their conformation and direction of migration during their passage through the gel. The time taken for a DNA molecule to alter its conformation and re-orientate itself in the direction of the new electric field is strictly size dependent. As a result DNA fragments up to several megabases in size can be fractionated efficiently. The resolution depends on the time of switching between the two fields (pulse time). Longer pulse times will resolve DNA fragments in the higher molecular weight range. Modifications of the original PFGE method have been developed which result in the DNA running straight in the tracks thus improving the resolution. The field is generated by the use of contour-clamped electric fields and the apparatus consists of an hexagonal array of electrodes to create a more homogeneous
field (Chu et al., 1986). The resolving power can then be optimised for fragments between 1 and 10 Mb.

A long-range restriction map of a chromosomal region can be generated by the same principle as conventional restriction mapping (Barlow and Lehrach, 1987). Physical linkage between two markers can be established if at least two restriction fragments are identical in size and a detailed map can be built up from information on single and double digests. Partial digests which are able to generate maps over much larger regions and to link more widely-spaced DNA probe have also been used in PFGE analyses. PFGE can also be used to identify candidate genes and assist in their isolation if the disease is caused by DNA deletion in some patients. The detection of deletions by PFGE has been useful in the diagnosis of carriers in Duchenne Muscular Dystrophy (DMD) (van Ommen et al., 1987).

1.4.2.2 Assembly of clone contigs

In order to provide a framework for the ultimate physical map a series of cloned DNA fragments need to be assembled which collectively provide full representation of the sequence of interest. For a complete representation, with no gaps, the series of clones should contain overlapping inserts forming a comprehensive clone contig. In principle, contig assembly is facilitated by the way in which genomic DNA libraries are constructed: as part of the strategy for maximising the representation of a library, the genomic DNA is deliberately subjected to partial digestion with a restriction endonuclease. As a result, individual genomic DNA clones usually contain DNA sequences that partially overlap with those found in at least some other clones in the library. In order to identify clones with overlapping inserts, a specific DNA probe from one clone is used to screen a DNA library. With genomic DNA libraries, this permits the assembly of a clone contig by bidirectional chromosome walking from a fixed starting point.

Using arrayed libraries of YAC clones representing multiple copies of the genome, individual clones are isolated either by screening gridded arrays by hybridisation with single-copy probes (Nizetic et al., 1991; Bentley et al., 1992), or by testing pools of clones in multiple tiers or multiple arrays for the presence or absence of a sequence tagged site (STS) using a PCR assay. In early studies using these approaches, YACs were isolated and overlapped on the basis of their shared STS or probe content, and contiguous regions of several megabases were obtained (Anand et al., 1991; Monaco et al., 1992). Where landmarks did not detect overlaps between clones directly, methods to isolate new landmarks from the ends of YAC were developed to facilitate chromosome walking in YACs. Frequently used methods include inverse-PCR (Ochman et al., 1988) and vectorette PCR (Riley et al., 1990). A variety of clone fingerprinting
1. Introduction

Techniques have been used to type clones at random and then integrate the information in order to identify clone contigs over large regions of a genome (Bellanne-Chantelot et al., 1992). This approach has been used to great effect in assembling YAC contigs over significant amounts of the human genome (Cohen et al., 1993). Sequence-tagged site content mapping has also been used effectively to identify YACs with overlapping sequences (Cole et al., 1992; section 1.2.2).

1.5 Identification of gene sequences from a defined genomic interval

To progress from chromosomal localisation of a disease locus to gene identification requires the identification and characterisation of disease genes. Definition of the map location of a disease locus may suggest previously characterised genes as candidates for disease-causing mutations. Alternatively, positional cloning techniques are often required to identify novel genes that can be subsequently evaluated for disease causing mutations. A primary goal for physical mapping is identifying the locations of genes within a clone contig that has been generated at a specific chromosomal region. In principle, two major features permit the DNA of genes to be distinguished from DNA that does not have a coding function.

1) Expression: all active genes are capable of making an RNA product, which in the vast majority of cases is mRNA. Mammalian genes usually contain exons and so the initial RNA transcript usually needs to undergo splicing. 2) Sequence conservation: the sequence of coding DNA and important regulatory sequences are more strongly conserved in evolution than that of non-coding DNA. In addition premature termination codons in coding DNA are selected against. In addition, vertebrate genes are often associated with CpG islands (Cross and Bird, 1995). These features have permitted a variety of different methods for identifying genes in cloned vertebrate DNA of which the most commonly used are described below.

1.5.1 Hybridisation to RNA/cDNA

A candidate DNA clone can be hybridised against a Northern blot containing a panel of mRNA or total RNA samples isolated from a variety of different tissues. Positive hybridisation may indicate the presence of a gene within the cloned fragment and may suggest a suitable cDNA library for screening. They may fail, however, for two reasons. First, significant expression of the gene may be restricted to a cell population or developmental stage which is not represented in the Northern blot panel or those cDNA libraries which are selected for screening. If the gene is not strongly expressed in the
relevant tissue (so that it is not well represented in the RNA samples or cDNA libraries which are being screened), detection of positive hybridisation signals may be difficult. Occasional transcribed repeats may also be problematic.

1.5.2 Zoo blot hybridisation

Coding DNA sequences are subject to considerable selection pressure to conserve biologically important sequences. By contrast, noncoding DNA sequences accumulate mutations comparatively rapidly and are not well conserved between species. A zoo blot is a Southern blot of genomic DNA samples from a wide variety of different species. Genomic clones can be used as hybridisation probes onto such zoo blots as was done for the isolation of the Duchenne Muscular Dystrophy (DMD) gene (Monaco et al., 1986). In this report a clone pERT 87-25 from human Xp21 hybridised to all tested zoo blot DNAs indicating that it may be from a portion of a gene that encodes a conserved amino acid sequence. Conserved sequences, which are probably protein coding sequences, hybridise at a higher stringency with genomic DNAs of other mammalian species than do less conserved, putative non-coding sequences. Some mammalian genes, often with a crucially important function in development, are so highly conserved that they will show significant hybridisation signals with evolutionarily distant species such as yeast, *Drosophila* and *Caenorhabditis elegans*. Others may only show significant hybridisation signals to mammals (Claudio et al., 1994).

1.5.3 CpG island identification

CpG islands (sometimes known as HTF islands; Cross and Bird, 1995) are short (~1 kb long) hypomethylated GC-rich sequences which are often found at the 5' ends of vertebrate genes. In the human genome, an estimated 56% of genes are associated with such sequences (Antequera and Bird, 1993). They include all examples of housekeeping genes and genes that are widely expressed, and a significant portion, perhaps about 40% of genes which show tissue-specific or restricted expression patterns (Larsen et al., 1992). Rare-cutting enzymes which recognise GC rich sequences can be used to identify potential CpG islands (Bird, 1986). The utility of the HTF island for genome mapping and the identification of novel genes was demonstrated by the characterisation of the MHC Class II locus (Sargent et al., 1989), where of the 20 new transcripts described, 18 were identified through their association with HTF islands. There are two drawbacks with this approach: first, identifying and subcloning the HTF islands can be laborious, and second, depending on the source of the cDNA library, the gene may or may not be represented. Also the method is not applicable to the substantial number of genes which have no associated CpG islands.
Island rescue PCR (IRP) can permit selective amplification of CpG island sequences from human YACs. The method depends on the high copy number of Alu repeats (so that there is a high chance of an Alu repeat in the vicinity of a CpG island) and the frequent occurrence of restriction sites for rare-cutters such as BssHII in CpG islands. The YAC DNA is cut with a suitable rare-cutter restriction nuclease and fragments are ligated to a bubble-linker primer with a suitably complementary overhang. CpG island-Alu PCR is then possible using an Alu-specific primer and a bubble-linker-specific primer (Valdes et al., 1994).

1.5.4 Direct screening of cDNA libraries

Genomic clones can be used directly as hybridisation probes on mRNA or cDNA libraries. Genomic fragments cloned into lambda vectors, PAC and cosmid vectors have been used successfully as hybridisation probes (Kendall et al., 1990) on a cDNA library using a pre-annealing procedure to prevent cross hybridisation with repetitive DNA sequences. Larger clones such as YACs occasionally can be used as hybridisation probes (Elvin et al., 1990), but are generally less amenable to such approaches due to very poor signal-to-noise ratios because of the high sequence complexity of the YAC and the large amount of repeat blocking that is necessary to avoid detecting spurious repeat-containing cDNAs.

This technique allows the possibility of identifying the maximum number of coding regions, including non CpG-island associated genes, with the minimum number of probings of the cDNA library. Once isolated, the cDNA inserts can be mapped back to the relevant cosmid, PAC or YAC inserts to define the location of the corresponding gene. This method requires the region of interest to be available as a collection of arrayed cloned genomic fragments, hence being applicable in analysing large arrayed genomic libraries into contiguous segments (Craig et al., 1990). It can also integrate data from many tissues. It is not limited by the structure of the gene or by expression patterns. By direct screening of a human duodenal mucosa cDNA library, seven new genes within the HLA-A region were localised (Kahloun et al., 1992). Also cDNAs were isolated from a retinal cDNA library using a YAC probe from the OATL1 region at Xp11.2 (Geraghty et al., 1993). Recently novel cDNAs were isolated from the Cri-du-chat critical region by direct screening of a chromosome 5-specific cDNA library (Simmons et al., 1995). The limits of the technique are that deriving coding sequences from the genomic clone can be time consuming and the problem of expression level bias where sensitivity to detecting genes whose transcripts are present at less than 0.01% is uncertain.
1.5.5 Exon trapping/exon amplification technique

The gene identification methods described above have their limitations: the restricted expression patterns of some genes may make them difficult to identify; genomic DNA clones may give very weak hybridisation signals if the percentage of exon signals is very low and many genes which are expressed in a tissue-specific or restricted manner do not have associated CpG islands. An alternative is to identify a gene by the ability of its exons to engage in an artificial RNA splicing assay. RNA splicing involves fusion of exonic sequences at the RNA level and excision of intronic sequences. Spliceosomes are able to accomplish this in vivo by recognising certain sequences at exon/intron boundaries. A cosmid or other suitable genomic DNA clone containing an internal exon flanked by intronic sequences will contain functional splice donor and acceptor sequences.

Exons can be identified in cloned genomic DNA by subcloning the DNA into a suitable expression vector and transfecting into an appropriate eukaryotic cell line in which the insert DNA is transcribed into RNA and the RNA transcript undergoes RNA splicing. Such techniques are known as exon trapping (often called exon amplification if a PCR reaction is employed to recover the exons from a cDNA copy of the spliced RNA). In one application of this principle the vector provides the splice donor site and the genomic insert is tested for the presence of a splice acceptor site (Duyk et al., 1990). A drawback of this method is the relatively high frequency of cryptic splice acceptor sites, necessitating cumbersome re-screening of candidates to identify true exons. Two other vector systems reduce this problem by requiring, in one case, both splice donor and acceptor sites from the putative exon (Buckler et al., 1991) or, in the other case, the interaction of splice donor and polyadenylation signals from 3' terminal exons using the specific exon trapping system, SETS (Hamaguchi et al., 1992). In these latter two systems, spliced mRNA populations are expanded (exon amplification) by a combined reverse transcription and polymerase chain reaction (RT-PCR) technique and can be cloned to provide exon libraries. The exon amplification system (Buckler et al., 1991) utilises the HIV-tat intron to replace an intron from the rabbit β globin gene, in the vector pSPL1, while the SETS system utilises the p53 gene intron 10 and its flanking exons, selected for accuracy of splicing and ease of recognition of the spliced products, in the pMHC2 vector. Modifications to the original pSPL1 vector have been described (Church et al., 1994) that increase both the sensitivity and efficiency of exon amplification and its ability to isolate expressed sequences from highly complex sources of genomic DNA. The new version, pSPL3, has been modified to eliminate vector only splice products and the false positives resulting from cryptic splicing in the HIV-tat intron.
This technique has been important in the identification of the Menkes disease gene (Vulpe et al., 1993), the neurofibromatosis type 2 tumour suppressor gene (Trofatter et al., 1993), the glycerol kinase gene (Walker et al., 1993), and the Huntington's disease gene (The HD Collaborative Research Group, 1993). Recently the cloning of 559 potential exons of genes of human chromosome 21 was identified by exon trapping (Chen et al., 1996). The exon trap/amplification system is a rapid, efficient and sensitive method to identify transcribed sequences and allows an easy transition from genomic to cDNA clones. The method has no biases on gene structure, apart from the requirement for an intron and the requirement that the splice specificity be compatible. There is also no apparent dependence on the level of expression or on the tissue, hence it is useful in identifying candidates for disease genes when the tissue in which the gene is expressed is uncertain. The problems associated with this method are one of a structural bias, as this method is incapable of specifically detecting genes with fewer than three exons. Another problem is the possibility that false positives may arise by activation of cryptic splice sites; these false positives have to be distinguished from true exons by conventional criteria such as interspecies sequence conservation or hybridisation to discrete mRNA species.

### 1.5.6 cDNA selection/capture

The techniques described have been popular when dealing with inserts from cosmid clones or subclones. Increasingly, however, large-scale physical mapping procedures rely on building YAC contigs. For any YAC, cognate cosmids (cosmid clones containing sequences from the same loci as those present in the YAC) can be identified by various procedures, including using (small) YACs as hybridisation probes to screen chromosome-specific cosmid libraries (Davies et al., 1981) or preparing cosmid libraries from the DNA of a purified YAC. An alternative is to try to identify expressed sequences directly from the YACs.

Direct cDNA selection techniques (also called direct selection and cDNA selection) involve forming genomic DNA/cDNA heteroduplexes by hybridising a complex cloned DNA, such as the insert of a YAC, to a complex mixture of cDNAs, such as the inserts from all cDNA clones in a cDNA library (Lovett, 1994). The principle underlying the technique is that cognate cDNAs corresponding to genes found within the YAC will bind preferentially to the YAC DNA; several rounds of hybridisation should lead to a huge enrichment of the desired cDNA sequences, enabling the identification of the corresponding genes. The first successful hybridisation selections described (Parimoo et al., 1991; Lovett et al., 1991) were essentially inverted screening of cDNAs with YACs. In these schemes, the YAC DNA was bound to a filter and a library of cDNAs was hybridised to this filter. After hybridisation, the filter was washed.
extensively and the specifically hybridising cDNAs were eluted. Because the yield from this type of selection is low, a PCR amplification step was incorporated to generate sufficient eluted material for additional steps or for molecular cloning. In both initial reports, purified YACs and cosmids were used as genomic targets and cDNAs were derived from conventional libraries. However, cDNA inserts were amplified using the PCR and flanking vector primers to yield a population of inserts essentially devoid of contaminating vector sequences. Repetitive elements were suppressed by blocking either the cloned genomic DNA (Parimoo et al., 1991), or the population of cDNAs (Lovett et al., 1991). It is a testimony to the robustness of the technique that both strategies worked well, and enriched the desired cDNAs a few thousand-fold after two rounds of selection.

Filter hybridisation is however not as easily quantitated or as sensitive as solution hybridisation (Young and Anderson, 1987). Due to the lack of control over the hybridisation it is likely to be irreproducible (Lovett, 1994). This is attributed to the pseudo-first-order hybridisation kinetics of filter hybridisations (Britten and Davidson, 1985) in contrast to the more easily controlled second-order kinetics in a solution hybridisation reaction. To better quantitate the repeat suppression and hybridisation steps, the selection system was modified to include biotin labelling and subsequent capture steps (Korn et al., 1992; Morgan et al., 1992). This biotin-streptavidin system (Figure 1.3) was tested and applied successfully to the isolation of cDNAs encoded by the Xq28 region in one study (Korn et al., 1992). When the same scheme (Figure 1.3) was applied to the regions surrounding the human interleukin 4 and 5 genes there was successful isolation of cDNAs in the study by Morgan et al., (1992), with enrichments achieved in the range of a few thousand-fold to more than 100,000 fold.

The direct cDNA selection method described here is a very simple, rapid and effective tool for the generation of a regional transcription map. It screens the genomic region studied as a whole and does not require its precise analysis, selection of single probes, subcloning into special vectors and the screening of large conventional cDNA libraries, if the full cDNA is not requested. cDNAs or cDNA libraries from multiple tissues can easily be mixed or used in parallel, thus increasing the probability of detection of tissue-specific transcripts. Furthermore, no specific demands such as evolutionary conservation, presence of a CpG island or favourable distribution of exon/intron boundaries must be fulfilled to allow the isolation of the gene. The analysis of the sub-library is very rapid as many identical sub-library filters can be spotted and hybridised in parallel with many probes and the resulting short or medium size cDNA clones are good substrates for DNA sequencing.

Direct selection can be affected by factors associated with nucleic acid hybridisations. One problem is that short exons fail to hybridise to their genomic locus efficiently. Another possible disadvantage is the isolation of transcripts from other regions of the genome based on the presence of a pseudogene in the region studied, in
1. Introduction

Mixture of complex amplified cDNA pools (random primed and dT synthesis)

Clone genomic DNA labeled with biotin

Block highly repetitive sequences to Cot = 20

Hybridize to Cot > 20

Capture genomic DNA plus hybridized cDNAs on streptavidin-coated paramagnetic beads

Elute, PCR amplify, reselect and clone

B: Biotin
S: Streptavidin
: Repetitive sequence element
--- cDNA inserts with amplification cassettes

Figure 1.3 A biotin-streptavidin capture system for direct cDNA selection. Lovett, M. (1994)
contrast to the CpG island approach (non-transcribed pseudogenes lose their CpG islands) or exon trapping (processed pseudogenes are not isolated). Conversely, the region-specific gene locus can serve as a template for the isolation of a transcribed pseudogene from another region of the genome in parallel with the isolation of the region-specific gene transcript. Complications arise due to the bias of the PCR amplification steps for shorter fragments (Korn et al., 1992; Sedlacek et al., 1993), leading often to a preponderance of relatively short clones in the sub-libraries, as well as the possibility of introducing sequence changes during the PCR amplification. Both problems can be overcome by the use of the probes to screen conventional cDNA libraries. Sub-libraries with large insert sizes can, however, be constructed, if appropriate precautions are taken. A summary of some advantages and disadvantages of direct selection is given in Table 1.2.

**Table 1.2 Summary of some advantages and disadvantages of direct cDNA selection** | Lovett, M. (1994)
---|---
**Advantages** | **Disadvantages**
Large target size: > 1Mb with ease | PCR-based: sequence and length bias
Very sensitive: can enrich 100,000 fold | Sensitive to exon length
Can use complex, uncloned primary cDNAs | Must have an appropriate source of cDNA
Can sample many tissues in a single experiment | Gene families and pseudogenes can complicate analysis
Some level of abundance normalisation | Low level repeats, mitochondrial cDNAs and ribosomal sequences can complicate analysis

Many groups have applied direct selection methods to various positional cloning projects and to intensive studies of defined genomic intervals, leading to the generation of transcript maps to several areas of the human genome including Xq28, around the human G6PD locus (Sedlacek et al., 1993), Xq13.3, for the transcriptional mapping of the DXS556-PGK1, 1 Mb region (Gecz et al., 1993) and the region containing the HLA class 1 region (Fan et al., 1993). The method has also been applied in the identification of genes of medical importance, especially for those involved in genetic diseases where the basic
biochemical defects are unknown. Some examples of this are the Huntington disease gene region on 4p16.3 (Rommens et al., 1993), the isolation of the gene involved in X-linked agammaglobulinaemia (Vetrie et al., 1993), five novel genes from the cri-du-chat critical region (Simmons et al., 1995), expressed sequences within the 1 Mb region flanking BRCA1 on human chromosome 17q21 (Osborne-Lawrence et al., 1995) and the recent isolation of conserved cDNAs and a novel CDC45-like gene from the DiGeorge critical region (McKie et al., 1998). In addition direct selection has also been used recently in the identification of 37 novel transcripts using a cosmid contig at the 'Down syndrome critical region' of human chromosome 21q22 (Guimera et al., 1997) and the isolation of chromosome 18 specific brain transcripts as positional candidates for bipolar disorders (Yoshikawa et al., 1997).

Direct selection also holds promise in building comparative interspecific genome maps, for example in selecting mouse and human cDNAs across a particular genomic region. The availability of cDNAs from mouse embryos at various stages of development makes this a particularly attractive approach for mapping genes that are developmentally regulated. Direct selection of genomic DNA fragments from one species using target genomic DNA from another species is technically difficult, but has already been achieved (Sedlacek et al., 1993a). Using human cosmids from the Xq28 region, sub-libraries were generated with homologously conserved sequences from mouse and pig DNA. In another study, a group of cDNA segments was selected by direct hybridisation of mouse cerebellar cDNAs against genomic DNA pools generated by microdissection of the mouse chromosome 16 C3-C4 region, enabling construction of a new, region-specific partial transcription map (Wei et al., 1996).

Coincident sequence cloning is a related PCR-based method which involves selecting for sequences that are co-incident between two complex populations of DNA sequences (Brookes et al., 1993). One application involves using YAC clones or other genomic DNA clones from a disease gene region as one DNA source and a second DNA source consisting of cDNA from cells in which the sought-after gene is thought to be expressed. In principle, this method is the PCR equivalent of the hybridisation-based cDNA selection technique: the objective being to identify expressed sequences in the genomic DNA clones because of their coincidence with transcripts found in the cDNA. The most modern version, end ligation coincident sequence cloning, has however the advantage of an end ligation reaction which endows the method with great specificity (Brookes, 1994).

Many other techniques have been developed using the direct selection framework such as the selection of hybrids by affinity capture (SHAC) (Chen-Liu et al., 1995) which involves the generation of cDNAs enriched in sequences from a cytogenetically defined chromosome region. An efficient method has also been described that combines flow-sorted whole chromosomes and cDNA hybrid selection (Rouquier
et al., 1995) and an entire human chromosome 5-specific genomic DNA cosmid library and cDNA hybrid selection (Del Mastro et al., 1995) for the large scale analysis of genomic regions. This method eliminates the requirement for YAC or cosmid contigs, while allowing direct identification of chromosome specific genes and could facilitate the construction of an expression map of the human genome. This is evident in a recent report, where 2006 ESTs were generated and analysed, derived from a collection of direct selection cDNA libraries that were highly enriched for human chromosome 7 gene sequences (Touchman et al., 1997), contributing to the construction of a chromosome 7 transcript map.

1.5.7 Subtractive Selection Cloning and Differential Hybridisation methods

Subtractive selection cloning has been useful in enriching for cell- or tissue-specific clones from cDNA libraries of some human tissues. Subtracted human retina and retinal pigment epithelium (RPE) cell line cDNA libraries have been generated by eliminating most of the constitutively expressed genes using a biotin-based procedure (Swaroop et al., 1991). Identification of ESTs for novel cDNAs from the subtracted RPE library and chromosomal localisation of these clones has also been reported (Gieser and Swaroop, 1992). Identification of a neural retina-specific gene (NRL) encoding a putative DNA binding protein of a leucine zipper family has also been reported from expression analysis of clones from the subtracted retinal library (Jackson et al., 1993). The subtraction-selection strategy has been employed to identify putative candidate gene regions for X-linked retinal diseases, using a human X-chromosome-specific genomic library (Yan and Swaroop, 1994).

Differential hybridisation schemes have been used to identify photoreceptor-specific genes. In earlier work, Bascom et al., (1992) utilised a differential hybridisation screen of retinal cDNA clones to identify genes likely to be important in retinal biology and identified ROM1, a photoreceptor membrane protein that localises to disc rims, and Chx-10 (Liu et al., 1994), a retinal homeobox gene. Recently, a differential hybridisation screening strategy was devised to identify novel retinal-specific genes with an emphasis on cone-specific genes (Swanson et al., 1997). Clones that hybridised exclusively or preferentially to a cDNA probe from the retina of the 13-line ground squirrel (13GS; which has a cone-predominant retina), as compared to a probe made from human fibroblast cDNA, were selected for further investigation.
1. Introduction

1.5.8 Computer-based DNA sequence analysis

Efficiency is the main reason for using computational tools for elucidating the structure of, and assigning tentative function to genes. If the DNA sequence is available, almost any kind of computational analysis is cheaper and faster than almost any kind of experimental analysis. Experimental methods for locating genes have grown greatly in power (Church et al., 1994). Yet, in many areas, experimental and computational methods still provide complementary information (Caskey et al., 1992). Complementary gene identification has recently played prominent roles in, for example, identifying candidate 'disease genes' (Brody et al., 1995), compiling inventories of possible genes in large-scale genomic sequencing projects (Sulston et al., 1992), and helping to assign tentative functions as the first ever organism-wide gene inventories progress (Casari et al., 1995).

Once a DNA clone has been sequenced, computer analyses can be used to determine whether the sequence is likely to represent part of a gene. Two major types of software are used: 1) homology searches: the nucleotide sequence and the inferred amino acid sequences for all three reading frames are compared against all available DNA and protein sequences which have been recorded in electronic databases. Any significant matching between the test sequence and the sequence of a known gene, cDNA or protein, whether of human or non-human origin, indicates a gene associated sequence (either a functional gene, pseudogene or gene fragment) (Rouleau et al., 1993). 2) Exon prediction: this approach is becoming increasingly important as large scale DNA sequencing project gather momentum. The programs are designed to scan a DNA sequence in order to identify the locations of likely exons by screening for conserved sequences found at exon/intron junctions and the splice branch site, and the presence of comparatively long ORFs. As yet however, even the best of such programs, example, as the Gene Recognition and Analysis Internet Link (GRAIL) software, have been only moderately successful in identifying exons when tested against genes whose exon organisations had previously been established (Lopez et al., 1994). When the relevant gene is GC-rich however, exon prediction can be quite accurate. A recent successful application has been the characterisation of the adult polycystic kidney disease gene PKD1, which proved to be problematic because of the existence of several transcriptionally active copies of closely related PKD1-like sequences mapping centromeric to PKD1 on chromosome 16p13.1 (Burn et al., 1995).
1.5.9 Summary, biases and limitations of gene identification methods

A summary of the experimental steps necessary for analysing a YAC for transcribed sequences by exon trapping, cDNA selection or direct screening is outlined in Table 1.3. Table 1.4 summarises the biases and limitations of these different gene identification methods. These limitations arise from three inter-related, practical problems: verification, completeness and redundancy. Every technique is affected by these to varying degrees. The problem of verification, validating true positives and discarding false positives, is greatest for techniques which identify genes independent of expression in any particular tissue (e.g. exon trapping, GRAIL). Techniques which depend upon expression in a given tissue or set of tissues (e.g. cDNA selection) are not as severely affected by the problem of verification (there is the strong prediction that the gene should be expressed in the tissue(s) used to make the cDNA), but there is the problem of completeness: to ensure that all genes in a region have been identified would require analyses with a large number of different cDNA sources. In this respect, the problems of completeness and verification are reciprocal. But the problem of completeness goes further, to include the biases inherent in particular techniques. For example cDNA screening has an expression level bias, detection of genes expressed at less than 0.01% being uncertain, and exon trapping has a structural bias, being incapable of specifically detecting genes with fewer than three exons. The last problem, redundancy, or efficiency, is the hardest to assess. For example, in cDNA selection, genes encoding abundant, ubiquitous transcripts will be isolated many times in each selected library, hence efficiency being frequently inversely related to completeness. There are two ways to address these problems using existing techniques: the first is to rely on a number of different techniques with different biases, in hopes that the sum of these partial solutions will approach completeness, while the second is to combine techniques in such a way that their shortcomings are cancelled out in the combination.

1.5.10 Similarity searches using sequence databases

Similarity searching is the process of comparing a new sequence against all other known sequences, then attempting to infer the function of the new sequence by assessing the matches and their biological annotations as described in the databases themselves and the literature.

There are a number of important issues in searching DNA and protein sequence databases (Altschul, et al., 1994), but the most important is access to a comprehensive and up-to-date data repository. GenBank distributed by the National
### Table 1.3 Experimental steps necessary to analyse a YAC for transcribed sequences

<table>
<thead>
<tr>
<th>Method</th>
<th>Exon trapping</th>
<th>cDNA selection</th>
<th>Direct cDNA screening</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• subclone &gt; arrayed cosmid library</td>
<td>• purify</td>
<td>• subclone &gt; arrayed cosmid library</td>
</tr>
<tr>
<td></td>
<td>• subclone pools of cosmids into splice vector</td>
<td>• immobilise (filter/beads)</td>
<td>• deplete cDNA probe</td>
</tr>
<tr>
<td></td>
<td>• transfect COS cells</td>
<td>• block repeats, yeast</td>
<td>• hybridise replica filters of arrayed cosmids</td>
</tr>
<tr>
<td></td>
<td>• extract mRNA</td>
<td>• hybridise with cDNA library inserts</td>
<td>• Southern blot positive cosmids</td>
</tr>
<tr>
<td></td>
<td>• reverse transcribe/PCR</td>
<td>• wash/elute</td>
<td>• hybridise with depleted cDNA probe</td>
</tr>
<tr>
<td></td>
<td>• clone PCR products in plasmid &gt; array clones</td>
<td>• PCR</td>
<td>• subclone individual positive fragments in plasmids</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• clone PCR products plasmid &gt; array clones</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Initial collection of gene candidates:</th>
<th>redundant collection of small products</th>
<th>redundant collection of small PCR products</th>
<th>non-redundant collection of genomic fragments carrying transcribed sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Follow-up analysis</td>
<td>• reduce redundancy by screening random collection with individual members to exclude multiple isolates</td>
<td>• reduce redundancy by screening random collection with individual members to exclude multiple isolates</td>
<td>• sequence</td>
</tr>
<tr>
<td></td>
<td>• screen several cDNA libraries from different tissues/times to validate exons and to link exons through a common cDNA</td>
<td>• check chromosomal location to exclude non-specific isolates</td>
<td>• GRAIL analysis</td>
</tr>
<tr>
<td></td>
<td>• screen cDNA libraries from specific tissue/time to link partial overlapping clones</td>
<td></td>
<td>• design oligo probe from exon</td>
</tr>
</tbody>
</table>
### Table 1.4  Biases and problems of different gene identification methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Structural bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exon trapping</td>
<td>genes with $\geq$3 exons variation</td>
</tr>
<tr>
<td></td>
<td>variation with position of exons relative to restriction sites used for cloning</td>
</tr>
<tr>
<td>cDNA selection</td>
<td>No</td>
</tr>
<tr>
<td>Direct cDNA</td>
<td>No</td>
</tr>
<tr>
<td>screening</td>
<td>Limited to sensitivity of hybridisation reaction ($\sim 10^{-5}$)</td>
</tr>
<tr>
<td>GRAIL</td>
<td>?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Expression bias</th>
<th>Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>Tissue and time</td>
</tr>
<tr>
<td>no</td>
<td>? tissue specific splicing</td>
</tr>
<tr>
<td>variable dependent upon number of clones picked</td>
<td>limited to tissue and time in cDNA library</td>
</tr>
<tr>
<td>Limited to tissue and time used in experiment</td>
<td>redundant collection of initial candidates</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>validation requires extensive genomic sequence</td>
<td>redundant isolation of genes expressed in more than one tissue/time</td>
</tr>
</tbody>
</table>

Center for Biotechnology Information (NCBI), the EMBL nucleotide sequence database (36%) facilitated by the European Bioinformatics Institute (EBI), and the DNA Database (4%) of Japan (DDBJ) are three partners in a long-standing collaboration to collect and distribute all publicly available sequence data. Sites in these respective countries exchange new sequence data and updates over the Internet on a daily basis and make this information available to the public by a variety of means including e-mail, anonymous ftp and the World Wide Web (Harper, 1994) giving a comprehensive up-to-date database. Timely access to complete and "nonredundant" sequence databases has become relatively simple and inexpensive.

Strong biases exist in protein and nucleic acid sequences and sequence databases. Many of these reflect fundamental mosaic sequence properties that are of considerable biological interest in themselves, such as segments of low compositional complexity or short-period repeats. Databases also contain some very large families of related domains, motifs or repeated sequences, in some cases with hundreds of members. These biases commonly confound database search methods and interfere with the discovery of interesting new sequence similarities. Problems may include the occurrence of misleading, spuriously high scores, ambiguities in the phase of sequence alignments and can result in chance similarities being claimed significant, or biologically important relationships being overlooked. Large improvements in the efficiency of searching databases have been implemented using new strategies, such as preprocessing a query sequence to identify known domains and motifs, dispersed repeats, low complexity segments and other regions of compositional bias such as potential membrane spanning and α-helical coiled-coil regions. Another complementary strategy was to reduce the redundancy in the target database(s) to be searched (Altschul et al., 1994).

The computer programs used to search the sequence database itself is also of importance. A number of different search algorithms have been developed over the years and further information about them may be found in Altschul et al., (1994). The BLAST family of programs ("Basic Local Alignment Search Tool") is a set of similarity search programs designed to explore all of the available sequence databases regardless of whether the query is protein or DNA. The BLAST programs have been designed for speed, with a minimal sacrifice of sensitivity to distant sequence relationships. The scores assigned in a BLAST search have a well-defined statistical interpretation, making real matches easier to distinguish from random background hits. It therefore offers a good combination of speed, sensitivity, flexibility and statistical rigor. BLAST uses a heuristic algorithm which seeks local as opposed to global alignments and is therefore able to detect relationships among sequences which share only isolated regions of similarity (Altschul et al., 1990).
A recent paper by Altschul et al., (1997) provides modifications to the current algorithms in the BLAST program that save the user time and provide far greater return by allowing the production of gapped alignments (Gapped BLAST) and motif searching (Position-specific Iterated BLAST - PSI-BLAST) within the BLAST system. Gapped BLAST and PSI-BLAST are useful search tools provided by the BLAST server (version 2.0, http://www.ncbi.nlm.nih.gov/BLAST/blast) (Altschul et al., 1997). The Gapped BLAST algorithm allows gaps (deletions and insertions) to be introduced into the alignments that are returned. Allowing gaps means that similar regions are not broken into several segments. The scoring of these gapped alignments tends to reflect biological relationships more closely. Position-Specific iterated BLAST (PSI-BLAST) provides an automated, easy-to-use version of a "profile" search, which is a sensitive way to look for sequence homologues. The program first performs a gapped BLAST database search. The PSI-BLAST program uses the information from any significant alignments returned to construct a position-specific score matrix, which replaces the query sequence for the next round of database searching. PSI-BLAST may be iterated until no new significant alignments are found. At this time PSI-BLAST may be used only for comparing protein queries with protein databases.

There are five implementations of BLAST, three designed for nucleotide sequence queries (BLASTN, BLASTX, TBLASTX) and two for protein sequence queries (BLASTP, TBLASTN). The former are used for the analysis of genomic sequence (including putative exons) and "single-pass" cDNA data, the latter usually come into play when one has identified discrete gene products from complete sequence. BLASTN compares a nucleotide query sequence against a nucleotide sequence database, BLASTX compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database and TBLASTX compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. Protein sequence query programs BLASTP compares an amino acid query sequence against a protein sequence database and TBLASTN compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands).

The fundamental unit of the BLAST algorithm output is the High-scoring Segment Pair (HSP). An HSP consists of two sequence fragments of arbitrary but equal length whose alignment is locally maximal and for which the alignment score meets or exceeds a threshold or cut-off score. A set of HSPs is thus defined by two sequences, a scoring system and a cut-off score; this set may be empty if the cut-off score is sufficiently high. The HSP consists of one segment from the query and one from the database sequence. In addition an MSP or Maximal-scoring Segment Pair has also been defined. It is defined by two sequences and a scoring system and is the highest-scoring of all possible segment pairs that can be produced from the two sequences.
The task of finding HSPs begins with identifying short words of length $W$ in the query sequence that either match or satisfy some positive-valued threshold score $T$ when aligned with a word of the same length in a database sequence. $T$ is referred to as the neighborhood word score threshold (Altschul et al., 1990). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Extension of the word hits in each direction are halted when the cumulative alignment score falls off by the quantity $X$ from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments, or when the end of either sequence is reached. The sensitivity and speed of the program can be adjusted via the standard BLAST algorithm parameters $W$, $T$, and $X$ (Altschul et al., 1990); selectivity of the programs can be adjusted via the cut-off score.

1.6 Inherited retinal degenerations

Inherited diseases that cause the retina to degenerate, leading to either partial or total blindness, affect approximately 1 in 3,000 people. Rapid progress is being made towards identification of the genetic causes of inherited retinal diseases, such as retinitis pigmentosa and macular degeneration. Linkage studies of large families and candidate gene screening of known retinal genes have already identified 59 independent genetic loci that can cause retinal degeneration (Fig 1.4) (Sullivan and Daiger, 1996). The online database of human genetic diseases OMIM (section 1.2.6) lists over 100 distinct diseases that include some form of retinal degeneration. These can be classified by mode of inheritance: autosomal dominant, autosomal recessive, X-linked and digenic forms are well documented in addition to a small number of families where retinal degeneration is associated with other symptoms for example RP and hearing loss in Usher's syndrome.

The heterogeneity observed in inherited retinal degeneration includes genetic heterogeneity (different genes causing the same disease), allelic heterogeneity (different mutations in the same gene causing either the same disease or different diseases), and clinical heterogeneity (the same mutation causing different symptoms in different individuals even within the same family). The genetic heterogeneity observed in these diseases suggests that retinal degeneration is a common end point for many biochemical abnormalities, and may be a reflection of the limited response of the retina to disease. Clinical heterogeneity in retinal diseases is poorly understood, but implies that genetic background and environment must play significant roles in clinical expression. One reason for this complexity is the broad definition of 'retinal degeneration'. This term encompasses diseases of the peripheral retina such as retinitis pigmentosa (RP) and
FIG 1.4 Cloned and/or mapped genes causing retinal degeneration in humans (red) and retina-specific genes (green). Disease genes are labeled to the right of the chromosome and retina-specific genes are to the left. Cases where the gene causing the disease has been identified are shown boxed in gold. Abbreviations for cloned disease-causing genes are listed below, in alphabetical order. References for all of the genes labeled can be found in the Genome Data Base, available on the World Wide Web at: http://gdbwww.gdb.org/gdb/browser/docs/topq.html. Cloned retinal genes: BCP, blue cone pigment; CHM, choroideremia protein; CNCG1, cyclic-GMP-gated channel, α subunit; DMD, dystrophin; GCP, green cone pigment; MYVI A, myosin VI A; NDP, Norrie disease protein; OAT, ornithine aminotransferase; PDE6A, cyclic GMP phosphodiesterase, rod-specific α subunit; PDE6B, cyclic GMP phosphodiesterase, rod-specific β subunit; PLCB4, phosphoinositide-specific phospholipase Cβ4; RCP, red cone pigment; RDS, peripherin/RDS, rhodopsin, ROM1, rod outer-membrane protein; RPGR, retinitis pigmentosa GTPase regulator; SAG, S-antigen (arrestin); TIMP3, tissue inhibitor of metalloproteinase. Retinal degeneration loci: adRP, dominantly or digenic retinitis pigmentosa; adRP, digRP, MD, dominant or digenic retinitis pigmentosa or macular degeneration; AGS, Alagille syndrome; arRP, recessive retinitis pigmentosa; BCM, blue cone monochromatism; CBD, colorblindness, deuteranopic (non-progressive); CBP, colorblindness, protanopic (non-progressive); CBT, colorblindness, tritanopic (non-progressive); CHM, choroideremia; CSNB1, congenital stationary night blindness (non-progressive); DMD, Oregon eye disease; GA, gyrate atrophy; ND, EVR2, Norrie disease or familial exudative vitreoretinopathy (non-progressive); OD, recessive Oguchi disease; RP3, retinitis pigmentosa; SFD, Sorsby fundus dystrophy; USH1B, Usher syndrome 1B.

congenital stationary night blindness (CSNB), diseases of the central retina such as macular degeneration, and many others in which the pattern of degeneration is complex. The key to unravelling the complexities of retinal degeneration is to first understand the structure and function of the normal retina in the visual process.

1.6.1 Structure of the eye

The function of the eye is to produce a clear image of the external world and to transmit this to the visual cortex of the brain. In order to produce a clear image the eye must have constant dimensions, a clear optical pathway and the ability to focus light on to the retina. These requirements, and the need for protection of the globe, determine the special structure of the eye and its associated apparatus (Figure 1.5).

1.6.1.1 The globe

The wall of the eye consists of three layers. The outer, inelastic layer, comprising the transparent cornea and the opaque sclera, provides the necessary rigidity of the eye when distended by the intraocular pressure. The middle, vascular layer is known as the uvea and comprises the choroid, the ciliary body and the iris. The choroid is responsible for the nutrition of the outer part of the retina. The inner layer is the retina which extends forwards to within 6 mm of the limbus (junction of cornea and sclera). The eye contains three optically clear spaces. The anterior chamber lies in front of the iris and the posterior chamber immediately behind the iris, bounded by the lens and ciliary body. These two chambers, which communicate through the pupil, are filled with clear aqueous humour. Behind the lens lies the vitreous cavity filled by the vitreous humour, a clear, gel-like structure composed of collagen fibrils suspended in a viscous liquid.

1.6.1.2 The neural retina

The retina is the neural sensory layer of the eye which houses a dense array of light-sensitive photoreceptors containing visual pigment molecules that initiate the neural response to light which has been focused onto the retina by the cornea and lens. Before light can reach the photoreceptors in the innermost layer of the retina, it must first penetrate the many nerve cell bodies and processes of the remaining retinal layers. Structurally and functionally the retina consists of two distinct layers: the non-neural retinal pigment epithelium (RPE), and the tightly opposed neural retina (Fig. 1.6).

The vertebrate retina has major neuronal cell types organised into 5 distinct layers; three layers of cell bodies and two layers of synaptic connections. The neurons of the retina are divided into three layers: (1) the most external (the outer or scleral layer) is
Figure 1.5 Transverse section of the Eye
the photoreceptor cell layer, which includes the outer and inner segments and the layer of photoreceptor cell bodies (or outer nuclear layer), (2) the layer of intermediate neurons (or inner nuclear layer), and (3) the layer of ganglion cells. The synapses are confined to the two synaptic, or plexiform layers—the outer and inner plexiform layers.

The receptor layer of the retina is intimately apposed to the pigment epithelium; light rays pass through the ganglion cells and inner layers to reach the photoreceptor cells, where the light initiates an electrochemical cascade. The elongated axonal processes of the photoreceptor cells synapse in the outer plexiform layer with the processes of the bipolar cells and horizontal cells, whose cell bodies are located in the inner nuclear layer. The horizontal cells form a means of lateral communication between photoreceptor cells. The bipolar cells synapse with amacrine and/or ganglion cells. Amacrine cells are located primarily in the vitreal portion of the inner nuclear layer. They extend processes to adjacent amacrine or bipolar cells, and their axons synapse with ganglion cells. The axons of ganglion cells converge to form the nerve fibre layer and exit the eye at the optic nerve. Another cell type located among the amacrine cell perikarya, are the inner plexiform cells. Unlike amacrine cells however, the inner plexiform cells extend processes into both synaptic layers. The various retinal cell types and their connections are presented in Fig. 1.6.

The external (outer) limiting membrane is formed by junctional complexes between cell membranes of the major glial cells, the Muller cells, and photoreceptor inner segments. The internal (inner) limiting membrane consists of a basement membrane, which is actually a surface modification of the vitreous body, and the expanded vitreal processes of Muller cells. Within the retina fine processes of Muller cells partially envelop all the neurons. These processes extend laterally from the main portion of the Muller cells, which extend from the external to the internal limiting membranes as radial fibres.

The retinal pigment epithelium (RPE) comprises a single layer of cells with villous processes projecting from its apical surface interdigitating with the photoreceptor outer segments. The basal surface has numerous infoldings and is functionally linked to the choroid via Bruch's membrane. The RPE contains melanin granules which control light scatter and permit precise image formation. Tight junctions connecting the RPE cells also render the RPE an important protective barrier between the choroidal circulation and the neural retina. Other important functions served by the RPE layer include supply of nutrients to the neural retina, production of interphotoreceptor space glycosaminoglycans, uptake and storage of retinoids and phagocytosis of rod, and to a lesser extent, cone outer segments.
Figure 1.6  Schematic diagram of cell types and histologic layers of the human retina. Shown are Bruch’s membrane (considered as the innermost layer of the choroid, but is in fact the basement membrane of the RPE) and the edge of the vitreous (V ). The basic relationship between rod (R ) and cone (C ) photoreceptors as well as bipolar (B ), horizontal (H ), amacrine (Am ), inner plexiform cell (I ), displaced amacrine (DA ), and ganglion (G ) neurons are depicted. Note that the Muller cell (M ) extends almost the width of the retina; the apical processes of the Muller cell form the external limiting membrane (ELM ), while the foot processes of the Muller cell partially form the inner limiting membrane (ILM ). Astrocytes (As ) are found primarily in the nerve fiber layer (NFL ).

Dowling, J.E. and Boycott, B.B. (1966)
1.6.2 The photoreceptors

The primate retina is a "duplex" retina consisting of both rod and cone photoreceptor cells (Fig. 1.7). Rods mediate vision in dim light, whereas cones function in bright light. Rods provide great sensitivity, especially to blue-green light (scotopic vision), whereas cones provide visual acuity for pattern detection as well as colour vision (photopic vision). All photoreceptors have an outer segment that contains the photosensitive pigment molecules. The outer segment is the site of phototransduction, where conversion of light energy into an electrical signal occurs. Both rods and cones contain an elaborate system of stacked membranous discs that arise during development as a series of invaginations of the cell's plasma membrane. In rods, most of these sacs eventually become separated from the outer membrane, whereas cone discs remain connected to the surface membrane. This rod/cone difference in outer segment geometry is thought to reflect the fact that cones function in bright light and thus require rapid recycling of their pigment molecules; the continuum between the disc and plasma membrane providing rapid access of the regenerated chromophore from the cell exterior to the bleached pigment. The inner segments contain densely packed mitochondria and the usual cellular machinery, including Golgi apparatus, endoplasmic reticulum and nucleus. Inner and outer segments are linked by a narrow ciliary connection which transmits components from the inner segment and cell nucleus to the discs and their plasma membrane. At the synaptic endings are numerous vesicles.

Rod cells are sensitive to light because they contain a visual pigment called rhodopsin, which is capable of trapping photons. Rhodopsin is arranged in a single molecular layer in the discs of the outer segment and is composed of two essential parts. One is the light-absorbing chromophore, vitamin A aldehyde, which is called 11-cis-retinal. Chemically, this component is closely related to vitamin A. Retinal is attached to the second moiety of rhodopsin, known as opsin (Wald, 1968). Like rhodopsin, the visual pigments found in cone cells are composed of two parts. The light-absorbing molecule appears to be similar to retinal. In the primate retina there are three types of cone cells that, unlike rod cells, are sensitive to either blue, green, or orange light. The sensitivity of the retinal molecule in each cone cell to a particular wavelength is determined by the specific type of protein (opsin) to which it is bound. The major rod pigment rhodopsin peaks at about 500 nm, whereas the cone pigments in primates absorb maximally in the blue (~450 nm), green (~530 nm) and yellow (~565 nm) regions of the spectrum (Wald and Brown, 1965). Cone opsins are collectively referred to as iodopsin. Within the macula lies the fovea, where the photoreceptors are exclusively cones, and at the centre of the fovea is the foveola, a depressed pit resulting from displacement of the inner cell layers so that light falls directly on the cones. This is the point of greatest visual acuity.
Figure 1.7 Illustrations of a typical rod and cone. OS, outer segment; IS, inner segment; RD, rod discs; CI, cone invaginations; C, ciliary connection between inner and outer segments; M, mitochondria; N, nucleus; S, synaptic ending. The synaptic ending contains vesicles, some apposed to ribbon structures that are found in both rods and cones. The ciliary connection is a slender structure connecting inner and outer segments. In addition to the labelled organelles there are also rough and smooth endoplasmic reticulum, Golgi apparatus, and the other usual components of cells. Dowling, J.E. and Boycott, B.B. (1966)
and colour perception. Cone density decreases sharply from the foveal centre, with a 90% reduction by 1mm in humans. Rod coverage follows an inverse pattern, increasing from 0% at the fovea to a maximum of 55-65% in far peripheral retina. Thus rods are essential for 'peripheral vision', while cones are required for 'central' vision.

1.6.3 The rod phototransduction pathway

The primary function of the retina is phototransduction. The first steps in the phototransduction pathway occur in the cells of the photoreceptor layer, the rods and cones. Here, opsin proteins, bound to the light-responsive chromophore 11-cis-retinal, are embedded in the plasma membrane and in stacked disc structures within the outer segment of each cell. When opsin absorbs a photon of light it becomes activated and, in turn, activates the G protein transducin. Upon activation, transducin activates cyclic GMP (cGMP) phosphodiesterase (PDE), which hydrolyses cGMP. As cGMP levels in the cytoplasm drop, the cGMP-gated ion channels (CNCGs) within the plasma membrane close, leading to hyperpolarisation of the membrane. Numerous other specialised proteins are also required to modulate and inhibit this cascade of reactions, ultimately returning the opsin molecule to its inactive state, ready to intercept another photon. This basic pathway takes place in both rods and cones. However, for many photoreceptor proteins, there are rod-specific and cone-specific versions of the same enzyme.

The phototransduction process has been extensively studied (Stryer, 1991; Yau, 1994). Figure 1.8 (Polans et al., 1996) shows the main components and biochemical reactions which are involved. The initial step is the activation of rhodopsin \( (R \rightarrow R^*) \) caused by the absorption of a photon of light. In turn, \( R^* \) activates transducin \( (T) \), a trimeric G protein comprised of \( \alpha, \beta, \) and \( \gamma \) subunits. In the dark, \( T\alpha \) is associated with \( T\beta/\gamma \) and has bound GDP, but after the absorption of light, \( R^* \) binds to \( T\alpha \), inducing the exchange of GDP for GTP and separation of \( T\alpha\)-GTP from \( T\beta/\gamma \). \( R^*-T\alpha\)-GTP dissociates into \( T\alpha\)-GTP and \( R^* \), and the latter can then activate another molecule of transducin. The activation of hundreds of molecules of transducin by \( R^* \) is the first step of amplification in the visual process. Next, \( T\alpha\)-GTP activates PDE, which is composed of four subunits, two catalytic \( (\alpha \text{ and } \beta) \) and two inhibitory \( (\gamma) \), by binding the \( \gamma \) subunits, therefore relieving the inhibitory constraint on the enzyme. The activated PDE hydrolyses many molecules of cyclic GMP. Because the high level of cGMP keeps the cyclic GMP-gated cation channels of the photoreceptor plasma membrane open in the dark, the reduction in the concentration of cyclic GMP causes the channels to close and leads to hyperpolarisation of the plasma membrane. After closure of the ion channels, the free \( \text{Ca}^{2+} \) concentration in the outer segment decreases, because \( \text{Ca}^{2+} \) ions are transported out of the cells through the \( \text{Na}^+\text{-Ca}^{2+} \) exchanger.
Figure 1.8  Key components of the phototransduction cascade in vertebrate rod photoreceptor cells (from Polans et al., 1996)

$hv$: incident photon

$R$: rhodopsin (* active)

$Rec$: recoverin

$T$: transducin

$T\alpha^*: $ activated $\alpha$ subunit of transducin

$GC$: guanylate cyclase

$RK$: rhodopsin kinase

$PDE$: cGMP phosphodiesterase (* active)

$GCAP$: guanylate cyclase activating protein

$Arr$: arrestin

$CaM$: calmodulin
Rhodopsin, in the meantime, is deactivated as a combined result of phosphorylation by the enzyme rhodopsin kinase, and binding of a 48 Kd protein called arrestin which block the further activation of transducin and therefore terminate phototransduction. All-trans retinal is reduced and recycled. Blocking the activated receptor is not sufficient to arrest the cascade; the Ta-GTP already formed will maintain the PDE activity by remaining bound to the γ subunit inhibitors until they lose this capacity due to the intrinsic GTP hydrolytic activity of Ta. The regenerated Ta-GDP has a low affinity for the PDE γ subunits which immediately rebind to the catalytic α and β subunits of PDE and block its activity. Ta-GDP then recombines with Tβγ and reassociates with the disc membrane. The system is reset to the dark-adapted state when rhodopsin is regenerated by recombining with 11-cis-retinal and when its bound phosphate is removed by a phosphatase.

1.6.4 Calcium feedback and the recovery of the dark state

In darkness, there is a steady circulation of Ca\(^{2+}\) across the membrane of the outer segment, which consists of an influx through the cGMP-gated channels and an efflux through a transport mechanism that involves the exchange of extracellular Na\(^{+}\) with intracellular Ca\(^{2+}\) and K\(^{+}\) (Cervetto et al, 1989). In the light, the closure of the cGMP-gated channels stops the influx of Ca\(^{2+}\), but the efflux continues, reducing the concentration of free Ca\(^{2+}\) in the outer segment (Gray-Keller and Detwiler, 1994). This decrease in Ca\(^{2+}\) triggers a negative feedback that results in light adaptation. If the Ca\(^{2+}\) mediated feedback is largely removed by minimizing changes in the concentration of Ca\(^{2+}\) in the outer segment, rods (and cones) show hardly any active adaptation to the background illumination.

While the role of Ca\(^{2+}\) as a mediator of light adaptation has been accepted for some time, the underlying mechanisms have only been determined recently. Several Ca\(^{2+}\)-mediated feedback pathways have been discovered. First, guanylate cyclase (GC), which synthesises cGMP, is inhibited by Ca\(^{2+}\), so that when the concentration of Ca\(^{2+}\) decreases in the light the guanylate cyclase activity increases, which counteracts the light-stimulated hydrolysis of cGMP (Koch and Stryer, 1988). A membrane-associated Ca\(^{2+}\)-binding protein termed guanylate cyclase activating protein (GCAP) activates the membrane bound photoreceptor guanylate cyclase(s) RetGC1 and RetGC2 (Gorczyca et al., 1995). GCAP1 has been shown to directly activate guanylate cyclase (Laura et al., 1996). A reduction of calcium results in uninhibition of GCAP, activation of RetGC and consequent conversion of GTP to cGMP, thus restoring the dark state (Dizhoor and Hurley, 1996). A second feedback pathway involves the light-stimulated phosphodiesterase activity which seems to be enhanced by Ca\(^{2+}\) (Kawamura and Murakami, 1991). This effect is mediated by another Ca\(^{2+}\)-binding protein, called
recoverin or S-modulin, which acts by inhibiting the phosphorylation of activated rhodopsin at high concentrations of Ca$^{2+}$ (Gorodovikova et al., 1994). Thus, high concentrations of Ca$^{2+}$ prolong the lifetime of photoexciited rhodopsin, which leads to the activation of a larger number of transducin molecules and hence greater light-stimulated phosphodiesterase activity. Conversely, when the concentration of Ca$^{2+}$ falls in the light, the lifetime of photoexcited rhodopsin decreases, which results in lower phosphodiesterase activity. Finally, the cGMP-gated channel is a target of Ca$^{2+}$ feedback. It has been shown (Hsu and Molday, 1993) that Ca$^{2+}$ decreases the apparent affinity of the cGMP-gated channels for cGMP, so that when the concentration of Ca$^{2+}$ falls in the light, some of the channels tend to re-open despite the decrease in concentration of cGMP.

In short, several pathways are activated when the concentration of Ca$^{2+}$ decreases in the light, and they all lead to negative regulation of the effect of light (Koutalos and Yau, 1996). Apart from speeding up the recovery of the cell's response to light, these feedback mechanisms serve to diminish the effect of illumination, thus enabling the cell to adapt to background light. Phototransduction in cones is thought to closely parallel that in rods, with many key rod proteins having cone counterparts such as opsin, transducin, cGMP PDE, recoverin (visinin), cGMP-gated channel and the Na$^+$/Ca$^+$ exchanger (Yau, 1994).

### 1.6.5 Apoptosis-a cause for photoreceptor cell death

When a photoreceptor cannot perform its function because of a structural or biochemical defect, it may degenerate. If the disease affects rods primarily, visual field loss progresses from the mid-periphery towards the middle of the retina, leaving the patient with night blindness and 'tunnel vision'. Conversely, diseases that primarily affect cones have the opposite effect, destroying the central region of the retina (the macula), where cones are most abundant. The steps leading to retinal degeneration, of either rods or cones, are not well understood, although several scenarios have been proposed. A current theory is that apoptosis, a form of programmed cell death, is ultimately responsible for retinal degeneration (Lolley et al., 1994).

#### 1.6.5.1 Pathogenesis

Apoptosis is an active and highly regulated cellular activity that occurs normally in physiological situations in which the elimination of particular cells is necessary for the well being of an organism, for example during embryonic development. It has been well recorded in many tissues of both vertebrates and invertebrates as a cause of cell loss (Portera-Cailliau et al., 1994). A cell committed to apoptosis withdraws from its
neighbours and is characterised morphologically by disintegration of the nucleolus and generalised condensation of the chromatin resulting from incision of most of the nuclear DNA into short chains of nucleosomes by an endogenous non-lysosomal nuclease (Wyllie et al., 1981). Inability to repair multiple breaks in the DNA strands is the lethal event which probably kills the cell (Schwartzman and Cidlowski, 1993). The cell breaks down into membrane bound apoptotic bodies which are engulfed by neighbouring cells or sometimes by circulating macrophages, neatly removing them without leakage of the cell contents which may be detrimental to surrounding cells (as shown by inflammation in necrosis). In contrast to necrosis, the process affects individual cells within a tissue, with its neighbours remaining healthy, and takes place in the absence of inflammation (Gregory and Bird, 1995). In the normal developing retina approximately 50% of ganglion, amacrine and bipolar cells and around 5% of photoreceptor cells die by apoptosis (Cepko, 1996). Apoptosis has been shown to be the cause of cell death in all animals with genetically determined retinal degeneration examined to date, whether naturally occurring or induced by transfection by mutant genes (Chang et al., 1993; Tso et al., 1994). Thus, it appears that cell loss is not a direct consequence of the intrinsic metabolic environment of the tissue.

1.6.5.2 Animal models of retinal degeneration

Apoptosis has been found to be the final common pathway of cell death due to mutations in three genes whose primary site of action is in the photoreceptor (Chang et al., 1993). Mutations in three photoreceptor-specific genes have been identified as the cause of disease in individuals affected by retinitis pigmentosa (RP). These are the β subunit of cyclic GMP (cGMP) phosphodiesterase (McLaughlin et al., 1993), rhodopsin (Dryja et al., 1990), and the retinal degeneration slow (rds)/peripherin genes (Farrar et al., 1991). Mouse models for RP involving mutations in these three genes are available, and the major feature of the phenotypes is photoreceptor degeneration.

For the β subunit of cGMP phosphodiesterase gene, the retinal degeneration (rd) mouse is an animal model that has been extensively studied (Chader et al., 1988). The mutation in the rd mouse is recessive and results in no enzyme production (Pittler and Baehr, 1991). The photoreceptors develop normally up to the early part of the second week of life at which time photoreceptor degeneration starts. Degeneration is rapid and by postnatal day 17, virtually all the rod photoreceptors have disappeared (Carter-Dawson et al., 1978). No spontaneous mutations of the rhodopsin gene have been identified in animals. Transgenic technology has been used however to create several lines of mice expressing mutant rhodopsin (Humphries et al., 1997), simulating a form of autosomal dominant RP seen in humans (Dryja et al., 1990a). In one line of transgenic mice, 50% of the nuclei in the outer nuclear layer (ONL) had degenerated by 7 weeks.
Degeneration was virtually complete by 12 months, with very few photoreceptor nuclei remaining in the retina (Chang et al., 1993). In another line of homozygous rhodopsin deficient mice, rod outer segments lose their photoreceptors over 3 months (Humphries et al., 1997). In heterozygotes outer segments are present but shortened and disorganised in older mice. The rds/peripherin gene was first studied in the rds mouse (Sanyal et al., 1980). In homozygous rds mice, the outer segment of the photoreceptor appears normal. Beginning at about 3 weeks of age, the photoreceptor nuclei in the ONL start to disappear, and by about 12 weeks, 50% of the nuclei have been lost. The degeneration is virtually complete by 12 months (Sanyal et al., 1980). Heterozygous rds mice also demonstrate abnormal photoreceptor morphology. However, the rate of degeneration is much slower than in homozygotes, and the process is never complete (Hawkins et al., 1985).

Although these three animal models represented different basic mutations, subsequent cell death was remarkably similar, with a lack of inflammatory response despite the large number of degenerated photoreceptors. Even in the Royal College of Surgeons (RCS) rat model, where the expression of the primary genetic defect is in the retinal pigment epithelium, apoptosis was observed in the photoreceptors (Tso et al., 1994). In every case, photoreceptors were shown to die via the apoptotic pathway, as evidenced by the histological picture, by terminal deoxynucleotidyl transferase-mediated biotin-dUTP nick end-labelling assays (TUNEL technique) (Gavrieli et al., 1992), and/or by direct demonstration of retinal-DNA nucleosomal laddering by gel electrophoresis.

1.6.6 Proteins associated with various retinal degenerative diseases

Several genes causing inherited retinal degeneration have been cloned and shown to encode photoreceptor-specific proteins, as described below. Their identification was accomplished primarily by way of the candidate gene or positional candidate gene approach (section 1.1.3.2; eg. rhodopsin, Dryja et al., 1990) on the basis that mutations in such indispensable photoreceptor-specific genes would be likely to cause retinal dysfunction. Mutations in these genes can either cause RP, CSNB or macular degeneration as outlined in several excellent reviews (Daiger et al., 1994; Dryja and Li, 1995; Bird, 1995; Sullivan and Daiger, 1996). Such mutations can contribute to our understanding of both abnormal and normal visual processes and are summarised below.
1.6.6.1 Rhodopsin

The gene encoding rhodopsin was the first gene identified as a cause of retinal degeneration (Dryja et al., 1990) and remains the best characterised retinal gene. The rhodopsin protein is 348 amino acids long and is a member of the G-protein coupled receptor family, with seven membrane spanning domains. The protein is localised to the discs and plasma membranes of rod outer segments. Each human rod contains approximately $4 \times 10^7$ rhodopsin molecules accounting for 80% of the protein content of the outer segment (Nathans, 1992). Because of its abundance and its primary role in phototransduction, rhodopsin was an obvious choice as a possible cause of retinal degeneration.

Close to 100 mutant alleles of the rhodopsin gene have been identified as causes of RP (Dryja and Li, 1995, review). Almost all are dominant and most of these are missense mutations, although a few are frame-shift mutations resulting in truncation of the carboxy terminus or in-frame deletions. Only three recessive alleles have been reported to date (Rosenfield et al., 1992; Kumaramanickavel et al., 1994). The mechanism(s) by which recessive or dominant rhodopsin mutations lead to photoreceptor degeneration have not yet been established. Two of the three recessive alleles, a nonsense mutation and a splice-site mutation, are not likely to encode a functional rhodopsin (Rosenfield et al., 1992), hence it appears that photoreceptor viability cannot be maintained without functional rhodopsin. The pathogenic mechanism for the third recessive allele (a missense mutation) is not known (Kumaramanickavel et al., 1994).

For dominant mutations, haploinsufficiency is unlikely to be the pathogenic mechanism since carriers of at least one apparent null allele are phenotypically normal (Rosenfield et al., 1992). Furthermore, transgenic mice expressing dominant missense mutations in a genetic background with two wild-type murine rhodopsin alleles develop photoreceptor degeneration. Thus, most dominant rhodopsin mutants appear to be gain-of-function or dominant-negative alleles. The majority affect either the normal folding of the molecule, its capacity to bind 11-cis-retinal, its effective transportation from the endoplasmic reticulum following synthesis or its incorporation into rod outer segments. (Dryja and Li, 1995).

Another proposed pathogenic mechanism for certain dominant mutations invokes the 'constant equivalent light' model. This was inspired by studies of mutants in which the lysine residue at codon 296 was altered. Lys296 is the attachment site for 11-cis-retinal and also participates in holding opsin in an inactive conformation when it is not bound to a chromophore (Robinson et al., 1992). Mutations altering this residue, such as Lys296Glu and Lys296Met, encode opsins that cannot incorporate 11-cis-retinal.
1. Introduction

and that constitutively activate transducin in vitro. Such mutations cause cell death by over-stimulating the phototransduction pathway, analogous to exposure to constant light which is known to cause photoreceptor cell death. However, Lys296Glu opsin in transgenic mice is inactivated by phosphorylation and binding to arrestin. Thus, Lys296Glu opsin in vivo does not activate the phototransduction cascade and therefore cell death must be mediated by a mechanism other than over-stimulation (Li et al., 1995).

1.6.6.2 Rod cGMP phosphodiesterase α and β subunits (PDEA and PDEB)

cGMP phosphodiesterase (PDE), the third component of the cascade is a holoenzyme consisting of two large subunits α and β, which share considerable homology with each other, and two small, identical γ subunits. Nonsense and missense mutations in the genes coding for the α and β subunits have been identified as the cause of a small percentage of recessive RP (McLaughlin et al., 1995). The nonsense mutant alleles in either gene would result in the truncation of the putative catalytic domains in both subunits (Lipkin et al., 1990) and therefore are likely to be null alleles.

Such mutations in the PDEB gene and the resulting phenotype in human RP families parallel those found in the rd (retinal degeneration) mouse model of RP (Farber, 1995), where loss of cGMP PDE activity leads to persistently elevated cGMP levels which are cytotoxic and cause photoreceptor death. Similarly, a nonsense mutation in the β-PDE gene has been found in the Irish Setter dog (Suber et al., 1993). Mutations in PDEB are also associated with autosomal dominant congenital stationary night blindness where the retarded hydrolysis of cGMP is presumed to be less severe, resulting only in reduced sensitivity of the photoreceptors to light (Gal et al., 1994).

1.6.6.3 cGMP-gated channel

The cGMP-gated cation channel resides on the rod outer segment plasma membrane and is the final component of the phototransduction cascade. It plays a key role in phototransduction by controlling the flow of Na⁺ and Ca²⁺ into the outer segment in response to light induced changes in cGMP concentrations. A significant fraction of the rod Na⁺/Ca²⁺ -K⁺ exchanger copurifies with the channel as measured by western blotting suggesting that the channel can interact with the exchanger under certain conditions (Molday and Molday, 1998). Mutations of the gene encoding the α subunit of the channel have been found in a few families with autosomal recessive retinitis pigmentosa (Dryja et al., 1995). Since the mutations are either obviously null (eg. frameshifts early in the reading frame) or encode a channel protein that functions poorly, it appears that a paucity or absence of functional channels is deleterious to rod photoreceptors.
The peripherin/RDS protein is a 39 kDa integral membrane protein localised to the rim region of outer segment disks, hence the name peripherin (Molday et al., 1987). Its function is to anchor the discs to the cytoskeleton. RDS was originally cloned as the gene underlying the semidominant mutation observed in retinal degeneration slow (rds) mice (Travis et al., 1989). Homozygous rds mice are affected with non-development of photoreceptor outer segments, followed by slow degeneration of the rod and cone cell bodies. In rds heterozygotes, the outer segments although present are shortened and disorganised, with very slow photoreceptor degeneration (Hawkins et al., 1985). The mutant allele at the mouse rds locus is caused by the insertion of a repetitive element into an exon (Travis et al., 1989). This has shown to be a null allele by transgenic complementation (Travis et al., 1992), with the phenotype in the heterozygote due to haploinsufficiency.

To date more than 30 mutations in the peripherin/RDS gene have been implicated in various retinal degenerative diseases. Whereas some peripherin/RDS mutations cause ADRP (Gruning et al., 1994), others result in macular degeneration and pattern dystrophies that are classified as butterfly-shaped pigment dystrophy (Nichols et al., 1993), fundus flavimaculatus (Weleber et al., 1993) and cone-rod dystrophy (Jacobson et al., 1996). Different diagnoses have been assigned even to relatives with the same mutation. Objective measures of photoreceptor function, such as ERGs, rod and cone sensitivities and rates of dark adaptation, show similar abnormalities regardless of clinical diagnosis. Hence, it is likely that the clinical heterogeneity is due not to variations in the underlying physiological defects but to epigenetic factors modulating the response of the retina to the primary genetic defect.

Like peripherin/RDS, ROM-1 is an integral membrane protein which is abundant at the rim region of the outer segment discs. Both ROM-1 and peripherin form covalently linked homodimers which interact non-covalently to form a tetrameric complex (Bascom et al., 1992). Three RP families have been reported which segregate mutations in both the ROM-1 gene and the peripherin/RDS gene ('digenic' inheritance) (Kajiwara et al., 1994). The ROM-1 mutations were frameshift mutations early in the reading frame and are likely to be null mutations. All three families segregated the same peripherin/RDS allele, the missense change Leu185Pro. Affected individuals were double heterozygotes for mutations in both genes, while heterozygote carriers of either the ROM-1 or peripherin/RDS mutation were phenotypically normal. The missense peripherin/RDS
mutant, L185P, which segregates with instances of digenically inherited RP, has been reported to be conditionally defective with respect to its subunit assembly (Goldberg and Molday, 1996). Unlike wild-type peripherin/RDS, the L185P mutant did not form native-like homotetramers on its own; however this mutant could assemble with wild-type ROM-1 to form a structurally normal heterotetrameric complex. This findings provide a novel molecular-based rationale for the unusual digenic disease inheritance pattern and also offer insights into regions of peripherin/RDS and ROM-1 which contribute to subunit-subunit interactions. Recent results on comparative analysis of the rom-1 knockout mouse (Clarke et al., 1998) and the rds mouse leads to the conclusion that peripherin/RDS is the dominant subunit required for disk morphogenesis. ROM-1 is relegated to a more minor role, perhaps enhancing the stability of the outer segment and/or fine tuning the structure of disks.

1.6.6.6 Retinal guanylate cyclase (RETGC-1)

Missense and frameshift mutations in RETGC-1 have recently been discovered to cause Leber's congenital amaurosis (LCA), an autosomal recessive condition characterised by blindness at birth or shortly thereafter, due to either a defect in photoreceptor development or early photoreceptor degeneration (Perrault et al., 1996). As cGMP production in photoreceptor cells is abolished in this disease, the excitation process is impaired due to consistent closure of the cGMP-gated channels with hyperpolarisation of the plasma membrane. The cGMP concentration cannot be restored to the dark level, leading to a situation equivalent to consistent light exposure during photoreceptor development, and consequent toxicity. An animal model for human Leber's congenital amaurosis was recently reported in the retinal degeneration (rd) chicken (Semple-Rowland et al., 1998). It was shown that the low levels of cGMP in the rd chicken were a consequence of a null mutation in the photoreceptor guanylate cyclase gene, which prevented phototransduction and affected survival of rods and cones but did not interfere with normal photoreceptor development. Recently the RETGC-1 gene was also implicated in causing dominant cone-rod dystrophy (CORD6) where missense mutations were identified in four different dominant cone-rod dystrophy families (Kelsell, et al., 1998). Hence depending on the location of the mutation in the RETGC-1 gene, it can result in either an autosomal dominant cone-rod dystrophy or an autosomal recessive retinal degeneration (Leber disease) (Perrault et al., 1998).
1.6.6.7 Guanylate cyclase activating protein (GCAP1)

A mutation in guanylate cyclase activator IA (GUCA1A), the gene for guanylate cyclase activating protein (GCAP1; section 1.6.4)) was recently reported, in a family with autosomal dominant cone dystrophy mapping to a new locus on chromosome 6p21.1 (Payne et al., 1998). A screen of the GCAP1 gene in affected individuals showed a single base pair missense mutation (A—>G) at codon 99 in exon 2 of this gene generating a tyrosine-to-cysteine change in the GCAP1 protein. This change was proposed to disrupt the GCAP1 protein thereby preventing calcium binding and consequently interfere with activation. The resulting effect on cGMP production would modify the number of open cGMP gated cation channels, and could explain the ultimate demise of cone photoreceptor cells.

1.6.6.8 Retinal rod-specific ATP binding cassette transporter (ABCR/RIM)

The abundant high molecular weight rim protein identified and characterised by Papermaster and colleagues (1978) in frog photoreceptors is a member of the superfamily of ABC (ATP binding cassette) transporters (Uling et al., 1997). The rod ABC protein consists of two structurally related halves, each of which contains a multiple membrane-spanning domain followed by a cytoplasmic ATP binding cassette. The protein is abundantly expressed in the outer segments of rod photoreceptors, but is absent in cone cells. The ABCR protein and the rim protein were found to be one and the same (ABCR/RIM protein) after amino acid sequence analysis of human, bovine and mouse proteins (Uling et al., 1997). Mutations in the ABCR gene were found to be responsible for Stargardt's disease, an autosomal recessive macular dystrophy with a juvenile onset (Allikmets et al., 1997). Recent studies have suggested that some mutations in the ABCR gene are associated with age-related macular degeneration (Allikmets et al., 1997a). Homozygous and compound heterozygous mutations in the ABCR gene resulting in null alleles have also been implicated in autosomal recessive RP and cone-rod dystrophy (Martinez-Mir et al., 1998, Cremers, et al., 1998). The function of the ABCR/RIM protein is not currently known. A related protein, ABC1, has been implicated in the engulfment of cell bodies after apoptosis (Luciani and Chimini, 1996). It is possible that the ABCR/RIM protein is involved in the transport of retinal derivatives, phospholipids, peptides or other endogenous substrates across the disk membrane (Molday 1998).
1.6.6.9 Cone-rod homeobox (CRX)

Mutations within the photoreceptor-expressed gene CRX (cone-rod homeobox gene) have been reported (Freund et al., 1997; Swain et al., 1997) to cause dominant cone-rod dystrophy at the CORD2 locus on chromosome 19q13 (Evans et al., 1994). Mutations in the CRX gene were found to cause autosomal dominant cone-rod dystrophy (adCRD) either by haploinsufficiency or by a dominant negative effect, since both over- and underexpression of rhodopsin, one of the targets of CRX activation, can cause photoreceptor degeneration (Humphries et al., 1997). In addition de novo CRX mutations have been found in isolated cases of dominant Leber congenital amaurosis (LCA) (Freund et al., 1998) and a novel homozygous mutation in the homeodomain of CRX resulted in autosomal recessive LCA (Swaroop et al., 1999). CRX belongs to the orthodenticle homeobox (OTX) family of homeobox genes and shows photoreceptor-specific expression and regulates photoreceptor differentiation (Furukawa et al., 1997). CRX binds specifically to conserved sequences upstream of several photoreceptor-specific genes, including the opsins, and activates transcription of interphotoreceptor retinoid binding protein, arrestin, and β-phosphodiesterase (Chen et al., 1997). A recent study extended the range of phenotypes that may result from mutations in CRX, to include dominant retinitis pigmentosa and dominant Leber congenital amaurosis (Sohocki et al., 1998). Different CRX mutations may produce significantly different retinal disease phenotypes, and these diseases also have significantly different ages at onset. Thus, like those in peripherin/RDS (section 1.6.6.4), mutations in CRX are associated with several forms of retinal degeneration.

1.6.6.10 Cellular retinaldehyde-binding protein (CRALBP)

Cellular retinaldehyde-binding protein (CRALBP) is not expressed in photoreceptors but is abundant in the RPE and Muller cells of the neural retina, where it carries 11-cis retinol and 11-cis retinaldehyde. In RPE, CRALBP is proposed to play a role in the visual cycle. It preferentially binds 11-cis retinol, and promotes its oxidation to 11-cis retinaldehyde for rhodopsin regeneration rather than pushing it into storage (via esterification), an hypothesis supported by the fact that oxidation by 11-cis retinol dehydrogenase is increased about threefold and esterification diminished by an order of magnitude when CRALBP is present (Saari et al., 1994). Molecular genetic analysis of a consanguineous pedigree segregating for non-syndromic autosomal recessive retinitis pigmentosa (arRP) at chromosome 15q26 indicated that the affected siblings were carrying a mutation in RLBPl (retinaldehyde-binding protein), the gene encoding CRALBP (Maw et al., 1997). The loss of function of the mutated protein was shown by
reduced solubility and its inability to bind 11-cis retinaldehyde (Intres et al., 1994). The CRALBP defect could cause 11-cis retinol to fail to act as an efficient substrate for microsomal oxidation by 11-cis retinol dehydrogenase, thus preventing the regeneration of rhodopsin. The CRALBP mutation may therefore result in a depletion of 11-cis retinaldehyde and destabilise rod and cone opsins which account for 70% of outer segment protein and hence photoreceptor outer segments leading to a slow retinal degeneration. Alternatively, lack of visual pigment may lead to a constant depolarisation of the photoreceptor cell's outer-segment membranes, thereby exhausting it by a high energy demand and calcium load.

1.6.6.11 RPE65

The RPE expresses a tissue-specific and highly conserved 65 kD protein (RPE65) present at high levels in vivo (Hamel et al., 1993). RPE65 has been localised to human chromosome 1p31 (Hamel et al., 1994). By linkage analysis, certain autosomal recessive childhood-onset, severe retinal dystrophies (arCSRDs) were localised to the interval occupied by RPE65 and analysis of the RPE65 gene in these families revealed five potentially disease-causing mutations (Gu et al., 1997). In addition, candidate gene analysis of RPE65 identified two null mutations in patients with autosomal recessive Leber's congenital amaurosis (LCA) (Marlhens et al., 1997). Like the retinal dystrophies with a later onset, LCA is genetically heterogeneous; mutations have been also found in retinal guanylate cyclase, RETGC-1 (section 1.6.6.6). Although the precise function of RPE65 is not known, a role in vitamin A metabolism is suspected. Evidence for this is provided by its biochemical association with serum retinol-binding protein (Bavik et al., 1992) and with the RPE-specific 11-cis retinol dehydrogenase (Simon et al., 1995), both proteins which are concerned with retinoid metabolism. RPE65 has no significant homology with any other protein but its initial expression coincides with the appearance of rod photoreceptor outer segments, consistent with a possible role in the visual cycle (Hamel et al., 1993). Recent studies of Rpe65-deficient mice Rpe65−/− demonstrated severe changes in visual function as shown in their retinal physiology and biochemistry (Redmond et al., 1998). There was an evident lack of 11-cis-retinal chromophore together with an over-accumulation of all-trans-retinyl esters in the RPE which implicated a block in the RPE visual cycle.
1.6.6.12 Bestrophin

A novel retina-specific candidate gene named bestrophin was recently mapped and cloned by Petrukhin and colleagues (1998), and was shown to be responsible for Best's macular dystrophy (BMD), an autosomal dominant, early-onset form of macular degeneration in which the primary defect occurred at the level of the retinal pigment epithelium (RPE). Five independent disease-specific mutations were identified in Swedish and Dutch families and expression studies provided evidence that mutations within this candidate gene are a cause of BMD. The 3' UTR of the candidate gene contained a region of antisense complementarity to the 3' UTR of the ferritin heavy-chain gene (FTH1). Structural analysis of the bestrophin sequence predicted a protein with two putative transmembrane hairpins. Four of the five missense mutations affected the residues in two hydrophillic domains whereas one mutation altered a highly hydrophobic tyrosine located in the second transmembrane segment in close proximity to the membrane surface (Petrukhin et al., 1998). The functional consequences of the bestrophin mutations could not be determined since the function of bestrophin could not be inferred from amino-acid sequences. However from the known BMD pathology (abnormal accumulation of lipofuscin in RPE cells associated with progressive macular degeneration) and the RPE-specific expression of Best's disease gene (designated VMD2), it is suggested that bestrophin may mediate the transport or metabolism of an essential component of lipofuscin granules. In another report (Marquardt et al., 1998) the same gene was identified, and characterised as a novel gene TU15B, located close to the FTH1 gene within the critical region containing the Best's disease gene. The TU15B gene was expressed exclusively in the RPE, suggesting that the protein played a functional role in the RPE. Strong evidence was also provided showing that mutations in the TU15B is associated with Best's disease and that this gene represented the VMD2 gene.

1.7 The human X chromosome

The human X chromosome contains approximately 150 Mbp of DNA, which constitutes about 5% of the haploid genome. In terms of genetic organisation, the X appears to be quite similar to an average autosome, and genes on the X chromosome are no more likely to be connected with sexual differentiation than are autosomal genes. The X chromosome is one of the most extensively studied of all human chromosomes, a result of the wide interest in X-linked diseases and in the phenomenon of X chromosome inactivation (Lyon, 1988). Both these features are related to the different dosage of X in males and females; because males have only a single X, recessive diseases tend to be revealed, which accounts for the large numbers of known X-linked diseases. As a result, the X chromosome was the first to have a genetic map based on restriction fragment
length polymorphisms (RFLPs) and systematic approaches to physical coverage were undertaken that have expanded as part of the Human Genome Initiative. By 1993, 200 structural genes had been assigned to the X chromosome, together with 400 anonymous DNA segments (Sclessinger et al., 1993). Genes for two X-linked diseases (chronic granulomatous disease and Duchenne muscular dystrophy) were the first to be isolated by positional cloning (Royer-Pokora et al., 1986; Monaco et al., 1986).

1.7.1 Mapping X-linked diseases

The mapping of disease genes on the X chromosome is facilitated by their characteristic phenotypic pattern (female carriers, affected male offspring) and by the manifestation of maternal meiotic recombinations between X chromosomal loci in male offspring. X-linked diseases have features that facilitate positional cloning. Chromosomal assignment, which can be an arduous task for rare autosomal diseases, is obvious because of the inheritance pattern. For about ten diseases, rare affected females have been found with balanced X-autosome translocations. In these patients, the normal X chromosome is generally inactive, and the translocated X active, because of a selection process operating in early embryogenesis against cells carrying an inactive translocated X. Translocations that have a breakpoint within a gene will lead to expression of the corresponding disease, as the uninterrupted copy on the normal X is inactive. Such translocations have provided precise localisation for the relevant disease genes that could be confirmed by linkage analysis in affected families and have been instrumental in the cloning of several X-linked disease genes (Mandel et al., 1992), such as Duchenne muscular dystrophy (DMD) (Worton and Thompson, 1988).

Although males have only one X chromosome, rare male patients exist who survive with deletions (often detected cytogenetically) encompassing several megabases of DNA. In most cases the lack of function for genes in the deleted region results in a contiguous gene syndrome. In other words, several diseases are associated in a single patient, allowing very accurate mapping of the relevant genes if a series of overlapping deletions are available for analysis (Ballabio, 1991). This was first observed in the case of the BB deletion encompassing part of the DMD locus and genes for CGD (CYBB), McLeod syndrome (XK) and retinitis pigmentosa-3 (RP3) in Xp21 (Francke et al., 1985). The BB deletion was instrumental for the cloning of the DMD and CYBB genes in 1986, and XK more recently (Ho et al., 1994). Such regions have relatively low gene densities, or the deletions would be lethal. In the Xp21.2 region this can be accounted for in part by the huge size of the DMD gene, but more generally it illustrates the great heterogeneity in gene density throughout the genome, with gene-poor regions (in general AT-rich, Giemsa dark bands) and gene-rich regions (GC-rich, Giemsa light bands) (Bickmore and Sumner, 1989).
1.7.2 Mutations in X-linked diseases

The spectrum of mutations and incidence of severe X-linked diseases is due to the constant input of new mutations and directly reflects the mutational sensitivity of a given gene. Since affected males do not reproduce in cases of severe disease, at each generation the number of mutations decreases by one third (since males have one-third of the X-chromosomes in the population) and the particular mutation becomes extinct after a few generations. This does not apply to mutations with mild or no effect on reproductive fitness such as those responsible for glucose-6-phosphate dehydrogenase (G6PD) deficiency, colour blindness, or some cases of mild haemophilia A or B.

There is a striking difference in deletion frequency for various diseases (reviewed in Mandel et al., 1992). In X-linked ichthyosis, 80% to 90% of the mutations are large deletions encompassing the entire gene. Many of these are 1.9-Mb deletions that result from unequal recombination between flanking low-copy repetitive element. Duchenne muscular dystrophy is another disease with a high frequency of partial deletions (60% to 70%) and a significant level of partial duplications (6% to 7%). In part, this may be due to the huge target size of the dystrophin gene (2.4Mb). For most other diseases analysed thoroughly such as haemophilia A and B and HPRT deficiencies, the frequency of deletions or other re-arrangements detectable by Southern blots is in the order of 5% to 15%.

Expansion of a trinucleotide repeat sequence is a disease-causing mechanism (Caskey et al., 1992), originally described as the sole type of mutation in two X-linked diseases. The fragile X mental retardation syndrome is caused by an unstable expansion of a CGG repeat in a 5' exon of the gene FMR-1, which is correlated in patients with abnormal methylation of the adjacent CpG island leading to loss of expression of the gene transcript (Verkerk et al., 1991). In spino-bulbar muscular atrophy (SMA), the mutation is a more moderate expansion of a CAG repeat in the NH2-terminal coding region of the androgen receptor gene (AR) (La Spada et al., 1991) while other heterogeneous mutations in AR result in the completely different phenotype of testicular feminization.

1.7.3 The genetic, physical and transcriptional map of the human X chromosome

A unified genetic, physical and transcriptional map of the human X chromosome is being assembled through a concerted, international effort. Two large groups (Nagaraja et al., 1996 and Crollius et al., 1996) have recently published integrated maps specifically for the X chromosome, owing to its poor representation compared with most of the
autosomes in the whole-genome physical maps (Chumakov et al., 1995; Hudson et al., 1995). This is principally due to the low representation of the X chromosome in the CEPH YAC library used to construct these maps, and the fact that they are based on microsatellite STSs which are lower in frequency on the X chromosome (Dietrich et al., 1996). Portions of Xp11 have been found to be unstable and underrepresented in YAC libraries (Boycott et al., 1996). The use of panels of classical somatic cell hybrids and radiation-reduced hybrids complement genetic and high-resolution physical mapping and are essential in sublocalising new markers and confirming the location of YAC end-clones within a region that was particularly unstable in the yeast. Recently 101 DNA markers were localised on a hybrid mapping panel consisting of 10 radiation-reduced and 4 classical somatic cell hybrids which contain breakpoints in the p11 region of the human X chromosome (Boycott et al., 1997). These markers included polymorphic markers, STSs, ESTs, pseudogenes and genes providing an integrated map of Xp11 with particular emphasis on the central portion of Xp11. The X chromosome integrated database (IXDB) a repository for physical mapping data of the human X chromosome also helps support the construction of an integrated physical, genetic, transcript and sequence map of the X chromosome (Leser et al., 1998).

YAC contigs with an average clone depth of at least four-fold coverage now span more than 80% (125 Mb) of this 150 Mb chromosome (Crollius et al., 1996), with an average inter-STS resolution of 75 kb and a current total of around 2091 STSs in cognate YACs (Nagaraja et al., 1997). These STSs include 97 ESTs and 190 gene-specific STSs from known genes, as well as 192 dinucleotide and 38 tri- and tetranucleotide repeat markers, hence allowing integration of this YAC/STS map with transcriptional and genetic maps. Numerous YAC contigs have also been established in smaller regions surrounding specific disease genes through positional cloning strategies, and contigs are rapidly being linked by common STSs and clones such that the 'consensus' physical map is virtually at the point of completion (Nelson et al., 1995). The gaps that remain are due to the problems associated with YAC cloning, that is regions that are unstable or unrecovered in cloned DNA, sequence elements that are repeated along the X and are unstable in clones, and clones that are chimeric or contain large internal deletions. Closure will most likely accrue from efforts to construct higher resolution maps in bacterial cloning systems (eg. PI, BAC and PAC), which provide comparable supplements to delimited zones of poor YAC coverage.

Several yeast artificial chromosome (YAC) contigs from the Xp11.23-22 region have been reported, however none of these maps provide a full coverage. A YAC contig encompassing the distal part of the Xp11.23 region between the loci TIMP1 and OATL1 has been reported (Coleman et al., 1994; Knight et al., 1994). A zinc-finger gene cluster has been detected on the distal OATL1 YAC (Knight et al., 1994), which also contains breakpoints for synovial sarcomas (Chand et al., 1995) and the retinal-expressed
sequences MG61, MG81, MG21 and MG44 on the proximal OATL1 YAC (Geraghty et al., 1993). Other YAC contigs encompassing the region between TIMP1 and the WASP gene locus have also been reported (Derry et al., 1994; Kwan et al., 1995). However, due to deleted or chimeric YACs and gaps in the YAC contigs, the physical distances between the OATL1 pseudogene cluster and the GF-1 locus (Zon et al., 1990) as well as the distances between the SYP locus (Ozcelik et al., 1990) and the VNTR locus DXS255 remained uncertain. Recently a long-range map of the Xp11.23-22 region was established concomitantly with a YAC-cosmid contig for the entire region flanked by the loci TIMP-1 and DXS146 (Schindelnauer et al., 1996). This pulsed-field gel electrophoresis (PFGE) map revealed a CpG cluster in a 1100-kb region from which a sequence-ready map with a high density of expressed sequence tags (ESTs) and polymorphic markers was constructed.

There are 6 major genetic maps for the X chromosome (Murray et al., 1994, Matisse et al., 1994, Donnelly et al., 1994, Wang et al., 1994, Fain et al., 1995 and Dib et al., 1996) consisting of RFLP's and microsatellite markers. Basic features of these maps are summarised in Table 1.5. The maps are described as 'framework' (F) or 'comprehensive' (C) depending on the statistical support for marker order, 1000:1 (lod 3) is required for a framework map.

### Table 1.5 Genetic maps for the X chromosome

<table>
<thead>
<tr>
<th>Reference</th>
<th>Marker loci</th>
<th>Genetic map length</th>
<th>Resolution</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matise et al., 1994</td>
<td>57 mixed markers (24 framework)</td>
<td>191 cM</td>
<td>8.9 cM (F)</td>
<td>Comprehensive map; Multimap algorithm</td>
</tr>
<tr>
<td>Donnelly et al., 1994</td>
<td>62 PCR-based markers (30 framework)</td>
<td>236 cM</td>
<td>3.7 cM (C)</td>
<td>Comprehensive map; integration with cytogenetic map</td>
</tr>
<tr>
<td>Murray et al., 1994</td>
<td>150 microsatellites (35 framework)</td>
<td>247 cM</td>
<td>8.2 cM (F)</td>
<td>Framework map; integration with cytogenetic map</td>
</tr>
<tr>
<td>Wang et al., 1994</td>
<td>270 mixed markers (104 framework) (50 PCR-based)</td>
<td>211 cM</td>
<td>-</td>
<td>Comprehensive map; integrates genetic and physical data from XCW4 (Schlessinger et al., 1993)</td>
</tr>
<tr>
<td>Fain et al., 1995</td>
<td>243 mixed markers</td>
<td>190 cM</td>
<td>3.2 cM (C)</td>
<td>2D comprehensive map integrates maps of Matise et al., Gyapay et al., and Murray et al., (1994)</td>
</tr>
<tr>
<td>Dib et al., 1996</td>
<td>216 CA repeats</td>
<td>198 cM</td>
<td>~ 2 cM (F)</td>
<td>Framework map average marker. HET 65%</td>
</tr>
</tbody>
</table>
These genetic maps show a high degree of overall consistency and the genetic lengths are in good agreement with that postulated by Morton in 1991 (220cM). Differences in the resolution and genetic length among maps reflect differences in error screening, number of markers analysed, ordering criteria and map format. The higher resolutions of the 2D map (3.2cM; Fain et al., 1995) and the map of Wang et al., (1994) reflect the high marker density provided by map integration, and make them extremely valuable for positional cloning projects where it is more useful to know and refine the probable marker order in a dense map, both to refine disease gene intervals and to assist in locus assignment for X-linked disorders that display genetic heterogeneity (eg. retinitis pigmentosa, Teague et al., 1994; mental retardation; Lubs et al., 1996). The high-confidence framework maps serve as a guide to low resolution linkage mapping. Thus the human X chromosome genetic map is essentially complete in terms of marker density. Whole chromosome efforts to generate further markers would now offer diminishing returns; all that remains is consolidation of marker order and distances across the chromosome, which will most likely be established by integration with the developing physical map. With the YAC map near to completion, X chromosome mapping is now shifting towards STS-based map integration, EST assignment and long-range sequencing (Nelson et al., 1995).

Since there are an estimated 50,000-100,000 gene in the human genome, 2,500-5,000 would be expected to be present on the human X chromosome. The identification of candidate disease genes usually requires a targeted search of the region of interest by methods like direct selection or exon trapping (section 1.5) but systematic functional analysis of the genome should eventually provide a catalogue of all the candidate genes in a region. One such strategy initiated by Adams and coworkers is designed to generate expressed sequence tags (ESTs) by partial DNA sequencing of randomly selected human brain cDNA clones (Adams et al., 1991). Expressed sequence tags (ESTs) can in principle serve as specialised sequence tagged sites (STSs) to assemble a functional map of the human genome. The strategy of physically linking ESTs to the nearest genetic linkage markers should provide specific candidate genes for the X-linked diseases associated with these markers or loci (Mazzarella and Srivastava, 1994). Recently 59 ESTs were mapped in 31 intervals spanning the X chromosome (Gianfrancesco et al., 1997) where a somatic cell hybrid panel provided a resolution of about 5Mb for the short arm and 4Mb for the long arm, and further resolution was obtained using YACs mapped in localised contigs. ESTs provide a reliable strategy for the mapping of transcription units in YAC contigs and an integration of physical, transcriptional and genetic maps. Precise mapping information together with sequence analysis may provide further evidence as to the function of each of the EST genes and their relationship to disease states.
Until now, genomic sequencing on the X has been limited to a small number of genes. With the advent of computer algorithms capable of identifying candidate-coding sequences in genomic DNA (Uberbacher and Mural, 1991), large-scale sequencing projects become more attractive. They will probably first involve the disease-rich region in Xq28 or the deletion-prone regions in the DMD gene whose large introns may contain genes. A recent sequence-based exon prediction strategy to investigate the organisation of genes around the synaptophysin (SYP1) locus at Xp11.23, revealed a gene-rich area containing novel genes (Fisher et al., 1997), where the density of transcribed sequences in this area was >80% comparable to that found in Xq28. The construction of a 1.5Mb BAC contig was able to provide total physical coverage of the Xp11.23 region with 28 clones that are suitable for large-scale sequencing efforts (Boycott et al., 1998a). Sequencing may also be used in the study of recombination, for instance in the pseudoautosomal region, or in the analysis of control elements involved in the various phases of X inactivation.

The proximal short arm of the human X chromosome at Xp11.23-p11.22 (Figure 1.9) has been shown to code for several genes. These include the genes for the ubiquitin activating enzyme UBE1 (Lafreniere et al., 1991), the proto-oncogene A-raf-1 (ARAF1) (Huebner et al., 1986), the neuron-specific phosphoprotein synapsin 1 (SYN1) (Yang-Feng et al., 1986), properdin (PFC), a positive regulator of the alternative pathway of human complement (Goundis et al., 1989), the tissue inhibitor of metalloproteinases TIMP-1 (Kidd et al., 1989), the zinc finger proteins ZNF21 and ZNF81 (Hagemann et al., 1994), the ornithine aminotransferase pseudogenes at the OATL1 locus (Lafreniere et al., 1991a), the ets-related gene ELK1 (Rao et al., 1989), the erythroid-specific transcription factor GATA (Caiulo et al., 1991), the transcription factor for the enhancer μE3 (TFE3) (Lafreniere et al., 1991), synaptophysin (SYP) (Ozcelik et al., 1990), and recently the gene for Wiscott-Aldrich syndrome (Derry et al., 1994). The positions of several of these genes have been further refined through linkage analysis (Kirchgessner et al., 1991), somatic cell hybrid mapping (Berger et al., 1992) or physical mapping (Derry and Barnard, 1992; Hagemann et al., 1994; Knight et al., 1994). In the distal part of this region, several new genes have been identified, including MG21, MG44, MG61, MG81 (Geraghty et al., 1993) and DXS1011E (Parrish and Nelson, 1993). In addition, three eye disease genes have been mapped to the region between MAOA (Xp11.3) and DXS255 (Xp11.22), including retinitis pigmentosa, RP2 (Ott et al., 1990); X-linked congenital stationary night blindness, CSNB1 (Bech-Hansen and Pearce, 1993) and Aland Island eye disease, AIED (Glass et al., 1993) (Section 1.8).
Figure 1.9  The proximal short arm of the human X chromosome consensus map (not to scale) from Nelson et al., 1995.
1.7.4 Other aspects of the X chromosome

1.7.4.1 X-inactivation

X inactivation occurs early in mammalian development to transcriptionally silence one of the pair of X chromosomes in females, thereby achieving dosage equivalence with males (Lyon, 1961a). The X inactivation process is a unique cis-limited regulatory event that affects nearly an entire chromosome, turning off several thousands of genes. It is thought to occur in three sequential steps: firstly, initiation early in development in XX embryos at a site on the X chromosome called the X inactivation centre (Xic), which involves the cellular determination of how many (if any) and which X chromosomes to inactivate; secondly, spreading of the inactivation signal (propagation) along the length of the X chromosome selected to be the inactive X and thirdly, stabilisation and maintenance of the inactive state at individual gene loci on the X, so once the chromosome is inactivated, all descendants of that chromosome will be inactivated (Ballabio and Willard, 1992). DNA methylation at the CpG-rich islands of some X-linked genes has been implicated in this maintenance step (Riggs, 1990).

In addition to the genes from the pseudoautosomal region, which have long been anticipated to escape inactivation, genes from several other regions of the human X chromosome have been shown to escape inactivation and to be expressed from both the active and inactive X chromosomes, including not only genes which map close to the pseudoautosomal region but also genes which map to proximal Xp and proximal Xq regions, while genes known to map to intermediate locations are often subject to X inactivation (Disteche, 1995). Many of the human X-linked genes which map outside the major pseudoautosomal regions and escape X inactivation have functional homologues on the Y chromosome. Some, however, do not. For example, the UBE1 and SB1.8 genes escape inactivation but do not appear to have any homologues on the Y chromosome. Other genes such as the Kallman syndrome gene KAL1, and the steroid sulfatase gene STS do not have homologues on the Y chromosome, but these are non-functional pseudogenes. It is likely, therefore, that for some genes sex difference in gene dosage is not a problem and is tolerated (Disteche, 1995).

In the mouse, however, there are considerable differences in the pattern of X inactivation. For example, the human non-pseudoautosomal genes ZFX (Xp22.1), RPS4X(Xq13.1) and UBE1 (Xp11.23) all escape inactivation, but the murine homologues Zfx, Rps4 and Ube1X (which unlike the human UBE1 gene has a homologue on the Y) are all subject to X inactivation. Abnormal dosage of such genes may be involved in some cases of embryonic lethality and in other cases the distinct clinical phenotype of XO females (Turner syndrome) in humans, whereas XO mice are practically normal.
Three genes that escape X chromosome inactivation have been reported to be clustered within a 6 Mb YAC contig and STS map in Xp11.21-p11.22 (Miller et al., 1995). A novel X-linked gene, DX8237E, which mapped to within 20 kb upstream of UBE1 in Xp11.23 was found to have a different X-inactivation status (Coleman et al., 1996). Sequences in the vicinity of these closest mapped genes with discordant X inactivation may be important determinants of X inactivation status. In a recent study on the inactivation status of a 5.5Mb region at Xp11.21-p11.22, it was shown that 8 out of 23 expressed sequences escape inactivation (Miller and Willard, 1998). All 8 expressed sequences were located within a region of less than 370kb, where genes located both distal and proximal to this cluster are subject to inactivation, thereby defining a unique multigene domain that is transcriptionally active on the inactive X chromosome. There are significant implications for understanding the patterns of expression of X-linked genes in carrier females as counselling in such disorders is now based largely on the assumption that X-linked genes are subject to X inactivation.

Recent progress towards understanding X inactivation has been made through attempts to identify the X inactivation centre (Xic), a cis-acting locus on the X chromosome which is the master regulatory switch controlling this process (Rastan, 1994). A candidate for the Xic, termed XIST (X-inactive specific transcript) has been identified within the human minimum XIC interval (Brown et al., 1991). The XIST gene is unique among X-linked genes in being expressed exclusively from the inactive X chromosome. The equivalent mouse gene (Xist) was cloned through homology to the human gene. It was localised to the mouse Xic interval, and shown to be expressed exclusively from the inactive X chromosome (Brockdorff et al., 1991). This locus was mapped in both species to a region between the ectodermal dysplasia locus (EDA, or Tabby (Ta) in mouse) and the phosphoglycerate kinase gene (PGK1/pgk1 locus for man and mouse respectively) (Brown et al., 1991a, Brown, 1991), a region of 2.5 Mb in humans now largely covered by a YAC contig. The XIST gene includes several tandem repeats, the most 5' of which are evolutionarily conserved (Brown et al., 1992). The gene does not contain any significant conserved ORFs and thus does not appear to encode a protein, suggesting that XIST may function as a structural RNA within the nucleus. It is thought to be required for initiation of X chromosome inactivation, but not for its maintenance (Brown and Willard, 1994).

1.7.4.2 Human and mouse X-syntenic regions

The positioning of approximately 70 loci on the X chromosomes of both mouse and human has confirmed the prediction made by Ohno (1973) that X-linkage of genes is largely preserved in mammals (Willard et al., 1994). Comparative mapping of the X chromosome between human and mouse has so far revealed eight conserved regions
1. Introduction

(Blair et al., 1994). This results from several re-arrangements that must have occurred since the divergence of human and mouse over 100 million years ago. Detailed knowledge of these re-arrangements is useful in the elucidation of mouse models of human diseases, as an identical comparative map position is one criterion for inferring genetic homology between similar phenotypes.

There are four blocks of homology in the ~20Mb region of the proximal human X chromosome short arm; three lie in the proximal region of the mouse X chromosome, and the fourth lies in the distal region (Blair et al., 1995). The largest block runs from Cybb to Pfc and, on the human X chromosome, is flanked on the CYBB boundary by DMD and on the PFC boundary by GATA1. In the mouse, the Cybb boundary is flanked by Gata1/Tfe3 and the Pfc boundary by Lamp2. Recently the murine homologues of the loci for McLeod syndrome (XK), Dent's disease (C1CN5), and synaptophysin (SYN) have been mapped to the proximal region of the mouse X chromosome and positioned with respect to other conserved loci in this region using progeny from two separate Mus musculus X Mus spretus backcrosses (Blair et al., 1995), hence establishing a high-resolution genetic map for this ~3 cM interval. This mouse interspecific cross has also been used to demonstrate genetic linkage of Syn-1, Timp and Araf and also showed physical linkage, with Timp lying only 10kb from Araf, within an intron of the Syn-1 gene (Derry and Barnard, 1992). The localisation of the Syn1, TIMP and ARAF to within 70 kb of each other on the human X chromosome and the demonstration of an arrangement similar to that in mouse illustrate the strong conservation of the human and mouse X chromosomes.

The correspondence established between the maps of the two chromosomes has allowed the validation of mouse mutants as homologous models for human diseases with similar phenotypes, based on their equivalent map positions, and may be useful for positional cloning of such loci. For example, the Hyp, Ta and Mo mutants correspond to the genes for hypophosphatemic rickets (HYP), ectodermal dysplasia (EDA) and Menkes disease (MNK), respectively in humans. An interesting exception to the conservation of genes is the apparent absence in mouse of sequences homologous to the steroid sulfatase, Kallmann syndrome (KAL), MIC2 and GS1 genes, four genes that are closely located in Xp22.3 in humans (Ballabio and Willard, 1992). Deletion of this region in males results in a relatively mild phenotype, and it may have been similarly deleted in a mouse ancestor. Recently however, comparative mapping of the most distal band of the human X chromosome at Xp22.3 has extended the suggested model for the evolution of mammalian X chromosomes by Blair et al., (1994). Rather than supporting the assumption of eight conserved X chromosomal regions being rearranged during mammalian evolution, data on Xp22.3 loci implicate a much more complex sequence of events of massive reshuffling of X chromosomes during their evolution, leading to the divergence of the X chromosomes between man and mouse (Blaschke and Rappold, 1997). This
exceptional genetic instability of Xp22.3 has helped refine our understanding of mouse models for certain X chromosomal human genetic disorders and also permits the characterisation of additional breakpoints leading to the chromosomal rearrangements between man and mouse. It has also been found that all genes tested that are located on the short arm of the human X are autosomal in marsupials and monotremes. This is at variance with Ohno’s hypothesis, and suggests that the short arm was of autosomal origin and was added to the X chromosome in eutherian mammals (Watson et al., 1991).

1.8 Eye diseases on the X chromosome

The proximal short arm of the human X chromosome is home to a cluster of genes implicated in various inherited disorders of the eye (Figure 1.10). Several of these disorders share key symptoms with differing degrees of severity and overlapping map locations, and the growing examples of allelic heterogeneity suggest that different mutations in the same gene may have diverse phenotypic effects.

1.8.1 Norrie disease

Norrie disease (ND) is an X-linked recessive neurodegenerative condition with ocular atrophy as the most conspicuous sign. The disorder is characterised by congenital or early childhood blindness due to proliferative and degenerative changes in the vitreous body and the retina (Warburg, 1966). At birth, an intraocular mass is the most prominent sign, followed by ocular shrinkage with age. Extraocular features of ND are sensorineural deafness and mental disturbances, often with psychotic features, occurring in at least one-third of the cases, suggesting developmental and/or degenerative changes in other areas of the nervous system.

The initial localisation of the ND gene came from studies indicating close linkage between ND and the locus, DXS7, on Xp11.3-11.4 (Bleeker-Wagemakers et al., 1985). DNA analysis on several Norrie patients with complex and atypical phenotypes, including hypogonadism, growth retardation, immunodeficiency and epileptic seizures, revealed submicroscopic deletions, including the locus DXS7 (Gal et al., 1986), presumably representing a ‘contiguous gene syndrome’ whose phenotype reflects the deletion or disruption of more than one gene. Subsequently, MAOA and MAOB genes were found also to be deleted in these patients (Diergaarde et al., 1989). The absence of deletions and re-arrangements in the genomic DNA of typical Norrie patients, as well as normal MAOA and MAOB enzyme activities excluded the MAO genes as ND candidates (Sims et al., 1989). A recombination event between DXS7 and ND, but not between ND and OATL1, which is proximal to DXS7, established DXS7 as the distal boundary of the ND region (Lindsay et al., 1992). The delineation of the proximal
Figure 1.10 Genetic localisations of inherited eye disorders on proximal Xp (not to scale)
boundary of the ND gene was achieved through characterisation of a YAC clone (YL1.28), which contains DXS7, MAOA and MAOB. The retention of the proximal end of the YAC in one of the deletion patients indicated that all or part of the ND locus was contained within a 160 kb fragment of this YAC (Chen et al., 1992).

Positional cloning efforts by two groups identified a candidate cDNA of 1.8 kb (Berger et al., 1992; Chen et al., 1992a). Several small deletions within the corresponding gene were identified among typical Norrie disease patients (Chen et al., 1993). Berger et al., (1992a) used a cosmid probe from a cosmid contig extending 250 kb proximal to the MAOB gene to screen retina and cDNA libraries. In addition to mapping within the region of overlap in several microdeletion patients, this cosmid detected aberrant genomic DNA fragments on Southern blot analysis of non-syndromic ND males, and hence was considered likely to span at least part of the ND locus. A similar screening approach was undertaken by Chen et al., (1992a), using as a probe the 160 kb YAC subfragment containing the disease locus to screen human retinal cDNA libraries. Both groups identified an evolutionarily highly conserved cDNA whose expression was limited to retina, brain and choroid. Furthermore, the cDNA sequence was absent or disrupted in all Norrie-associated microdeletions.

Structural analysis of the ND gene revealed that it is 28 kb and consists of three exons, the first of which is not translated. An open reading frame of 399 bp is contained within exons 2 and 3 (Black and Redmond, 1994). A protein of 133 amino acids with a molecular weight of 15 kDa was predicted. The ND protein, termed norrin, or NDP, is cysteine rich with the cysteines in the C-terminal encoded by exon 3. Protein sequence comparisons revealed that this cysteine-rich domain (CT domain) shares homology with diverse extracellular proteins (Meindl et al., 1992). Since these proteins share an identical arrangement of disulphide bridges, this suggests that norrin may form dimers with itself, or possibly with other proteins. A signal peptide sequence at the N-terminal of the Norrie protein suggests that it is a secreted protein, involved in the regulation of cell differentiation and proliferation (Meindl et al., 1992). Computational modeling of NDP indicates that it shares homology and a predicted three-dimensional structure with a carboxy-terminal domain (termed the cysteine knot) found as a component in a variety of cysteine-rich neurotrophins such as transforming growth factor (TGF) and nerve growth factor (Meifinger et al., 1993). On this basis it has been predicted that the dimerisation of NDP may be an essential element in providing the recognition signals for targeting of neuronal/retinal connections.

A full spectrum of mutations (stop codons, missense mutations, and insertions, as well as deletions) have been identified in Norrie patients (Berger et al., 1992a; Meindl et al., 1992). A majority (over 70%) of missense mutations occur in the highly conserved 3' end of the gene encoding the C-terminal part of the protein. Expansion of a repeat motif (an insertion of five 2 bp repeats) located at the 5' end of the gene may affect the
transcription of the gene (Schuback et al., 1995). A survey of reported mutations has failed to identify any simple correlations between genotype and phenotype; however, it has recently been reported that missense mutations in non-conserved residues of the C-terminal domain lead to a more benign Norrie phenotype, presumably because they produce a partially functional protein (Meindl et al., 1995). This hypothesis is supported by the description of similar norrin mutations causing X-linked exudative retinopathy (XLFEVR), the symptoms of which overlap that of ND but are restricted to the eye and are less severe (Chen et al., 1993). It is possible that a temporal difference in pathogenesis may explain the clinical difference between these allelic disorders; the events causing ND may occur earlier in foetal life, so that by birth the disease process is more advanced.

In order to elucidate the cellular and molecular processes involved in Norrie disease, a mouse model has recently been generated by gene targeting (Berger et al., 1996). Hemizygous mice carrying a replacement mutation in exon 2 of the ND gene developed retrolental structures in the vitreous body and showed an overall disorganisation of the retinal ganglion cell layer with loss of photoreceptor outer segments. These ocular findings are consistent with observations in ND patients, thus this mutant mouse line should prove a faithful model for unravelling the early pathogenic events in this neurological disorder.

1.8.2 X-linked congenital stationary night blindness

Night blindness, or nyctalopia, is a symptom of a wide group of retinal disorders. It can be inherited and is either progressive or non-progressive (stationary). Congenital stationary night blindness (CSNB) comprises a group of disorders characterised by congenital onset of non-progressive night blindness, subnormal visual acuity, variable myopia and absence of pigmentary degeneration of the retina (Pearce, 1990). Three monogenic inheritance modes, autosomal dominant, autosomal recessive and X-linked recessive have been described (McKusick, 1988). In contrast with the autosomal forms, the X-linked form of CSNB (CSNBX) is frequently associated with high myopia, resulting in loss of visual acuity.

The basic defect in CSNB is known to lie in the retina. Electroretinography (ERG), which measures electrical responses of the outer and middle retina, has a key diagnostic role in CSNB (Miyake and Kawase, 1984). The ERG of affected males is abnormal, with an absent or severely reduced scotopic b-wave. Carrier status in females is not generally detectable clinically. Miyake et al., (1986) have classified CSNB of mixed genetic types into two clinical subgroups, termed complete and incomplete CSNB, on the basis of refractive error, ERG and dark adaptation responses. The complete type (CSNB1) lacks rod function and myopia is usually severe. The incomplete type (CSNB2) shows some rod function but also impairment of cone function. The refraction error ranges from
moderate myopia to hyperopia. In the study of Miyake et al., (1986) CSNB1 and CSNB2 appear to be distinct disease entities, which do not co-exist within the same pedigree. However, Pearce et al., (1990) suggested that CSNBX is a single disease entity with a highly variable clinical expression. It is hoped that the identification of the CSNB1 and CSNB2 genes will ultimately resolve the uncertainty that surrounds the clinical and genetic heterogeneity of X-linked CSNB as well as provide information about the basic defects in this X-linked retinal condition.

Genetic analyses over the last decade have established heterogeneity for CSNBX, implicating at least two distinct genes on the proximal short arm of the X chromosome. Linkage studies initially located CSNB1 close to DXS7 in Xp11.3 (Gal et al., 1989). Recent studies based on critical recombination events in a set of families with complete CSNB have localised the disease gene to the region between DXS556 and DXS8083 in Xp11.4-Xp11.3 (Boycott et al., 1998). A similar localisation was reported for a new CSNBX locus using haplotype analysis (Hardcastle et al., 1997), between RP2 and RP3 on Xp11.4 to Xp11.3 with flanking markers DXS556 and DXS8080 (interval of 5-6 cM). The interval in both reports overlaps the locus reported to contain the cone dystrophy (COD1) gene (section 1.8.4) indicating that both diseases may well be allelic. Tight linkage has been found between CSNB2 and markers in Xp11.3-11.22 (Musarella et al., 1992), and it has been placed proximal to DXS7 by a single recombination event (Bergen et al., 1994). The CSNB2 locus has recently been further refined using a set of families with incomplete CSNB between markers DXS722 and DXS255 (~1.2Mb) within Xp11.23 (Boycott et al., 1998).

Further genetic mapping information have been consistent with the two-locus hypothesis. Informative crossovers have placed CSNBX proximal to MAOA and distal to TIMP-1 (Bech-Hansen et al., 1992; Aldred et al., 1992), overlapping with proposed minimal regions for both the CSNB1 and CSNB2 genes. An affected male in a Dutch CSNB pedigree (Bergen et al., 1995) carried a double-recombinant chromosome that genetically mapped a novel CSNBX locus distal to DXS7, between markers DMD44 and DXS228 (disease interval Xp21.1-Xp11.3) providing evidence that CSNBX is genetically heterogeneous. This locus again overlaps with the CSNB1 minimal region. Since the CSNBX gene in this family is linked to the XLRP3 gene region distal to DXS7, they may be allelic disorders. Reported crossovers positioned CSNBX in a another Dutch family to the interval OTC-DXS1003 (Xp11.4-Xp11.23) (Berger et al., 1995) which overlaps that of the two previous localisations. This localisation is consistent with segregation of the CSNB1 gene reported by Boycott et al., (1998) between DXS556 and DXS8083. From such findings, it has been postulated that X-linked CSNB may be allelic with the RP2 and RP3 genes which similarly display genetic heterogeneity with map locations in Xp11.3-p11.22 and Xp21.1 (section 1.8.7.1) respectively.
However, the recently cloned RPGR gene responsible for XLRP3 which maps just distal to OTC, in Xp21.1 (Meindl et al., 1996) together with mapping information which localised the CSNB1 as proximal to OTC (Boycott et al., 1998); indicated that complete CSNB and RP3 are not allelic. The sublocalisation of RP2 to the region between DXS8083 and DXS6616 on the basis of critical recombination crossovers (Thiselton et al., 1996) and the recent isolation of the RP2 gene (Schwahn et al., 1998), together with mapping information localising the gene for incomplete CSNB (CSNB2) proximal to DXS6616-namely at DXS722 (Boycott et al., 1998) showed that CSNB2 and RP2 genes are also not allelic. Although neither XLRP2 and incomplete CSNB nor RP3 and complete CSNB are allelic, a mutation in the RPGR gene appears to be responsible for CSNB in at least one family (Herrmann et al., 1996). As in the autosomal forms of RP and CSNB which have been shown to be caused by different mutations in the rhodopsin gene on chromosome 3q (Dryja et al., 1993), it is possible that a few families with X-linked CSNB have disease-causing mutations in RP2 or RP3.

Recently a novel retina-specific gene mapping to the CSNB2 minimal region was characterised and found to have similarity to voltage-gated L-type calcium channel $\alpha_1$-sub-unit genes (Bech-Hansen et al., 1998). Mutation analysis of this new $\alpha_1$-sub-unit gene, CACNA1F, in 20 families with incomplete CSNB revealed six different mutations that are all predicted to cause premature protein truncation. Mutations in the same gene were reported by Strom and colleagues (1998), where mutation analysis revealed nine different mutations in ten families, including three nonsense and one frameshift mutation. These findings established that loss-of-function mutations in CACNA1F cause incomplete CSNB, making this disorder an example of a human channelopathy of the retina. These data also indicate that aberrations in a voltage-gated calcium channel, causing a decrease in neurotransmitter release from photoreceptor presynaptic terminals, are a frequent cause of CSNB2.

### 1.8.3 Aland Island Eye Disease (AIED)

Aland Island Eye Disease (AIED) is an X-linked form of ocular hypopigmentation, also known as Forsius-Eriksson syndrome (1964). This non-progressive condition is characterised in affected males by reduced visual acuity, reduced dark adaptation, colour vision abnormalities, infantile night blindness and high axial myopia. In AIED, female carriers are normal on clinical and electrophysiological examination. Initially, AIED was thought to be a variant of ocular albinism (OA1) on the basis of the fundal depigmentation in the original family described, and was given the locus symbol OA2, but this has since been disproved.
1. Introduction

Linkage analysis in the original AIED family enabled a localisation for the gene to be made to pericentromeric X (Alitalo et al., 1991). Subsequent refinement in a second AIED family placed the disease locus between DXS7 (Xp11.3) and DXSY1 (Xq21.3) with tight linkage to TIMP-1 and DXS255 (Schwartz and Rosenberg, 1991). Linkage analysis of a third family suggested a location between DMD and DXS255 (Glass et al., 1993). Assuming homogeneity of the three pedigrees, the combined information supported a localisation of the AIED gene between DXS7 and DXS255 (Glass et al., 1993). These findings contradicted a previous assignment of the AIED gene to Xp21.3-p21.2 based on a male patient with a contiguous gene deletion and complex phenotype of DMD, glycerol kinase deficiency, congenital adrenal hypoplasia and AIED. Further characterisation of the deletion in this patient showed that it spanned a region more proximal than those of several OA1 patients (Xp22), supporting the argument that OA1 and AIED are distinct (Pillers et al., 1990). The identification of dystrophin isoforms in retinal tissue has since raised the possibility that the impaired retinal neurotransmission in this patient was a direct consequence of dystrophin deficiency as observed in many DMD/BMD patients, and the phenotype was termed Oregon eye disease to distinguish it from AIED (Pillers et al., 1993).

It has been suggested that AIED is clinically indistinguishable from incomplete CSNB (Weleber et al., 1989). Studies of three AIED families (Alitalo et al., 1991; Schwartz and Rosenberg, 1991; Glass et al., 1993) defined recombinant chromosomes that together localised the AIED gene between DXS7 and DXS255, a region that overlaps with the minimal region for the CSNB2 gene (section 1.8.2). This lends further support to the notion that these conditions are most likely allelic.

1.8.4 X-linked progressive cone dystrophy (XLPCD)

X-linked progressive cone dystrophy (XLPCD) is characterised by mild to severe myopia, photophobia, night blindness, loss of visual acuity, abnormal colour vision and disturbed cone electroretinogram (ERG) (Pinkers and Deutman, 1987). The cone ERG is absent or severely attenuated in affected males, and the rod ERG may be abnormal. The severity of the fundus findings correlates with the degree of visual dysfunction and suggests an age-related progression of the condition (Hong et al., 1994). The retinas of some affected males may present with a tapetal-like sheen (Jacobsen et al., 1989). The clinical picture of heterozygous females ranges from asymptomatic to a wide spectrum of cone-mediated dysfunction; colour vision testing has been reported to allow detection of 87% of obligate carriers (Keunen et al., 1990). XLPCD can be distinguished from another form of X-linked cone dysfunction, blue cone monochromatism (BCM). In contrast with XLPCD, BCM is a congenital, non-progressive disorder in which the function of cones that are sensitive for the blue part of the light spectrum remain intact.
Since progressive X-linked cone-rod dystrophy (COD1) is an extremely rare disease entity, only limited linkage data are available from the literature, despite the fact that COD1 is located in a well-mapped part of the human genome. Linkage analysis in three families assigned an XLPCD locus, COD1, to the Xp21.1-p11.3 region, between the loci DXS84 and ARAFL (Bartley et al., 1989; Hong et al., 1991, 1994; Bergen et al., 1993). The COD1 locus has recently been refined to a limited region of Xp11.4 (Seymour et al., 1998). In this study haplotype analysis of two family branches, containing three obligate recombinants, two affected and one unaffected, defined the COD1 locus as distal to DXS993 and proximal to DXS556, a distance of ~1.0Mb, which excluded COD1 as an allelic variant of RP3 or RP2 (section 1.8.7) and established a novel locus sufficiently defined for positional cloning. On the other hand, BCM and possibly another form of XLPCD were recognised as defects in the green and red cone pigment-gene cluster located on Xq28 (Nathans et al., 1989; Reichel et al., 1989). In a recent report, linkage analysis in an XLPCD family, previously described by Pinckers and Timmerman (1981), identified a novel locus for XLPCD on Xq27 (Bergen and Pinckers, 1997), establishing evidence for genetic heterogeneity in XLPCD.

1.8.5 X-linked optic atrophy

X-linked optic atrophy (XLOPT) is another form of optic atrophy. Several inherited forms of atrophy of the optic nerve are known. Most common are dominant optic atrophy (DOA) and Leber hereditary optic neuropathy (LHON). Aside from DOA and LHON, other hereditary forms of optic atrophy are rare. In contrast to DOA and LHON, in which the optic nerve is usually the only affected organ, the recessive optic atrophies are multisystemic diseases; the CNS is affected along with the optic atrophy and other organs. In a large Dutch pedigree with XLOPT described by Volker-Dieben et al., (1974) and Went et al., (1975), the affected males have low visual acuities, severe bilateral optic atrophy with early onset and a slow progression of the disease. The macula and the periphery of the retina showed no abnormalities and no indication of night blindness. Ophthalmic study of the female carriers did not reveal any abnormalities. Recombination and linkage analysis on this Dutch family using 26 markers that spanned the entire X chromosome, placed the optic atrophy-causing gene in the Xp11.4-p11.21 interval between markers DXS993 and DXS991 also showing close linkage without recombination at the MAOB locus (maximum LOD score [Zmax] 4.19) (Assink, et al., 1997). Since the region of the X chromosome identified for the XLOPT gene also contained the genes involved in XLRP2 (section 1.8.7.3.2), CSNB2 (section 1.8.2) and Norrie disease (section 1.8.1), mutations in these genes could be involved in XLOPT. The localisation of an X-linked disease-causing gene in XLOPT could also have important implications for the pathogenesis of LHON and other optic atrophies as well.
1.8.6 Choroideremia

Choroideremia is an X-linked eye disorder characterised by progressive dystrophy of the choroid, retinal pigment epithelium and retina. Affected males develop night blindness in early adulthood, followed by a progressive constriction of visual fields. They show loss of the retinal pigment epithelium and the underlying choriocapillaris, which starts in the equator region of the fundus, with the retinal degeneration progressing towards the anterior retina and posterior pole. As the process is insidious, patients are frequently seen in a moderately advanced state with the posterior pole having the only remaining normal tissue, allowing the patient good central but little side vision. Eventually, complete blindness occurs by the fourth to fifth decade in life. The majority of choroideremia carriers show some fundus signs, including subretinal pigment clumping and an appearance of granularity to the retinal pigment epithelium as described in Figure 1.11 (Fundus photograph of choroideremia carrier). Usually the ERG, dark adaptation study and visual field examination are normal, but carriers with extensive fundus changes will demonstrate abnormalities on testing and may show progression with time. Usually the rate of progression does not functionally affect carriers, who rarely lose their vision. Retinitis pigmentosa carriers with retinal pigment epithelium and choriocapillaris atrophy often have bone spicule formation in equator regions, and may demonstrate choroidal sclerosis and relative preservation of anterior retinal pigment epithelium compared with the loss of these structures in choroideremia patients.

The disease locus was mapped to band Xq21 by both linkage analysis of pedigrees and cytogenetic analysis of patients with detectable chromosomal translocations and deletions (Cremers et al., 1990; Merry et al., 1992). A positional cloning strategy (Merry et al., 1992) was successfully used to isolate the human choroideremia (CHM) gene. The gene was found to be expressed in various tissue types, including retina, choroid and retinal pigment epithelium (Cremers et al., 1990). It was also demonstrated that the CHM gene escapes X chromosome inactivation (Carrel and Willard, 1993). This is remarkable, since the patchy changes of the retinal pigment epithelium in carrier females had been generally taken for evidence that the CHM gene is subject to X-inactivation.

The open reading frame (ORF) of the CHM gene was found to consist of 15 exons (van Bokhoven et al., 1994), which are partly or completely lacking in male choroideremia patients with deletions and in two female patients carrying a balanced X-autosome translocation involving the Xq21.2 band (Cremers et al., 1994). It spans at least 150 kb of Xq21.2 and encodes an ubiquitously expressed protein of 653 amino acids. Insight into the function of the CHM protein came with the biochemical purification of
Figure 1.11  Fundus photograph of a retina in a choroideremia carrier: the subretinal pigment clumping and an appearance of granularity to the retinal pigment epithelium (RPE) is due to the progressive atrophy of the choriocapillaris and RPE.
Rab geranylgeranyl transferase (Rab GGTase) from rat brain (Seabra et al., 1992). This enzyme attaches geranylgeranyl groups to Rab proteins, a modification essential for their action in intracellular vesicular transport. Rab GGTase is a heterodimer composed of tightly associated α and β-subunits. It requires an accessory component, the Rab escort protein (REP), for activity. Rat REP-1 was found to be identical with the protein encoded by the human CHM gene (Andres et al., 1993). REP-1 binds to newly synthesised Rab proteins, presents them to Rab GGTase and delivers the geranyl geranylated Rabs to their target membranes (Alexandrov et al., 1994).

Lymphoblasts of CHM patients have a markedly decreased but still detectable Rab GGTase activity (Seabra et al., 1993) suggesting the existence of an additional REP protein. A homologue of the CHM/REP-1 gene has been identified which was designated CHML for choroideremia-like. This intronless gene on chromosome 1q is expressed in a wide variety of tissues (van Bokhoven et al., 1994). The CHML or REP-2 protein was shown to perform a function similar to that of REP-1. It was hence hypothesised that this protein partially compensates for the loss of REP-1 activity in choroideremia patients, thereby preventing symptoms in tissues and organs other than the eye (Cremers et al., 1994a). Ocular symptoms might be due to the presence of Rabs that are preferentially prenylated by REP-1. In order to understand the pathogenesis of the disease in more detail, a mouse model was generated for choroideremia by targeted inactivation of the rep-1 gene (van den Hurk et al., 1997). The disruption of the rep-1 gene gave rise to lethality in male embryos but in females embryos was only lethal if the mutation was of maternal origin, therefore not being transmitted through the female germline. In both heterozygous females and chimeras the rep-1 mutation caused photoreceptor cell degeneration.

1.8.7 X-linked retinitis pigmentosa

Retinitis pigmentosa (RP) is a group of progressive retinal degenerations primarily affecting the (rod) photoreceptor cells and pigment epithelial layer (Pagon, 1988). With a prevalence of about 1 in 4000 in the general population, it is one of the more common causes of blindness in early and middle life. The prominent symptoms are progressive development of night blindness and constriction of the peripheral visual fields ('tunnel vision'), with eventual loss of central vision in most cases (Heckenlively, 1988). Fundus examination in advanced disease shows attenuated retinal vessels and bone-spicule pigmentary deposits commencing in the periphery due to invasion of the retina by cells of the retinal pigment epithelium (Figure 1.12; Fundus photograph of a retina in the late stages of XLRP). Progressive deterioration of retinal function is revealed by scotopic ERG, which shows reduced amplitude and delayed response times and soon becomes unrecordable (Heckenlively, 1988).
Figure 1.12 Fundus photograph of a retina in the late stages of XLRP. Advanced disease shows attenuated retinal vesicles and bone-spicule pigmentary deposits commencing in the periphery due to invasion of the retina by cells of the retinal pigment epithelium (RPE).
RP shows exceptional genetic heterogeneity as a consequence of the complex protein networks within the retina: families illustrating autosomal dominant (ADRP), autosomal recessive (ARRP), X-linked (XLRP), digenic inheritance and syndromal forms are well documented, with ~20 different RP loci mapped to date (Dryja and Li, 1995; Berson, 1996; Sullivan and Daiger, 1996). Up to 50% of RP cases are isolated (simplex), with no family history of the disease. Most of these are ARRP, but some are likely to be ADRP or XLRP (preponderance of males), while others may represent new mutations or phenocopies (Jay, 1982). The prevalence of the various genetic forms varies considerably from country to country (Jay, 1982).

The X-linked form of the disease (XLRP) presents the most severe phenotype. XLRP males become symptomatic in the first decade of life with blindness occurring by the third or fourth decade, and female heterozygous carriers often develop phenotypic signs of the disease in middle to late life (Bird, 1975). In such females random inactivation of one or other of the two X chromosomes in each cell during embryogenesis (Lyon, 1988) causes a wide variation in phenotype depending on the proportion of retinal cells with an active disease-bearing chromosome (Bird 1975).

1.8.7.1 Genetic heterogeneity in XLRP

Two lines of evidence firmly established the genetic heterogeneity of XLRP; genetic linkage studies and the analysis of X chromosome deletions. There are to date potentially 5 XLRP loci; RP2, RP3, RP6, RP15 and recently RP24. The evidence for each of these is discussed below.

Close genetic linkage between XLRP and DXS7 in Xp11.3 was first demonstrated in 5 British families by Bhattacharya et al., (1984). However, ensuing reports were in conflict over the location of XLRP relative to DXS7. Friedrich et al., (1985) identified two recombinants in a large Danish family that positioned XLRP proximal to DXS7, while others favoured a more distal location between DXS7 and DXS84, tightly linked to OTC in band Xp21 (Denton et al., 1988; Musarella et al., 1988). This was consistent with a male deletion patient (BB; Francke et al., 1985) who suffered from DMD, McLeod syndrome, CGD and RP. Two further deletion patients SB, with CGD, McLeod syndrome and XLRP, and OM, with only McLeod syndrome and CGD, further localised the XLRP gene proximal to the marker CYBB. The result of a large multipoint analysis of 20 XLRP families essentially excluded the XLRP locus from the Xp21 region and favoured a location proximal to DXS7 and distal to DXS14 in Xp11 (Wright et al., 1987).

The situation was clarified by heterogeneity analysis of linkage data pooled from 62 XLRP pedigrees from many countries (incorporating most of the previously published families), whereby Ott and colleagues (1990) provided overwhelming evidence in favour of 2 loci versus a single XLRP locus with odds of $6.4 \times 10^9$. The majority of families (75%)
were associated with a locus 1cM distal to OTC (RP3) in Xp21.1, consistent with the earlier localisations for XLRP from both deletion patients and linkage analysis of large pedigrees. In 25% of families the most likely location for the gene was 2cM proximal to DXS14 in Xp11.22. The confidence limits for the latter locus were wide, extending 3cM proximal to DXS7 to 1cM distal to the centromeric marker DXZ1, so that it could correspond to the Xp11 locus proposed by earlier linkage studies (RP2). More recent heterogeneity analysis of 37 XLRP families has placed the RP3 gene 0.4cM distal to OTC (70% of families) and the gene for RP2 6.5cM proximal to DXS7 (30% of families) (Teague et al., 1994), and these have become the 'working' locations for the two loci.

Musarella et al., (1988) proposed another XLRP locus (RP6) located between DXS28 and the DMD locus in Xp21.3-p21.2 based on recombination data from one two-generation family. Ott et al., (1990) found some statistical evidence for a third locus located in the same region, but patients deleted for this region have not been reported to have RP (Wright, 1990), which argues against an XLRP gene in this location.

A fourth locus for a clinical variant of XLRP was recently reported in distal Xp (McGuire et al., 1995), following linkage to DXS989 (Xp22.11) in a family originally thought to be segregating an autosomal dominant cone-rod dystrophy owing to the early cone involvement and the presence of severely affected females. This locus may correspond to the putative RP6 locus which is not genetically well-defined (Musarella et al., 1988). The authors argued that the progressive cone-rod degeneration seen in this family, although less common than the typical rod-cone subtype, is a justifiable clinical variant of RP (it has also been reported in an RP2 family; Jacobsen et al., 1992), and proposed the locus symbol RP15. However, the 'dominant' nature of the disease has been questioned on the premise that manifesting females are likely to arise owing to the lottery of X-inactivation. In general, unless the gene in question escapes X-inactivation, the terms 'dominant' and 'recessive' would seem inapplicable. This paper demonstrates the inter-familial phenotypic variability evident in X-linked retinal degenerations strengthening the case that many may be allelic variants, and has raised important questions concerning the wider issues of nomenclature for inherited eye diseases (Inglehearn and Hardcastle, 1996).

In support of further heterogeneity within XLRP, Aldred et al., (1994) described 3 XLRP families which appear unlinked to either the RP2, RP3 or RP6 loci through haplotype data and multipoint linkage analysis using markers spanning Xp22.2-Xq21.3. It was tentatively suggested that there may be further XLRP loci on distal Xp or Xq, although the authors recommend caution in interpretation as both families had affected males in only one generation.
Recently the mapping of a novel locus, RP24 by haplotype and linkage analysis of a single XLRP pedigree was reported (Gieser et al., 1998). The RP24 locus was identified at Xq26-27 by genotyping 52 microsatellites spanning the entire X-chromosome. A maximum LOD score of 4.21 was obtained with a marker DXS8106. Haplotype analysis assigned RP24 within a 23-cM region between the DXS8094 (proximal) and DXS8043 (distal) marker. Other chromosomal regions and known XLRP loci were excluded by obligate recombination events between markers in those regions and the disease locus. The mapping of RP24 should lead to a reevaluation of genetic data, with additional markers and for a larger cohort of XLRP families. A gene for an X-linked cone dystrophy (COD2) has also recently been assigned to Xq27, in a 8-cM genetic interval between DXS292 and DXS1113 (Bergen and Pinkers, 1997). This interval has a partial overlap with the RP24 critical region and although clinical results differ between the RP24 and COD2 families, it is conceivable that RP24 and COD2 are allelic diseases.

In summary, although the genetic heterogeneity of XLRP is well established, the story is far from complete. However, the overall genetic data is supportive of two predominant loci, RP3 in Xp21.1 and RP2 in Xp11.3-p11.2. As their close proximity often precludes straightforward assignment of families to one or other locus by genetic mapping, researchers have sought clinical difference between RP2 and RP3, primarily driven by the need to provide genetic counselling and prenatal diagnosis for 'at risk' carrier females, but also to aid gene mapping efforts.

1.8.7.2 Phenotype/Genotype correlations in XLRP

Inter-familial variation in phenotype can be due to a variety of causes, including polygenic or environmental influences on a single disease locus, due to different alleles at the same genetic locus in different families, or due to different genetic loci. There are no unequivocal phenotypic differences between RP2 and RP3 that allow them to be distinguished on clinical grounds alone. However, clinical studies in families that can be distinguished genetically have led to some tentative correlations between genotype and phenotype. Features such as tapetal reflex (Denton et al., 1988; Bergen et al., 1995a), myopia (Wright et al., 1991), onset of symptoms (Kaplan et al., 1992) and level of cone dysfunction (Wright et al., 1991; Jacobsen et al., 1992) have all been variably associated with either RP2 or RP3, but there is no conclusive evidence to suggest that any are indicative of a specific XLRP loci. Once the precise number and location of the genes causing XLRP are established it may be possible to determine whether or not clinical manifestations of each disease provides some basis for predicting the genotype. The confounding factor of allelic differences within each genetic type make it more likely that the issue will only be clarified once all the XLRP genes are isolated.
1.8.7.3 Recent progress in positional cloning of the RP3 and RP2 genes

A comprehensive review of all published linkage data (Aldred et al., 1994a), incorporating data of Bergen et al., (1995a) suggests that 36% of European families are RP2, almost two and three times that found in Australia (22%) and the USA (14%). Overall, 30% of families are RP2, consistent with the heterogeneity analyses of Ott et al., (1990) and Teague et al., (1994). Within Europe, RP2 appears particularly prevalent in the UK. Although this may reflect the large numbers of British families reported, this may be a consequence of a founder effect. The genetic origin and drift of distinct RP mutations within a country or region may considerably influence the probability of encountering a particular XLRP subtype on DNA diagnosis, with important implications for genetic counselling. The following sections summarise recent mapping progress on the two XLRP loci for which genetic evidence is undeniable, RP3 and RP2.

1.8.7.3.1 RP3 - Physical mapping studies and isolation of candidate genes

Fine mapping of deletion patients has provided the ultimate route towards identification of the RP3 disease-causative gene. The RP3 gene was believed to lie in the proximal portion of the BB deletion (section 1.8.7.1) because of its coincidence with the critical region delineated by linkage analysis. The two genetic markers, DXS1110 and OTC that flank the RP3 locus span a physical distance of ~520kb (Nelson et al., 1995); the proximal end of the BB deletion is 40kb centromeric to DXS1110. Several groups have been searching for transcribed sequences in this region surrounding the proximal breakpoint for several years.

The BB deletion junction was initially done by Musarella et al., (1991) using a DMD cDNA as a starting point from which to isolate genomic clones and create a long-range physical map of the proximal part of the deletion. PFGE studies restricted the RP3 locus to a 205kb SfiI fragment within which a CpG island was identified. Segments of genomic DNA adjacent to the CpG island hybridised to discrete bands in digested DNA from several species, indicating evolutionary conservation, and these segments were subsequently used to isolate a cDNA of ubiquitous expression from retinal cDNA libraries (Roux et al., 1994). Given its high similarity to the murine tctex-1 gene (thought to be involved in spermatogenesis) the function of the novel human gene (TCTE1L) was speculative, but its location in Xp21 and complete deletion in patient BB prompted investigation as a candidate for RP3. However, no disease-associated changes in the coding portion of the gene were found in 20 RP3 patients (Roux et al., 1994).

At this point re-evaluation of the contiguous deletions associated with syndromes in Xp21.1 suggested that the deletion observed in patient BB may have misdirected the search for RP3. A deletion in patient NF who suffered from DMD and
CGD but did not have RP symptoms, extended proximal to that of BB, essentially excluding the proximal part of the BB deletion from containing RP3 (Brown et al., 1996). Gene cloning efforts were thus directed to an extended interval between the TCTE1L locus and the SB proximal deletion breakpoint 50kb distal to OTC (Meindl et al., 1995a).

In the autumn of 1995, two groups simultaneously reported the cloning of a candidate RP3 gene, SRPX/ETX-1, mapping within this interval (Meindl et al., 1995a; Dry et al., 1995). To identify transcribed sequences, the strategy adopted by Meindl et al., (1995a) was to screen retinal cDNA libraries with sub-fragments of a YAC containing both OTC and CYBB, paralleled by genomic sequencing of putative CpG islands and hybridisation of cosmid digests with a splice-site consensus sequence. Dry et al., (1995) used YACs covering the CYBB-OTC interval to isolate cosmids upon which to perform exon amplification (section 1.5.5). The SRPX/ETX-1 transcript showed highest expression in retina and heart, encodes a putative cell surface protein and contains sushi repeats similar to selectin cell adhesion proteins. SRPX/ETX-1 was found to lie 250kb proximal to the BB deletion junction (Meindl et al., 1995a) but was considered a candidate gene for RP3 on the grounds that large chromosomal re-arrangements may affect the expression of nearby genes (position effect; Bedell et al., 1996) or be associated with secondary deletions remote from the major site. The gene was completely deleted in XLRP patient SB, and a microdeletion was deleted in XLRP patient MO encompassing the promoter region and exon 1 of SRPX (Meindl et al., 1995a). However, no further functionally significant mutations were detected by SSCP screening of all exons in a panel of unrelated XLRP patients nor by full-length RT-PCR sequencing in two RP3 families (Meindl et al., 1995a). The role of this highly conserved retinally expressed gene in the pathogenesis of RP therefore remains to be determined.

A timely collaborative effort between the two groups in searching for further transcripts mapping within the MO deletion has since resulted in the cloning of a gene which is mutated in a proportion of RP3 patients (Meindl et al., 1996). The gene is evolutionarily conserved and is thought to be a 'housekeeping gene' on the basis of its ubiquitous expression and 5' CpG island (Dry et al., 1996). Interestingly, this gene shows very low levels of expression in the retina and RPE, which explains why it could not be isolated through direct screening of retinal cDNA libraries (Meindl et al., 1996). The gene was eventually identified by systematically subcloning and sequencing two cosmids covering the proximal part of the MO deletion, and detection of gene sequences by computational analysis (section 1.5.4.8). Mutations were found in highly conserved residues which segregated with disease in 7 XLRP families from a pool of 74 unrelated patients, providing evidence that this gene underlies at least a percentage of RP3. The predicted protein contains a tandem repeat structure similar to RCC-1 (regulator of chromosome condensation) which regulates the GTPase Ran (Ras-related nuclear protein), known to play a role in cell cycle progression, membrane transport and RNA
1. Introduction

processing. The high rates of membrane turnover in the retina and RPE (see section 1.6.1.2) have led to speculation that this novel gene, termed RPGR (retinitis pigmentosa GTPase regulator) acts to regulate this process (Meindl et al., 1996). It is interesting that in another X-linked eye disorder, choroideremia, the defective gene plays a role in the geranylgeranylation of different Rab proteins, another family of Ras-related GTPases (Seabra et al., 1993; van Bokhoven et al., 1994). In view of the clinical similarities between choroideremia and RP3 (Bird 1975), it is speculated that RPGR may be a guanine-nucleotide exchange factor for retina-specific Rab proteins.

Shortly after the publication of the RPGR gene, Roepman et al., (1996) cloned the same gene via a novel method called 'YAC representation hybridisation' and found disease-associated mutations in 5 out of 28 XLRP patients. It is surprising that so few of the XLRP patients screened, about 20% (Buraczynska et al., 1997) revealed disease-associated mutations when RP3 accounts for ~70% of all XLRP cases (section 1.8.7.1). Perhaps mutations may lie in unidentified parts of the gene or alternative transcripts, common mutations were not identified by SSCP analysis and/or there is heterogeneity within RP3, with the major locus still to be identified. Fujita et al., (1996) have recently reported a recombination event in a large RP3 family which localises the causative mutation proximal to the BB deletion. This was further supported by Brown et al., (1996) whose results suggested that RP3 lies outside the BB and NF deletions and within a 380-kb region between the proximal NF and SB deletion breakpoint junctions, from which region the RPGR gene was isolated (Meindl et al., 1996). If RPGR is the major RP3 gene, questions remain as to the presence of XLRP in patient BB. It is feasible that his large deletion may include another RP3 locus, or an RPGR regulatory element situated some distance from the gene. Other likely explanations include a long-range chromosomal position effect, a small secondary rearrangement and without a family history, it is also difficult to exclude the possibility that patient BB may have had an autosomal form of RP.

Studies are currently underway to determine the precise location and function of RPGR in the retina and elucidate the mechanism by which defects within it lead to retinal degeneration. This will enhance our understanding of normal retina function and may provide clues as to the cause of other inherited retinopathies, in particular RP2. Investigation of the physiological function of the RPGR protein and generation of a mouse model of X-linked RP has recently been attempted by cloning and characterising the full-length and variant cDNA isoforms derived from the mouse homolog of the human RPGR gene; designated mRpgr (Yan et al., 1998). The discovery of RPGR will undoubtedly benefit women in families segregating the gene who request carrier testing, and will diagnose XLRP in a proportion of sporadic patients enabling more appropriate counselling and prognosis. Recent mutation screening of 10 CSNBX pedigrees has disclosed an RPGR mutation segregating in one family which is absent from 170 control
X chromosomes (Hermann et al., 1996). This provides evidence for the postulated allelism between XLRP and CSNB (section 1.8.2) and may assist functional analysis of the gene product.

1.8.7.3.2 RP2 - further genetic mapping studies to confirm and refine the RP2 locus and recent isolation of a candidate gene

The lack of associated cytogenetic abnormalities for RP2 has hampered localisation of this gene and its precise location by linkage analysis has been less well-defined. Ott et al., (1990) localised RP2 to a broad region extending from DXS7 in Xp11.3 to the centromere. Multipoint linkage analysis by Wright et al., (1991) on the large British kindred of Bhattacharya et al., (1984) gave a maximum likelihood location for RP2 close to DXS255 in Xp11.22 and TIMP-1 in Xp11.23 in an area extending from 2cM proximal to DXS7 to 1cM distal to DXS14. This is supported by the study of Teague et al., (1994), where RP2 had a maximum likelihood location 6cM proximal to DXS7. The recent heterogeneity analysis of Bergen et al., (1995a) produced a most likely location for RP2 at DXS255, with a confidence interval extending from DXS7 to DXS14.

In summary, all multipoint linkage studies have given differing most likely locations for RP2, albeit with varying confidence intervals which overlap to some degree. This may reflect detection of linkage by DXS7 to both RP2 and RP3 loci (only ~10-15cM separates them). More importantly, the relative rarity of the RP2-type family (30% of most XLRP populations) reduces the number of informative recombination events necessary for fine genetic localisation of the gene.

At the outset of this study, overall data indicated a location for the RP2 locus between DXS7 and DXS255 (Wright et al., 1991; Friedrich et al., 1992), a genetic interval of 13-18cM (Mahtani et al., 1991) which was refined to 5cM between markers DXS8083 and DXS6616 in this laboratory (Thiselton et al., 1996). Thus there is still scope for refining the localisation of RP2 further by genetic mapping, but given the problems of classifying families and obtaining large pedigrees with informative recombinants, other approaches are being pursued. Since a physical map was being constructed in the laboratory using YACs centred around the DXS426 locus, then cDNA selection and isolation strategies using these YAC reagents to generate new ESTs was conducted (Chapters 4 & 5). Also investigation of candidate genes mapping to the critical interval was carried out (Chapter 6). Families segregating CSNBX mapping to Xp11 may also be useful for refining the localisation of RP2 if these represent allelic disorders.

Aldred et al., (1994b) have described a family in which XLRP cosegregates with mental retardation and which appears to be RP2 by linkage analysis (maximum likelihood location 0.5cM distal to TIMP). This may represent a new genetic syndrome due to a locus that is fortuitously located in the same region, or, more interestingly,
raises the possibility of a contiguous gene deletion syndrome involving the RP2 gene in this family, which could significantly reduce the region of search for the RP2 gene.

Recently a novel gene RP2, which is mutated in approximately 18% of the patients with X-linked retinitis pigmentosa 2 was isolated (Schwahn et al., 1998). The gene was identified by positional cloning from a 5-cM linkage interval in Xp11.3, through the detection of an LINE1 retrotransposition in intron 1 of the RP2 gene. This was found in 1 of 26 patients screened with the YAC representation hybridisation (YRH) technique, which was used previously for the identification of a 6.4-kb microdeletion that was instrumental in the isolation of the RP3 gene (Roepman et al., 1996a). In addition to the L1 retrotransposition, two nonsense, two frameshift and one missense mutation as well as one in-frame deletion have also been identified (Schwahn et al., 1998). Through expression studies and RT-PCR the RP2 gene is found to be expressed ubiquitously and subject to X chromosome inactivation. A domain of 151 aa within the N-terminal portion of the RP2 polypeptide revealed significant homology with cofactor C, a protein involved in the biogenesis of the β-tubulin molecule, which could suggest a possible role of the RP2 gene in tubulin folding. Thus, RP2 may be due to a novel mechanism not previously implicated in the pathogenesis of RP.

1.9 Aims of this thesis

The primary aim of this thesis was to further progress the positional cloning and identification of the gene causing one form of XLRP (RP2) on proximal Xp. This was carried out using genetic mapping strategies to further define the RP2 gene critical region, and using YAC reagents within this interval to establish experimental procedures for the efficient enrichment of expressed sequences from this target region for cDNA transcript isolation and analysis.

Initial research was aimed at genetic characterisation of 13 new XLRP families using an extensive set of markers spanning Xp22.13-p11.22 in order to distinguish between those families segregating either the RP2 or RP3 (or other XLRP) loci, and further delineate the RP2 critical region. The main focus of subsequent research was the successful establishment and application of the magnetic bead capture technique for cDNA selection in the 500kb region distal to DXS426, within the RP2 critical interval, using two well characterised retinal cDNA libraries. The enrichment, isolation and characterisation of potential novel retinal cDNA clones from selected sub-libraries would help in the assembly of a transcriptional map of this target region and in the identification of the gene responsible for RP2.
SUMMARY OF INTRODUCTION

1.1 Principles and strategies for identifying disease genes

Several of the possible strategies for identifying disease genes aim initially to identify a number of candidate genes, which then have to be tested individually for evidence that implicates them as the disease-causative gene. Mapping the disease to a specific subchromosomal localisation is generally the first step, which can be done by genetic mapping followed by physical mapping and the subsequent identification of gene sequences from a defined genomic interval. Recently accumulating DNA sequence data in the form of sequence tagged sites (STSs) and expressed sequence tags (ESTs) generated by the human genome project have proven enormously valuable in identifying and mapping genes.

1.1.1 The Human Genome Project

The major scientific thrust of the Human Genome Project concerns constructing high resolution genetic and physical maps as a prelude to the ultimate physical map, the complete sequence of the human genome. There is also extensive interaction with research focusing on mapping disease genes, a commitment to map and sequence the genomes of a variety of model organisms and to develop ancillary technologies including data analysis.

1.1.1.1 Human genetic maps

Several major efforts to construct genetic maps of the human genome have occurred during the past few years, and recent emphasis has been on merging these efforts to forge consensus maps. The ideal framework map will have markers spaced uniformly at distances around 3 to 5 cM, since this is the most efficient sort of map to use to try to find additional markers or disease genes with current technology. The latest genetic map with more than 5,000 highly polymorphic simple sequence repeats (Dib, C et al., 1996) represents the last version of the Genethon map with an average marker density every 0.7 cM or about every 700,000 base pairs.
1.1.1.2 Physical maps of the human genome

Using STS landmarks to identify and order clones, a framework physical map has been constructed (Cohen D et al., 1993) for most of the human genome from large inserts of human DNA cloned into yeast artificial chromosomes (YACs). More recently using the techniques of genetic mapping (1- to 10-Mb resolution), fluorescence in situ hybridisation (~1-Mb resolution), and radiation hybrid mapping (down to 50-kb resolution), combined with the results of STS-content screening against a YAC library, markers have been placed on average every 200 kb across the genome producing an integrated physical map of the human genome (Hudson, TJ et al., 1995).

1.1.1.3 Transcriptional maps of the human genome

Transcript maps of the human genome are an important part of the sequencing infrastructure as well as a critical resource for the positional candidate approach to disease gene cloning. The mapping of 318 cDNAs using the CEPH mega-YAC library (Berry, R et al., 1995) and 13,600 ESTs on the human physical map (Hilllier et al., 1996) has enabled regional gene mapping on a genome-wide scale. The screening and usage of cDNA libraries has been facilitated by using gridded master arrays of clones representing unique genes, the development of which has been fostered by the IMAGE consortium.

1.1.1.4 Sequencing the human genome

Recent technological developments and experience with large-scale sequencing provide increasing confidence that it will be possible to complete an accurate, high-quality sequence of the human genome by the end of 2003. To date only about 6% of the human genome sequence has been completed (Collins F et al., 1998). Only sequence of high accuracy and long-range contiguity will allow a full interpretation of all the information encoded in the human genome. A "working draft" human genome sequence produced sooner, will be of lower accuracy and contiguity. It will nevertheless be very useful, especially for finding genes, exons, and other features through sequence database searches and a growing number of prediction programs. However, because this sequence will have gaps, it will not be as useful as finished sequence for studying DNA features that span large regions or require high sequence accuracy over long stretches such as gene regulatory regions which may lie some distances from the ORF.
1.1.1.5 The genetic, physical and transcriptional map of the human X-chromosome

A unified genetic, physical and transcriptional map of the human X chromosome is being assembled through a concerted, international effort. The X chromosome integrated database (IXDB) a repository for physical mapping data of the human X chromosome helps support the construction of an integrated physical, genetic, transcript and sequence map of the X chromosome (Leser et al., 1998).

YAC contigs with an average clone depth of at least four-fold coverage now span more than 80% (125 Mb) of the 150 Mb X-chromosome (Crollius et al., 1996), with an average inter-STS resolution of 75 kb and a current total of around 2091 STSs in cognate YACs (Nagaraja et al., 1997). Several yeast artificial chromosome (YAC) contigs from the Xp11.23-22 region have been reported, however none of these maps provide a full coverage. A YAC contig encompassing the distal part of the Xp11.23 region between the loci TIMP1 and OATL1 has been reported (Coleman et al., 1994; Knight et al., 1994). There are 6 major genetic maps for the X chromosome (Murray et al., 1994, Matise et al., 1994, Donnelly et al., 1994, Wang et al., 1994, Fain et al., 1995 and Dib et al., 1996) consisting of RFLP's and microsatellite markers. These genetic maps show a high degree of overall consistency and the genetic lengths are in good agreement with that postulated by Morton in 1991 (220cM).

Recently 59 ESTs were mapped in 31 intervals spanning the X chromosome (Gianfrancesco et al., 1997) where a somatic cell hybrid panel provided a resolution of about 5Mb for the short arm and 4Mb for the long arm, and further resolution was obtained using YACs mapped in localised contigs. The construction of a 1.5Mb BAC contig was able to provide total physical coverage of the Xp11.23 region with 28 clones that are suitable for large-scale sequencing efforts (Boycott et al., 1998a).

1.1.2 Genetic mapping

1.1.2.1 Principles of linkage analysis

The aim of linkage analysis is to determine the recombination frequency between two loci, which may be either two genetic markers or a genetic marker and a disease locus. The further two loci are on a chromosome, the more likely it is that a crossover will separate them. Thus the recombination fraction \( \theta \) is also a measure of the distance between two loci and define genetic distance. Two loci which show 1% recombination are defined as being 1cM apart on a genetic map. The maximum recombination fraction between any two loci is 0.5 or 50% which corresponds to unlinked loci which segregate independently. Where two markers lie close together
on a chromosome, the number of crossovers will be minimal, and the $\theta$ value will be less than 0.5, and approaches 0 for tightly linked loci. A genetic haplotype refers to the combination of alleles observed at a linked locus. Haplotypes mark recognizable chromosomal segments which can be tracked through pedigrees and through populations. Such data are used to determine the minimal region of interest.

1.1.2.1 Calculation of lod scores

The lod score, $Z$, is a logarithm of the odds that two loci are linked (with recombination fraction $\theta$) rather than unlinked (recombination fraction 0.5). Being a function of the recombination fraction, lod scores are calculated for a range of $\theta$ values and the maximum value $Z$ estimated. The lod score at a given recombination fraction is therefore calculated as $Z = \log_{10} \left[ \frac{L(\theta)}{L(0.5)} \right]$ where $L$ is the likelihood of an observation of recombination at a given $\theta$. The $\theta$ which gives the highest lod score obtained ($Z_{\text{max}}$), provides the best estimate of linkage between two loci.

1.1.3 Physical mapping

Physical mapping of complex genomes can be accomplished by one of two approaches: site mapping or clone mapping. Site mapping uses small DNA hybridisation probes to analyse genomic DNA digested with rare cutting restriction enzymes that contain one or more CpG motifs in their recognition sequence. The DNA fragments are then separated by pulse field gel electrophoresis (PFGE) and detected by hybridisation of DNA probes on Southern blots. Site mapping enables large regions of genomic DNA to be mapped and characterised without reliance on cloned DNA. In contrast, clone mapping aims to form contiguous sets of overlapping clones or 'contigs' representing the genome, and the locations of sequence structures such as restriction enzyme sites are determined by the analysis of individually cloned fragments. Clone mapping results in a detailed map of all restriction enzyme sites in the genome and provides complete templates for the isolation of expressed sequences. While site maps can be constructed fairly rapidly and are valuable in providing an initial framework to direct gene searches, clone mapping has become a widely adopted approach since the introduction of yeast artificial chromosome (YAC) cloning (Burke et al., 1987).

Although YACs can accommodate large DNA fragments so that fewer clones are required to establish physical coverage, YAC cloning have several limitations which include low cloning efficiencies, a high percentage of chimeric and rearranged clones (40-60%), and the difficulty of purifying insert DNA away from endogeneous yeast DNA. This has led to the development of alternative bacteria-based cloning
systems such as bacteriophage P1, bacterial artificial chromosomes (BACs) and P1-derived artificial chromosomes (PACs), to complement the YAC system (reviewed in Monaco and Larin, 1994).

1.1.4 Identification of gene sequences from a defined genomic interval

1.1.4.1 Cross species sequence homology and selection of CpG islands

A number of strategies exist for the isolation of genes from cloned DNA. The earliest method involves screening anonymous genomic DNA fragments for sequences that are evolutionarily conserved. This strategy has been instrumental in the cloning of genes for the DMD gene (Monaco et al., 1986) and more recently (in combination with other methods) to isolate genes in the BRCA1 genomic region (Jones et al., 1994). Another conventional strategy that is frequently used involves the identification of CpG islands as signposts for the 5' ends of genes (Antequera and Bird, 1993; Cross and Bird, 1995). CpG islands differ from bulk genomic DNA by being relatively (G+C)-rich (>60%) and having a high concentration (10-20 times above average) of the CpG nucleotide (Larsen et al., 1992). CpG islands are often associated with expressed sequences and their selection provides a rapid and efficient way to identify potential coding regions. However, this approach is limited by the fact that not all genes are associated with CpG islands.

1.1.4.2 Direct YAC screening of cDNA libraries

As YAC-based mapping and cloning becomes increasingly important in positional cloning, several different strategies for identifying transcribed sequences within YAC clones have been devised. One of these involves labelling of gel-purified YAC inserts and using them as hybridisation probes to screen cDNA libraries (Elvin et al., 1990). This method has the advantage of speed, since it avoids subcloning of the YAC and multiple transcripts can be identified with a single screening. However, in practice there are a number of disadvantages to this approach. These include isolation of YAC DNA in sufficient quantity and purity, and the requirement of large amounts of radioactivity for labelling. The complexity of the probe usually result in poor signals, increased background and false positives. In addition, multiple cDNA libraries would have to be screened to take into account tissue specificity as well as developmentally regulated gene expression. Despite these difficulties, several genes have been isolated using this approach (Kahloun et al., 1992; Geraghty et al., 1993).
1.1.4.3 Exon amplification

Exon amplification or trapping is an expression independent method for isolating coding sequences from cloned DNA (Duyk et al., 1990; Buckler et al., 1991). It is based on in vivo selection of functional splice sites flanking sequences in the genomic DNA. Genomic DNA fragments are subcloned into a specially constructed vector that facilitates transcription and splicing of exons. Cytoplasmic mRNA derived from mammalian cells transfected with these constructs, are analysed by RNA-based PCR amplification for the presence of an exon from the genomic insert. This method has successfully isolated the genes for Menkes disease (Vulpe et al., 1993), NF2 (Trofatter et al., 1993) and the Huntingdon's disease (The Huntingdon's Disease Collaborative Research Group, 1993).

1.1.4.4 cDNA selection

The original 'cDNA selection' method involves hybridisation of an amplified cDNA library to the target DNA (YAC or cosmid DNA), immobilised on nylon membranes after suppression of repetitive sequences (Parimoo et al., 1991; Lovett et al., 1991). Nonspecifically hybridising sequences are removed by stringent washing conditions whereas specific cDNAs that remained are eluted and amplified by PCR before cloning. Improvements to the original cDNA selection using (i) pools of normalised cDNA libraries, (ii) hybridisation in solution rather than membrane-bound templates, and (iii) the use of biotin-streptavidin and magnetic beads to immobilise cDNA inserts that specifically hybridised to the target DNA (reviewed in Lovett, 1994). This procedure has led to the construction of transcriptional maps over large areas of the human genome (Sedlacek et al., 1993; Gezc et al., 1993) and the identification of the human disease gene for X-linked agammaglobulinaemia (Vetrie et al., 1993) and expressed sequences within the 1Mb region flanking BRCA1 on human chromosome 17q21 (Osborne-Lawrence et al., 1995).

1.1.4.5 Sequence based methods

Recent computational advances have enabled the development of sequence-based methods for identifying potential exons in genomic sequence data. Most notably, is the development of GRAIL, which recognises coding regions in DNA sequence through a technology generally referred to as 'pattern recognition'; (Uberbacher and Mural, 1991). Basically the system recognises characteristics of sequence found to occur in protein coding regions. The limitation of the current system is that it will not identify exons that contain noncoding sequence. In practice,
only small genomic regions will be analysed by this approach because of the requirement for sequence data. However, with the availability of automated sequencing machines, this obstacle will become less relevant and more gene searches could be conducted in this way. This approach was used in the positional cloning of the gene for Kallman syndrome, which involved sequencing 60kb of genomic DNA (Legouis et al., 1991).

1.2 Inherited retinal degeneration

Linkage studies of large families and candidate gene screening of retinal genes have identified 115 independent genetic loci that can cause retinal degeneration (RETNET, 1999; website RetNet/sum-dis.htm). The heterogeneity observed in inherited retinal degeneration includes genetic heterogeneity, allelic heterogeneity and clinical heterogeneity. The genetic heterogeneity observed in these diseases suggests that retinal degeneration is a common end point for many biochemical abnormalities, and may be a reflection of the limited response of the retina to disease.

1.2.1 Structure of the retina

The retina is the neural sensory layer of the eye which houses a dense array of light-sensitive photoreceptors containing visual pigment molecules that initiate the neural response to light which has been focused onto the retina by the cornea and lens. Structurally and functionally the retina consists of two distinct layers: the non-neural retinal pigment epithelium (RPE), and the tightly apposed neural retina. The primate retina is a "duplex" retina consisting of both rod and cone photoreceptor cells. Rods mediate vision in dim light, whereas cones function in bright light. Rods provide great sensitivity, especially to blue-green light (scotopic vision), whereas cones provide visual acuity for pattern detection as well as colour vision (photopic vision). All photoreceptors have an outer segment that contains the photosensitive pigment molecules. The outer segment is the site of phototransduction, where conversion of light energy into an electrical signal occurs. Both rods and cones contain an elaborate system of stacked membranous discs that arise during development as a series of invaginations of the cell's plasma membrane. Rod cells are sensitive to light because they contain a visual pigment called rhodopsin, which is capable of trapping photons. Rhodopsin is arranged in a single molecular layer in the discs of the outer segment and is composed of two essential parts. One is the light-absorbing chromophore, vitamin A aldehyde, which is called 11-cis-retinal. Chemically, this component is closely related to vitamin A. Retinal is attached to the second moiety of rhodopsin, known as opsin (Wald, 1968).
1.2.2 Phototransduction

The primary function of the retina is phototransduction. The first steps in the phototransduction pathway occur in the cells of the photoreceptor layer, the rods and cones. Here, opsin proteins, bound to the light-responsive chromophore 11-cis-retinal, are embedded in the plasma membrane and in stacked disc structures within the outer segment of each cell. When opsin absorbs a photon of light it becomes activated and, in turn, activates the G protein transducin. Upon activation, transducin activates cyclic GMP (cGMP) phosphodiesterase (PDE), which hydrolyses cGMP. As cGMP levels in the cytoplasm drop, the cGMP-gated ion channels (CNCGs) within the plasma membrane close, leading to hyperpolarisation of the membrane. Numerous other specialised proteins are also required to modulate and inhibit this cascade of reactions, ultimately returning the opsin molecule to its inactive state, ready to intercept another photon. This basic pathway takes place in both rods and cones and has been extensively studied (Stryer, 1991; Yau, 1994; Polans et al., 1996).

1.2.3 X-linked eye diseases

The proximal short arm of the human X chromosome is home to a cluster of genes implicated in various inherited disorders of the eye. Several of these disorders share key symptoms with differing degrees of severity and overlapping map locations, and the growing examples of allelic heterogeneity suggest that different mutations in the same gene may have diverse phenotypic effects.

1.2.3.1 X-linked retinitis pigmentosa

Retinitis pigmentosa (RP) is a group of progressive retinal degenerations primarily affecting the (rod) photoreceptor cells and pigment epithelial layer (Pagon, 1988). The prominent symptoms are progressive development of night blindness and constriction of the peripheral visual fields ('tunnel vision'), with eventual loss of central vision in most cases (Heckenlively, 1988). Two lines of evidence firmly established the genetic heterogeneity of XLRP; genetic linkage studies and the analysis of X chromosome deletions. Genetic analysis of XLRP families has shown a high degree of heterogeneity. There are two major loci (RP2 and RP3), which have been mapped to the short arm of the X chromosome. Other loci include RP6, RP15, RP23 and RP24 (Ott et al., 1990; McGuire et al., 1995; Gieser et al., 1998; Hardcastle et al., in preparation). Retinitis pigmentosa type 3 (RP3) has been shown to be the predominant form of XLRP from genetic linkage studies and segregates with the
disease in 60%-90% of affected families (Ott et al., 1990; Teague et al., 1994; Fujita et al., 1997) and overall, 30% of families are RP2 based on heterogeneity analyses of Ott et al., (1990) and Teague et al., (1994).

1.2.3.1.1 Retinitis pigmentosa 3

Fine mapping of deletion patients has provided the ultimate route towards identification of the RP3 disease-causative gene. The RP3 gene was believed to lie in the proximal portion of the BB deletion at Xp21 because of its coincidence with the critical region delineated by linkage analysis. The two genetic markers, DXS1110 and OTC that flank the RP3 locus span a physical distance of ~520kb (Nelson et al., 1995); the proximal end of the BB deletion is 40kb centromeric to DXS1110. Several groups have been searching for transcribed sequences in this region surrounding the proximal breakpoint.

In 1996, a gene named RPGR (retinitis pigmentosa GTPase regulator) was cloned from the RP3 region (Meindl et al., 1996; Roepman et al., 1996). The RPGR gene is ubiquitously expressed and consists of 19 exons. The putative RPGR gene product of 815 amino acids contains a tandem repeat structure similar to the RCC1 (regulator of chromosome condensation) protein which regulates the GTPase Ran (Ras-related nuclear protein), known to play a role in cell cycle progression, membrane transport and RNA processing. Recently, it has been shown that the RCC1-like domain of RPGR interacts with the δ subunit of rod cyclic GMP phosphodiesterase (PDEδ) and that disruption of this interaction may be the cause of, or may contribute to, retinal disease (Linari et al., 1999). Various groups have independently reported disease-causing mutations within the RPGR gene in XLRP patients. The percentage of mutations found is 15%-20% of XLRP (Zito et al., 1999; Buraczynska et al., 1997; Meindl et al., 1996; Roepman et al., 1996), which is a remarkably low frequency compared with the high percentage of familial disease linked to the RP3 locus. Investigation of splice variants, promoter/enhancer mutations, and as yet unidentified exons may account for the low mutation rate detected in this gene; however, an equally likely explanation for these findings could be microheterogeneity.

1.2.3.1.2 Retinitis pigmentosa 2

The lack of associated cytogenetic abnormalities for RP2 has hampered localisation of this gene and its precise location by linkage analysis has been less well-defined. At the outset of this study, overall linkage analysis data indicated a location for the RP2 locus between DXS7 and DXS255 (Wright et al., 1991; Friedrich et al., 1992), a genetic interval of 13-18cM (Mahtani et al., 1991) which was refined to
Summary of Introduction

5cM between markers DXS8083 and DXS6616 (Thiselton et al., 1996). Recently a novel gene RP2, which is mutated in approximately 18% of the patients with X-linked retinitis pigmentosa 2 was isolated by use of the YAC representation hybridisation technique (Schwahn et al., 1998). The gene was identified by positional cloning from a 5-cM linkage interval in Xp11.3, through the detection of an LINE1 retrotransposition in intron 1 of the RP2 gene. In addition to the L1 retrotransposition, two nonsense, two frameshift and one missense mutation as well as one in-frame deletion have also been identified (Schwahn et al., 1998).

The RP2 gene consists of five exons, encodes a polypeptide of 350 amino acids and is ubiquitously expressed. A domain of 151 aa within the N-terminal portion of the RP2 polypeptide revealed significant homology with cofactor C, a protein involved in the biogenesis of the β-tubulin molecule, which could suggest a possible role of the RP2 gene in tubulin folding. Thus, disease may be due to a novel mechanism not previously implicated in the pathogenesis of RP. Recently mutations in the RP2 gene was shown to cause disease in 10% of families with familial X-linked retinitis pigmentosa (Hardcastle et al., 1999). The spectrum of RP2 mutations has a trend towards severe protein truncation (Mears et al., 1999; Hardcastle et al., 1999), with only one missense mutation identified in two families (Schwahn et al., 1998; Hardcastle et al., 1999). Since genetic mapping studies suggest that a maximum of 25% of XLRP cases are linked to this interval (Teague et al., 1994) a large increase in the percentage of cases of XLRP attributable to RP2 gene mutations is not anticipated.

1.2.3.1.3 Choroideremia

Choroideremia is an X-linked eye disorder characterised by progressive dystrophy of the choroid, retinal pigment epithelium and retina. Affected males develop night blindness in early adulthood, followed by a progressive constriction of visual fields. They show loss of the retinal pigment epithelium and the underlying choriocapillaris. The disease locus was mapped to band Xq21 by both linkage analysis of pedigrees and cytogenetic analysis of patients with detectable chromosomal translocations and deletions (Cremers et al., 1990; Merry et al., 1992). A positional cloning strategy (Merry et al., 1992) was successfully used to isolate the human choroideremia (CHM) gene. The CHM gene was found to consist of 15 exons (van Bokhoven et al., 1994), and spans at least 150 kb of Xq21.2 and encodes a ubiquitously expressed protein of 653 amino acids which is identical to rat Rab escort protein-1 (REP-1) (Andres et al., 1993). Rab geranylgeranyl transferase (Rab GGTase) from rat brain (Seabra et al., 1992) attaches geranylgeranyl groups to Rab proteins, a modification essential for their action in intracellular vesicular transport. REP-1 binds
to newly synthesised Rab proteins, presents them to Rab GGTase and delivers the geranyl geranylated Rabs to their target membranes (Alexandrov et al., 1994).

1.2.3.1.4 Other eye diseases on proximal Xp

A range of other X-linked retinal phenotypes have been mapped to this region including Norrie's disease, congenital stationary night blindness (CSNB), Aland Island eye disease and cone dystrophy, all with loci overlapping the XLRP intervals and which might or might not be allelic with them.

Norrie disease (ND) is an X-linked recessive neurodegenerative condition with ocular atrophy and is characterised by congenital or early childhood blindness. ND was found to be localised between DXS7 (Lindsay et al., 1992) and the proximal end of a YAC clone (YL1.28) of 160kb (Chen et al., 1992). Positional cloning identified a candidate cDNA of 1.8kb (Berger et al., 1992; Chen et al., 1992a) with microdeletions of this gene in several Norrie disease patients (Chen et al., 1993). Computational modeling of the ND protein, termed norrin, or NDP indicates that it shares homology and a predicted three-dimensional structure with a carboxy-terminal domain (termed the cysteine knot) found as a component in a variety of cysteine-rich neurotrophins such as transforming growth factor (TGF) and nerve growth factor (Meitinger et al., 1993). On this basis it has been predicted that the dimerisation of NDP may be an essential element in providing the recognition signals for targeting of neuronal/retinal connections.

Congenital stationary night blindness (CSNB) comprises a group of disorders characterised by congenital onset of non-progressive night blindness, subnormal visual acuity, variable myopia and absence of pigmentary degeneration of the retina (Pearce, 1990). Miyake et al., (1986) have classified CSNB of mixed genetic types into two clinical subgroups, termed complete and incomplete CSNB, on the basis of refractive error, ERG and dark adaptation responses. The complete type (CSNB1) lacks rod function and myopia is usually severe. The incomplete type (CSNB2) shows some rod function but also impairment of cone function. Genetic analyses over the last decade have established heterogeneity for CSNBX, implicating at least two distinct genes on the proximal short arm of the X chromosome. Based on critical recombination events in a set of families, complete CSNB has been localised to the region between DXS556 and DXS8083 in Xp11.4-Xp11.3 (Boycott et al., 1998) and a similar localisation reported for a new CSNBX locus using haplotype analysis (Hardcastle et al., 1997), between RP2 and RP3 on Xp11.4 to Xp11.3 with flanking markers DXS556 and DXS8080 (interval of 5-6 cM). The interval in both reports overlaps the locus reported to contain the cone dystrophy (COD1) gene indicating that both diseases may well be allelic. The CSNB2 locus has recently been refined
using a set of families with incomplete CSNB between markers DXS722 and DXS255 (~1.2Mb) within Xp11.23 (Boycott et al., 1998). This form of CSNB was reported to be derived from certain mutations in the RP3 gene RPGR (Apfelstedt-Sylla et al., 1996).

Aland Island Eye Disease (AIED) is an X-linked form of ocular hypopigmentation, also known as Forsius-Eriksson syndrome (1964). This non-progressive condition is characterised in affected males by reduced visual acuity, reduced dark adaptation, colour vision abnormalities, infantile night blindness and high axial myopia. It has been suggested that AIED is clinically indistinguishable from incomplete CSNB (Weleber et al., 1989). Linkage analysis and subsequent refinement studies in three AIED families (Alitalo et al., 1991; Schwartz and Rosenberg, 1991; Glass et al., 1993) defined recombinant chromosomes that together localised the AIED gene between DXS7 and DXS255, a region that overlaps with the minimal region for the CSNB2 gene. This lends further support to the notion that these conditions are most likely allelic.

X-linked progressive cone dystrophy (XLPCD) is characterised by mild to severe myopia, photophobia, night blindness, loss of visual acuity, abnormal colour vision and disturbed cone electroretinogram (ERG) (Pinkers and Deutman, 1987). The COD1 locus has recently been refined to a limited region of Xp11.4 (Seymour et al., 1998). With haplotype analysis of two family branches the COD1 locus was defined as distal to DXS993 and proximal to DXS556, a distance of ~1.0Mb, which excluded COD1 as an allelic variant of RP3 or RP2 and established a novel locus sufficiently defined for positional cloning.
CHAPTER TWO

Materials and Methods
CHAPTER TWO

MATERIALS AND METHODS

2.1 Genomic DNA Preparation and Analysis

2.1.1 DNA extraction

Blood samples were collected in 10ml EDTA tubes then either extracted at the time of their arrival in the laboratory, or stored at -80°C until required. Two methods were used to extract the DNA from peripheral blood lymphocytes.

2.1.1.1 Standard protocol

Individual blood samples (approximately 10ml) were thawed slowly on ice, and their volumes increased to 25ml with sterile distilled water (SDW). Erythrocyte lysis was achieved by incubation with lysis buffer (25ml; section 2.9) for 30 minutes on ice. Intact lymphocytes were pelleted at 3500g for 15 minutes, and the supernatant reduced to 8ml by aspiration. A further 42ml of lysis buffer was added to the supernatant on ice for 10 minutes to ensure complete lysis of the erythrocytes. Intact lymphocytes were pelleted at 3500g for 15 minutes and resuspended in 3ml of suspension buffer (section 2.9). Overnight rotary incubation with 10% SDS (section 2.9) and proteinase K (0.4mg/ml) at 37°C was followed by a further 2 hour rotary incubation at 37°C with 0.6M sodium perchlorate. The suspension was extracted with an equal volume of phenol, phenol:chloroform (1:1, v/v) and twice with chloroform respectively at 5000g. The DNA in the aqueous layer was precipitated with 3M sodium acetate (0.1 volume; section 2.9) and 100% ethanol (2.5 volumes). The precipitated DNA which formed on tube inversion was removed with a sterile needle and washed in 70% ethanol. The DNA was dissolved in an appropriate volume of TE buffer (100-300μl depending on the size of the pellet) at 4°C over a period of 1-2 days.

2.1.1.2 Rapid protocol (Nucleon II DNA extraction kit, Scotlab)

Thawed blood samples were transferred to 50ml Falcon tubes and their volume increased to 50ml with reagent A (section 2.9). After mixing, the tube was centrifuged at 4000 rpm for 5 minutes, the supernatant discarded and 2ml of reagent B (section 2.9) was
added to the pellet. The mixture was then transferred to a 5ml tube and 500µl of 5M sodium perchlorate added, and mixed slowly in a rotary mix for 15 minutes at room temperature, followed by 25 minutes at 65°C. Tubes were cooled on ice, then 2ml of chloroform and 300µl of silica suspension were added. Samples were mixed for 5 minutes and then centrifuged for 6 minutes at 4000 rpm. The upper DNA-containing layer was transferred to a universal tube where two volumes of ethanol were added to precipitate the DNA. After mixing, the DNA pellet was picked out using a sterile needle and transferred to an eppendorf tube. It was then washed with 70% ethanol, allowed to dry, and resuspended in a suitable amount of sterile distilled water or TE buffer (usually 300-400µl).

Integrity and yield of genomic DNA was estimated by electrophoresis of a 2µl aliquot on a 1% agarose gel (section 2.1.3). DNA samples were kept at 4°C for subsequent use, or at -20°C for long term storage. On average the yield of DNA from 10ml of whole blood was determined by spectrophotometry to be between 250-300µg.

2.1.2 Restriction Endonuclease Digestions

Endonuclease digestions were performed on genomic DNA and cloned DNA using the appropriate buffers and at the temperatures recommended by the suppliers (BRL, NBL, New England Biolabs and Pharmacia). The amount of enzyme used and the length of digestion were determined from the assumption that one unit of enzyme digests approximately one microgram of DNA in one hour. For the preparation of Southern blots of genomic DNA, a two to five fold excess of enzyme was used and digestions were allowed to proceed for 8-16 hours. Digestion reactions were checked for completion on agarose gels prior to Southern blot preparation. Reactions in high salt buffers were often subjected to the addition of spermidine (an enzyme concentrator) to a 5mM concentration to avoid partial digestion of the DNA. Once digestion was complete, reactions were stopped by the addition of EDTA in the loading buffer.

2.1.3 Agarose Gel Electrophoresis

DNA fragments were size-separated through agarose (Sigma) gels in 1X TAE buffer (section 2.9). For optimal resolution, the concentration of the agarose gel was varied according to the average fragment sizes; genomic DNA restriction digests (1-20kb) were separated on 1.0%- 0.8% (w/v) agarose gels and PCR fragments (0.1-1.0kb) on 2.0% (w/v) agarose gels (1% ordinary agarose, 1% low melting temperature (LMT) agarose. DNA size markers were either λ/HindIII molecular standard or ØX174/HaeIII (Promega) molecular size marker. Digested genomic DNA (10µg) was loaded in 1X loading buffer (section 2.9) and was subjected to electrophoresis at 30V for 16hours in
1X TAE buffer. In the same way undigested DNA products of the Polymerase Chain Reaction (PCR) were prepared with an appropriate amount of 10X loading buffer (1/10 volume of the reaction, section 2.9) and loaded into wells. Following electrophoresis size fractionated DNA was visualized under UV light (UV transilluminator) after gels were first stained in ethidium bromide (0.5μg/ml, section 2.9) for 30 minutes and then destained in electrophoresis buffer for 15 minutes. Gels were photographed under UV light with a Polaroid MP4 camera using a Kodak, plus-X, Estar film.

2.2 Blotting Procedures

In order to identify an individual fragment in a complex mixture that has been resolved by gel electrophoresis, the technique first described by Southern (1975) using capillary transfer of DNA to a membrane has been used. Although based on the same principles, the techniques described below have since evolved from this original article and details have become membrane specific.

2.2.1 Southern (Alkali) Blotting

Agarose gels containing DNA that has been subjected to electrophoresis were firstly immersed in 0.25M HCl (section 2.9), to depurinate the DNA till the loading dyes had changed colour, and left for an additional 10 minutes. This step was only necessary for DNA fragments<10Kb in size. After rinsing in distilled water the gels were immersed in denaturing solution (section 2.9) for 30 minutes. DNA fragments were transferred from the agarose gel to a positively charged nylon membrane (Hybond N+, Amersham) by capillary action using the alkali blotting procedure described below.

Rinsed gels were placed on a wick of 3MM Whatman paper which was assembled over a reservoir of transfer buffer of 0.4M NaOH (section 2.9). Hybond N+ membranes were cut to exactly the same size as the gel, pre-treated according to manufacturers instructions and placed on top of the gel. This assembly was then covered with a further two pieces of 3MM Whatman paper soaked in transfer buffer. Care was taken throughout not to introduce bubbles between the various layers. A stack of paper towels was then placed on top to a height of several centimetres and a uniform weight (500-750g) was added. Transfer took place overnight, after which the gel lane positions were marked on the membrane which was then air dried. Since there is no need to fix DNA after alkali blotting onto Hybond N+, the blots were then used for hybridisation as described in Section 2.3.
2. Materials and Methods

2.2.2 Bidirectional Southern Blotting

Bidirectional blotting is a modification of the Southern blotting protocol which involves capillary transfer of DNA from the gel in opposite directions to two Hybond N+ membranes using 0.4M NaOH contained within the gel as a transfer buffer reservoir. The gel was treated as described in section 2.2.1 and then immersed in 0.4M NaOH for 1 hour prior to blotting. Paper towels were stacked on the bench and covered with three sheets of Whatman paper (3MM) pre-soaked in 0.4M NaOH transfer buffer. A pre-soaked sheet of Hybond-N+ membrane was then placed on top of the assembly. The treated gel was then carefully positioned on top of this membrane and another sheet of pre-soaked Hybond-N+ (0.4M NaOH) was placed on the gel. Care was taken throughout not to introduce bubbles between the various layers. Three pieces of pre-soaked Whatman paper (3MM) were positioned on the membrane and paper towels were then stacked on top. A uniform weight covered the assembly and transfer took place overnight. Following transfer, the assembly was dismantled and lane positions were marked on the two filters. Each filter was rinsed in 2 X SSC (section 2.9) to remove any residual agarose and air dried. The blots were then used for hybridisation as described in Section 2.3

2.3 Hybridisation procedures

2.3.1 Probe preparation and purification of inserts

For Southern blots, the probes used were generated by PCR from plasmid/cosmid clones, genomic DNA or retinal cDNA templates. A 25μl aliquot of the PCR product was then purified through a Sephacryl S-400 MicroSpin Column (Pharmacia) following manufacturers instructions. This was to remove any excess dNTPs, DNA oligos of less than 30bp, primer-dimer products and mineral oil present in the PCR reaction.

2.3.2 Radioisotopic oligolabelling of probes

"Oligolabelling" was developed as a method for labelling DNA restriction fragments to high specific activity, for use as hybridisation probes (Feinberg and Vogelstein, 1983). The DNA to be labelled is first denatured and then mixed with hexadeoxyribonucleotides of random sequence. These "random hexamers" anneal to random sites on the DNA and then serve as primers for DNA synthesis by the Klenow
2. Materials and Methods

fragment of E.coli DNA polymerase I. With labelled nucleotide(s) present during this synthesis, highly labelled DNA is generated. The Klenow fragment lacks the 5' - 3' exonuclease activity of DNA polymerase I, hence labelled nucleotides incorporated during oligolabelling are not subsequently excised as monophosphates. Advantages of this method are its ability to label very small quantities of DNA, which permit probes of very high specific activity to be produced, and its ability to label DNA fragments which have been separated on agarose gels without first purifying them from the agarose.

The reagents were supplied in kit form by Pharmacia. Briefly, 50-100ng of purified DNA template was denatured by incubation at 90°C for 5 minutes and then quenched on ice. The template (maximum volume of 25μl) was then mixed with unlabelled random hexanucleotides (100μM each of dATP, dGTP, dTTP) and 30μCi [α³²P]-dCTP of specific activity 6000Ci/mmol. The labelling reaction was initiated with the addition of 5-10 units of Klenow fragment in a total volume of 50μl and incubated at 37°C for 1 hour. Probes were heat denatured at 90°C for 10 minutes before use.

2.3.3 Removal of repetitive sequences from labelled DNA

For probes containing repetitive sequences, pre-annealing of the denatured DNA with sonicated human placental DNA (Sigma) was carried out. 20μl of 20 X SSC (section 2.9) and 20μl of placental DNA (9.5mg/ml) was added to the labelled DNA, boiled and chilled on ice. The mixture was then incubated at 65°C for 90 minutes and then used without denaturation.

2.3.4 Estimation of incorporation of radioisotope into the synthesised probe

After incubation, 1μl of the labelling reaction was carefully spotted onto the centre of a glass microfibre filter (Whatman GF/B) and a hand held Geiger counter clamped above it such that the meter reading was approximately 100 counts per second. The filter was then washed with 10ml of ice cold 5% (v/v) Trichloroacetic acid (TCA) which was drawn through the filter by a vacuum. The filter was then replaced in the same position underneath the monitor and the new meter reading recorded. The specific activity (SA) of the probe could then be estimated, using the following formula:-
2. Materials and Methods

\[
SA = \frac{(\mu Ci) \times (2.2 \times 10^9) \times (F)}{D_i + [(4 \times 325)(F)(\mu Ci/S)]}
\]

Where

- \(\mu Ci\) = \(\mu Ci\) of dCTP in reaction
- \(F\) = Fraction of input label incorporated into DNA
- \(D_i\) = Mass of input DNA template (ng)
- \(S\) = Specific activity of dCTP (Ci/mmol)
- 325 = Average molar mass of a deoxyribonucleotide

Thus, for 50ng of template at 50% incorporation of 50\(\mu Ci\) of dCTP, SA is approximately \(1\times10^9\) dpm/\(\mu g\) of DNA, whilst for 100ng, it drops to \(5.2\times10^8\) dpm/\(\mu g\). All probes used exceeded 50% incorporation.

2.3.5 Nucleic Acid Hybridisation to Southern Blots

Hybridisations were carried out in glass bottles or plastic boxes at 65°C in a rotating or shaking incubator respectively. Hybond N+ filters were prehybridised for at least 30 minutes at 65°C in 30 ml of a solution containing 0.3M sodium phosphate buffer pH7.2, 7% SDS, 1mM EDTA and 1% BSA (Fraction V, Sigma) (section 2.9; Church and Gilbert, 1984). 50-100ng of heat denatured oligo-labelled probe was then added to 20ml of the original prehybridisation solution covering the filters (the remainder was discarded). Hybridisation was carried out overnight at 65°C.

The next day filters were rinsed in 2 X SSC/0.1%(w/v) SDS (section 2.9) at room temperature for 10 minutes. This step was repeated, and depending on the required stringency, the filters were washed twice with 2 X SSC/0.1% SDS at 65°C for 15 minutes. This was followed by two 15 minute washes at 65°C with 1 X SSC/0.1% SDS. Stringency was further increased by two 15 minute washes at 65°C in 0.5 X SSC/0.1% SDS. Following exposure to X-OMAT AR film (Kodak) overnight, two similar washes with 0.2 X SSC/0.1% SDS were carried out if required. All filters were kept moist and not allowed to dry to prevent irreversible binding of the probe.
2.3.6 Membrane stripping protocols

Hybond-N+ membranes used for Southern blots were stripped by pouring 0.5% (w/v) SDS at 99°-95°C onto the filters and allowing to cool to room temperature. This was repeated if necessary. Satisfactory removal of the probe was confirmed by autoradiography for an appropriate time period.

2.3.7 Autoradiography

Washed filters were wrapped in cling film and exposed to Kodak X-OMAT AR autoradiographic film at -70°C with intensifying screens.

2.4 Manipulation of bacterial and bacteriophage cloned DNA

2.4.1 Vectors and bacterial strains

Two human retinal cDNA libraries were used to isolate the cDNA clones described in this thesis. These were the commercial human adult retinal cDNA library (from Stratagene), constructed in λZAPII and another human adult retinal cDNA library constructed in λgt10 (courtesy of J. Nathans). First strand cDNA was synthesised using oligo(dT) and random hexamer primers. After the addition of EcoRI linkers the cDNA was cloned into the EcoRI cloning site of λZAPII. The λgt10 library was constructed in a similar manner except that the first strand cDNA was synthesised using only oligo(dT) primers. For a description of the vectors and their uses see section 2.4.4.

Plasmid vector pDIRECT which was used for direct cloning of PCR products of second round enriched cDNA pools (Chapter 4) was supplied by Clontech. The pUC-derived linearised pDIRECT vector (Figure 5.2A, section 5.3) contains non-complementary 5' single-stranded ends (11-nucleotide long on one side and 12-nucleotides long on the other) for cloning PCR products directionally using the PCR-Direct system. In addition, pDIRECT contains the fl origin of replication for synthesis of single-stranded DNA upon co-infection with M13 helper phage. The cloning site and the multiple restriction enzyme sites flanking it are located in the middle of the lac Z gene, allowing for blue/white colour selection of recombinant plasmids. Once a gene of interest has been cloned, transcripts can be produced using the T7 and T3 promoters (Figure 5.2A, section 5.3) which flank the cloning site in opposite orientations. The pDIRECT plasmid was transformed into competent E. coli DH5α host cells (Gibco BRL or Invitrogen)
permitting $\alpha$-complementation with pUC based plasmids containing the N-terminal portion of $\beta$-galactosidase and allowing blue/white colour selection (section 2.4.2.3)

The pre-blocking (quenching) agents used in the cDNA selection procedure described in Chapter 4 were cloned into either plasmid vectors pBR322 or pGEM. pR7.3 and pR5.8: Human ribosomal RNA 45S precursor region EcoRI fragments (7.3 and 5.8kb) were cloned in pBR322. pK515: A ~7kb E.coli ribosomal RNA operon BamHI fragment in pBR322. Yeast ribosomal RNA clones of 2.6kb(H7) and 7.0kb(H15) BamHI fragments cloned in pBR322. A LINE 1 (Long Interspersed element) clone (pLA1) of 5.8kb AccI fragment cloned in pGEM vector. Both the pBR322 and pGEM vectors were transformed into E.coli DH5α host competent cells.

2.4.1 Storage and recovery of bacterial cultures in glycerol

Glycerol stocks of bacterial cultures were prepared for long term storage. To 0.5 ml of bacterial culture, 0.5 ml of 50% sterile glycerol (sterilised by autoclaving) was added in an eppendorf to give a final concentration of 25% glycerol. The culture was then vortexed to ensure that the glycerol was evenly dispersed. The culture was transferred to a labelled storage tube equipped with a screw cap and an air-tight gasket and frozen in a ethanol-dry ice mixture. The tube was then transferred to -70°C for long-term storage. To recover the bacteria, the frozen surface of the culture was scraped with a sterile inoculating needle and the bacteria streaked immediately onto the surface of an LB agar plate containing the appropriate antibiotics. The frozen bacterial glycerol stock was then returned to storage at -70°C. The plate was incubated overnight at 37°C.

2.4.2 Transformation

2.4.2.1 Preparation of competent cells using calcium chloride

A single E.coli bacterial colony of 2-3 mm in diameter (eg. of DH5α strain) was picked from a plate freshly grown for 16-20 hours at 37°C, transferred into 10 ml of LB broth (section 2.9) and grown up overnight at 37°C. 50 ml of LB-broth in a 1-litre flask was inoculated with 0.5 ml of the overnight culture. This culture was then incubated for about 2 hours at 37°C with vigorous shaking (300 cycles/minute in a rotary shaker). For efficient transformation, it is essential that the number of viable cells should not exceed $10^8$ cells/ml. To monitor the growth of the culture, the OD600 was determined every 20-30 minutes until it reached a reading of 0.4.

The culture was then aseptically transferred to sterile, disposable, ice-cold 50-ml polypropylene tubes (Falcon 2070). The cultures were then cooled to 0°C by storing the tubes on ice-water for 10 minutes and the cells were then harvested by centrifugation at
2. Materials and Methods

4,000 X rpm in a Sorvall GS3 rotor for 10 minutes at 4°C. The supernatant was discarded and the cells were then gently resuspended in 20 ml cold 0.1M calcium chloride buffer (section 2.9). Cells were left on ice for 30 minutes, then recovered by centrifugation as before at 4,000 X rpm for 10 minutes. The supernatant was decanted and the cells were gently resuspended in 2 ml of ice-cold 0.1M calcium chloride and left on ice at 4°C for about 18 hours. This is because the efficiency of transformation increases four to six-fold during the first 12-24 hours of storage and then decreases to the original level. The cells could be used at this stage for transformation. In order to store them for a longer period, an equal volume of glycerol i.e. 2 ml was added for a final concentration of 50%. The competent cells were then mixed gently, aliquoted into chilled sterile 1.5 ml microfuge tubes, then stored at -70°C for future transformations.

2.4.2.2 Transformation of bacterial strains with clone of interest

Frozen competent cells (e.g. DH5α cells) were thawed on ice immediately after removal from the -70°C freezer and used immediately after thawing. 100μl of competent cells were distributed to each ice-chilled sterile microfuge tube. 10-50 ng of DNA in a volume of 10μl or less was gently added to the cells gently in each tube. The contents in the tubes were mixed by gentle tapping. (Supercoiled pUC19 DNA and competent bacteria that received no plasmid DNA were used as an internal check of the transformation process). The tubes were then incubated on ice for 30 minutes before being transferred to a holder placed in a circulating water bath at 42°C and heat shocked for exactly 45 seconds. They were then rapidly transferred to an ice bath and the cells allowed to chill for 2 minutes before adding 900μl of SOC media (section 2.9) to each tube. The cultures were then incubated at 37°C for 1 hour with shaking (230 rpm) in order to allow the bacteria to recover and to express the antibiotic resistance marker encoded by the plasmid. An appropriate volume (up to 200μl per 90-mm plate) of transformed competent cells were spread onto L-agar plates (section 2.9) containing the appropriate antibiotic and selection components (section 2.4.2.3). The plates were left at room temperature until the liquid had been absorbed, inverted and then incubated at 37°C overnight for 12-16 hours. Plasmid DNA from bacterial colonies was prepared as described in section 2.4.3.

2.4.2.3 Growth media and selection of recombinants

LB broth (section 2.9) and L-agar (bacteriological, 15g/l) were used for the growth of all bacterial strains. Antibiotic selection of transformants was achieved with 50μg/ml ampicillin (section 2.9) (pDIRECT, pGEM and pBR322 plasmid vectors). When selecting for resistance to ampicillin, transformed cells were plated at low density (<10^4
colonies per 90-mm plate) and the plates were not incubated for more than 20 hours at 37°C, in order to avoid the appearance of ampicillin-resistant satellite colonies.

Blue/white colour selection of transformants/recombinants by the process of α-complementation was achieved by spreading on to the surface of an agar plate 40μl of a solution of X-gal (5-bromo-4-chloro-3-indolyl-β-D-galactoside) 20 mg/ml in dimethylformamide and 4μl of a solution of IPTG (isopropylthio-β-D-galactoside) 200 mg/ml in sterile, distilled water. The solvent was allowed to evaporate and the plates to dry for about an hour before plating out the transformation reaction (section 2.4.2.2).

2.4.3 Preparation of plasmid DNA

2.4.3.1 Midi-preparation of plasmid DNA

One technique for preparing plasmid DNA for digestion and sequencing reactions was a mid-range (midi) preparation technique, for the isolation and purification of larger amounts of DNA, modified from an alkaline-SDS method described by Birnboim and Doly (1979).

Colonies from a bacterial stab or glycerol stock were streaked out on L-agar plates supplemented with the appropriate antibiotic and incubated overnight at 37°C. A single colony was picked from the plate and used to inoculate 50 ml of sterile L-B broth containing the appropriate antibiotic in a sterile 200 ml flask, and incubated overnight with agitation. Glycerol stocks were prepared where an appropriate volume of culture was mixed with 0.25 volumes of 50% glycerol under sterile conditions and stored either at -20°C or -70°C. The culture was then transferred into 50 ml conical tubes (Greiner) and centrifuged at 3,000 X g for 10 minutes at 4°C. The resultant cell pellet was resuspended in 4 ml of Solution I (section 2.9) and left at room temperature for 5 minutes. Chromosomal and other macromolecular components were denatured on ice with 8 ml of Solution II (section 2.9) for 5 minutes and precipitated on ice with 4 ml of high salt Solution III (section 2.9) for 15 minutes. The sample was then centrifuged at 3,000 X g for 10 minutes at 4°C, and the supernatant strained into fresh 50 ml Falcon tubes. Plasmid DNA was precipitated from the supernatant with 0.6 volumes of isopropanol and pelleted at 3,000 X g for 10 minutes. The supernatant was removed and the pellet was dissolved in 4 ml of TE buffer (section 2.9) and extracted with an equal volume of phenol. DNA was then precipitated with 0.1 volumes of 3M sodium acetate (section 2.9) and 2 volumes of absolute ethanol.

The DNA was pelleted at 10,000 X g and resuspended in 400μl of TE buffer (section 2.9). At this stage contaminating RNA and protein was digested; firstly by incubation with RNAses A (0.5 mg/ml) and 3M sodium acetate (15μl) at 37°C for 30 minutes. An equal volume of phenol, phenol:chloroform (1:1, v/v) and chloroform
respectively were used to extract the plasmid DNA and the DNA was precipitated from
the resultant aqueous layer with 2.5 volumes of absolute ethanol and 0.1 volume of 3M
sodium acetate. The DNA was pelleted at 14,000 X g for 5 minutes, and washed in 70%
ethanol before air drying and resuspension in an appropriate volume of TE buffer
(section 2.9). DNA yield was estimated by agarose gel electrophoresis of 2μl of the DNA
preparation along with known quantities of molecular weight standards (λ, Hind III).

2.4.3.2 Mini-preparation of plasmid DNA

The method described below is appropriate for preparation of DNA from 5 ml
cultures of plasmid containing bacteria, and is essentially a scaled down version of the
midi-preparation method (Section 2.4.3.1).

5 ml of L-B broth supplemented with the appropriate antibiotic was inoculated
with a single colony, obtained as described in section 2.4.3.1 and incubated in a sterile 50
ml Falcon tube at 37°C overnight with agitation. 1.5 ml of this culture was transferred to
an eppendorf tube and spun in a microfuge for 3 minutes at 13,000 X g. The supernatant
was removed and the pellet resuspended in 100μl of Solution 1 (section 2.9). 200μl of
freshly prepared Solution II (section 2.9) was added and mixed gently with the
resuspended pellet (without vortexing) and then placed on ice for 5 minutes. 150μl of
Solution III (section 2.9) was then added and mixed thorougly. The sample was placed
on ice for at least 5 minutes followed by centrifugation at 13,000 X g for 5 minutes. The
supernatant was removed to another eppendorf tube, and an equal volume of a
phenol:chloroform (1:1, v/v) mixture was used to extract the plasmid DNA, after
vigorous shaking and centrifugation in a microfuge for 5 minutes at 13,000 X g. The
DNA was precipitated from the resultant aqueous layer with 0.1 volume of 3M sodium
acetate and 2.5 volumes of cold absolute ethanol, placed at -20°C for about 30 minutes.
The DNA was pelleted as before, washed with 70% ethanol, air dried and resuspended
in 20μl TE buffer. 2μl of the preparation was subjected to electrophoresis along with size
markers on a 0.8% agarose gel to check the quality of DNA preparation and to estimate
the yield of plasmid DNA.

Another method using the commercial Magic/Wizard Minipreparation kits
(Promega, USA) was available at a latter stage and was preferably used to the Birnboim
and Doly method, due to the use of a DNA recovery resin in comparison to the phenol
chloroform method of extraction. 5 ml of a culture grown overnight was pelleted in a 50
ml Falcon tube at 2,500 rpm, 4°C for 5 minutes in an MSE Mistral 3000i using a swing-
out rotor. The supernatant was discarded and the cell pellet resuspended in 300μl of Cell
resuspension solution (section 2.9) by pipetting several times and then transferred to a
1.5 ml eppendorf tube. 300μl of Cell lysis solution (section 2.9) was added to the cells and
shaken well for a minute. 300μl of Neutralisation solution (section 2.9) was then added to
2. Materials and Methods

the lysed cells and shaken well again. The neutralised cell lysate was left on ice for about 15-20 minutes and centrifuged at 13,000 rpm for 10 minutes. While the cell lysate was being centrifuged, 0.5 ml of DNA recovery resin (Wizard Miniprep DNA Purification System; Promega, USA) was well shaken and aliquoted into a fresh 1.5 ml eppendorf. The clear supernatant from the spun cell lysate was transferred into the DNA recovery resin and mixed well by inversion.

A mini-prep column was placed on the vacuum manifold and a 2 ml syringe barrel into the column. The DNA recovery resin/cell lysate supernatant mixture was added to the syringe barrel and drawn through by vacuum filtration. When no liquid was left in the syringe barrel, 2 ml of Column Wash (section 2.9) was added and drawn off by vacuum filtration. This step was repeated, the syringe barrel removed and the column left under vacuum for about 30 seconds. The column was then placed in a fresh 1.5 ml eppendorf and spun at 13,000 rpm in a microfuge for 2 minutes. The column was then placed in a fresh 1.5 ml eppendorf and 30μl of sterile preheated (75°C) distilled water was added to the resin in the column and allowed to stand for 60 seconds. The column was then spun at 13,000 rpm for 20 seconds. A further 35μl of preheated water was added to the column and allowed to stand for 60 seconds before being spun at 13,000 rpm for 30 seconds. The column was removed and discarded, and the quality and quantity of the mini-prep DNA checked by running 2-5μl on an agarose gel along with known quantities of molecular weight standards (λ Hind III).

2.4.4 Manipulation of the λZAPIII and λgt10 human adult retinal cDNA libraries

λZAPIII is an insertion vector equipped with multiple cloning sites within plasmid sequences that can be excised in vivo in the presence of f1 or M13 helper bacteriophages and converted to a plasmid vector, Bluescript SK (M13-). The plasmid vector is designed to facilitate DNA and RNA sequencing and synthesis of RNA probes utilizing bacteriophage T7 and T3 promoters flanking the cloned fragment. DNA fragments up to 10 kb in length can be inserted in unique XhoI, EcoRI, SpeI, XbaI, NsoI or SacI sites. XL1-Blue E.coli. strain was used during propagation of vector. This strain supports vigorous growth of λZAPIII. Since this strain is Rec-, its use may be advantageous if the bacteriophage λ clone is carrying rare cDNAs with repetitive sequences that can be deleted or rearranged by recombination between homologous sequences. XL1-Blue also has the added benefit of allowing improved colour selection of recombinant versus non-recombinant clones. As this strain carries a tetracycline marker on the F' episome it is essential to start liquid cultures from colonies grown on agar plates containing 25μg/ml tetracycline (section 2.9).
λgt10 is designed for cloning small (~6kb) cDNA fragments into the unique EcoRI sites in the immunity region. Cloning efficiency is high, and recombinants were selected using the C600hfl A (BNN102) E.coli strain which is tetracycline resistant. The high frequency lysogeny mutation (hfl A150) suppresses plaque formation by parental cl+ bacteriophages but allows plaque formation by recombinant cl- bacteriophages. Recombinants retain red and gam, which allows propagation on recA hosts.

2.4.4.1 Preparation of host cells.

50 ml of sterile L-B supplemented with 1 ml of 10% filter sterilised maltose (0.2%) and the appropriate antibiotic was inoculated with a single bacterial colony. Bacteria was grown in the presence of maltose as this helps the bacteria adsorb bacteriophage λ more efficiently. The sugar induces the maltose operon, which contains the lam B gene that codes for the bacteriophage λ receptor. The culture was grown overnight at 37°C with moderate agitation.

The next day the cells were centrifuged at 4,000 X g for 10 minutes at room temperature. The supernatant was discarded and the cell pellet resuspended by vortexing in ~20 ml of sterile 10mM magnesium sulphate. Magnesium ions are necessary for adsorption and phage integrity. The cell suspension was either used immediately or stored at 4°C for up to 3 weeks, before use. The highest plating efficiencies were obtained when freshly prepared cells were used.

2.4.4.2 Titration of λZAPII and λgt10 phage stock DNA

A plaque is derived from infection of a single bacterium by a single virus particle. The progeny particles infect neighbouring bacteria, which in turn release another generation of daughter virus particles. The result of successive rounds of infection is a spreading zone of lysis, visible to the naked eye as a clear area in an otherwise turbid area of bacterial growth. In this way it is possible to determine the population of bacteriophages since each plaque contains the progeny of a single bacteriophage particle. Most phage lysates contain between 10^9 to 10^11 plaque forming units (pfu) per ml.

Tenfold serial dilutions of bacteriophage stocks of both the λZAPII and λgt10 cDNA libraries were prepared in SM buffer (section 2.9). 100μl of each dilution of 10^2, 10^3, 10^4, 10^5 and 10^6 was dispensed into sterile tubes (13mm X100mm). Then 100μl of plating bacterial cell suspension in 10mM MgSO4 (section 2.9) was added to each tube. The contents in the tubes was mixed by shaking and incubated for 20 minutes at 37°C, to allow the bacteriophage particles to adsorb to the bacteria. Meanwhile 3 ml of molten top agarose (0.7%) was aliquoted into sterile tubes and equilibrated at 50°C in a water bath for 15-20 minutes. After the 20 minutes incubation the molten agarose was added to the
2. Materials and Methods

adsorbed bacterial/phage suspension, mixed and quickly poured evenly onto a pre-warmed (37°C) plate. This was repeated for second and subsequent tubes on fresh plates. L-agar plates to be used were either stored at room temperature for 2-4 days before use or were allowed to dry for 1 hour in a 37°C incubator with the lids ajar. L-agar plates with tetracycline (15µg/ml) were used for plating of the λZAPII and λgt10 bacteriophage stock using XL1-Blue and C600 hfl A (BNN102) cells respectively. After the top agarose had set, plates were incubated at 37°C. Plaques began to appear after about 7 hours of incubation and were counted or picked after 12-16 hours.

2.4.4.3 Amplification and preparation of plate lysate stocks of the bacteriophage λgt10 library.

The λgt10 adult retinal bacteriophage cDNA library was amplified and plate lysate stocks prepared since only an aliquot of this library was available (Gift from J. Nathans). Aliquots of the library containing ≥1kb cDNA inserts and ≤1kb cDNA inserts were amplified and lysate stocks prepared separately. The titre of the original aliquots was determined as described in section 2.4.4.2.

In order to have a representation of 1 X 10^6 recombinants of the library, 1 X 10^5 recombinants were plated out onto ten 20cm by 20cm L-agar tetracycline (15µg/ml) plates. This was done in two separate lots using the two different sized aliquots of the λgt10 cDNA library. Freshly poured L-agar tetracycline (15µg/ml) plates, dried and equilibrated to room temperature were used. Aliquots of the diluted λgt10 stock containing 1 X 10^5 bacteriophage particles in a volume of 250µl were mixed with 600µl of the C600 hfl A (BNN102) plating bacterial cells (Section 2.4.4.1) and incubated for 20 minutes at 37°C. 15 ml of molten (47°C) top agarose (0.7%) was then added to the infected bacteria, swirled gently and then poured onto an individual plate. This was repeated for all ten plates. Once the agarose had set the inverted plates were incubated at 37°C for about 7-8 hours or until the plaques had just begun to make contact with each other.

The plates were removed from the incubator and 30 ml of SM buffer pH 8.0 (section 2.9) was added to each plate and stored at 4°C for about 3 hours with intermittent gentle shaking. Using sterile pasteur pipettes the SM buffer containing the bacteriophage particles was harvested into sterile 50 ml Falcon tubes. Another 10 ml of SM buffer was added to each plate and stored for 30 minutes in a tilted position to allow all the fluid to drain to one area. This was again harvested and combined with the first harvest. 500µl of chloroform was added to the pooled SM buffer, vortexed briefly and centrifuged at 4,000 X g for 10 minutes at 4°C. The supernatant was then transferred to another sterile Falcon tube and 100µl of chloroform was added and stored at 4°C, for long-term storage of the stocks. The titre of the amplified stock was determined as
described in section 2.4.4.2, and usually remained unchanged as long as the stock is stored at 4°C.

### 2.4.5 Immobilization of bacteriophage λ plaques onto Hybond-N+

This method is modified from that of Benton and Davis (1977). Bacteriophage λ plaques were obtained on 20cm x 20cm plates as described in section 2.4.4.3. The plates were chilled at 4°C for at least 1 hour to allow the top agarose to harden. They were then removed to room temperature and a labelled Hybond-N+ filter was placed on the surface of the top agarose so that it comes in direct contact with the plaques for about 3 minutes. Orientation marks were made with a sterile needle, stabbing it through the filter and the agar beneath it. The filter was carefully removed using blunt-ended forceps and placed DNA side up into a tray containing a saturated piece of Whatman 3MM paper saturated with denaturing solution (section 2.9) for 5 minutes. This was repeated using neutraliser (section 2.9) for 5 minutes and 2 X SSC (section 2.9) for 1 minute. The filter was then air-dried and the DNA was fixed to the membrane by UV cross-linking (5 minutes at a wavelength of 254nm). The entire process was repeated for the production of replica filters, making sure that the orientation marks were correctly placed relative to those of the first filter.

### 2.5 Yeast Artificial Chromosomes (YAC) Methods

#### 2.5.1 Yeast strain and Vector

The ICRF (Larin et al., 1991) and ICI (Anand et al., 1990) human 48, XXXX YAC library were used in these studies. The cloning vector used to construct the libraries is pYAC4 which comprises the cis-acting yeast chromosomal sequences joined to segments of the bacterial plasmid vector pBR322. This allows propagation of the YAC vector in E. coli., by standard procedures.

#### 2.5.2 Preparation of YAC DNA in plugs

YAC clone DNA was prepared according to Vollrath and Davies (1987), with slight modification. Yeast clones were streaked out onto AHC agar plates (section 2.9) and incubated at 30°C overnight. A single large colony was picked into 10ml AHC broth (section 2.9) and grown at 30°C for 24 hours; the culture was subsequently used to seed a 100ml or 200ml culture in AHC broth. This was grown with agitation at 30°C for 40
hours after which the O.D.600 was measured and the number of cells were calculated (O.D.600 of 0.3 corresponds to $3.3 \times 10^6$ cells/ml).

Cells were pelleted by centrifugation for 10 minutes at 4K, and resuspended in 50ml SCE solution (section 2.9). The cells were re-pelleted and resuspended in SCEM (SCE with 30mM β-mercaptoethanol, made fresh each time) to a final concentration of $2.5 \times 10^9$ cells per ml. Lyticase (Sigma) was added to a final concentration of 120μg/ml and incubation was performed at 30°C for 1 hour. Low melting point (LMP) agarose (1%) was prepared in 1M sorbitol (section 2.9) and kept molten at 55°C. Yeast cells (0.5ml) were mixed with an equal volume of LMP agarose and 200μl aliquots were dispensed into plastic moulds and left to set at 4°C.

The plugs were placed in proteinase K/sarcosyl solution (section 2.9) and incubated at 50°C for 2 days. Following extensive rinsing with TE buffer, proteinase K was inactivated by treatment with 0.04mg/ml PMSF (phenyl-methyl-sulphonyl-fluoride) in TE (section 2.9) with incubation at 50°C for 30 minutes. The plugs were then rinsed three times in TE buffer and stored at 4°C.

### 2.5.3 Pulsed Field Gel Electrophoresis (PFGE)

Pulsed field gel electrophoresis is a powerful technique for resolving DNA molecules in the megabase size range (Schwartz and Cantor, 1984), and enables large regions of genomic DNA to be mapped and analysed. Its wide separation range has bridged the size gap between conventional gel electrophoresis (up to 50kb) and cytogenetic techniques (5Mb). It is hence capable of separating chromosome size DNAs from the yeast genome.

In PFGE, DNA molecules undergo continuous re-orientation caused by periodic changes in alternating electric fields. The duration of the alternating electric field is known as the switch time or pulse time. Each time the electric field is switched, the DNA molecules must re-orientate or change their direction of migration in the gel matrix. Larger DNA molecules require longer times to re-orientate with each change of the electric field whereas smaller molecules that can re-orientate quickly, will spend a larger portion of each switch interval migrating with the field. Hence, as the size of the DNA molecules increases, the pulse time must be increased to resolve these molecules.

The experiments described here were conducted on the CHEF-DR II apparatus (Biorad) based on the CHEF (clamped homogeneous electric fields) technique (Carle and Olson, 1984)
2.5.3.1 Casting and loading the gel

100ml of 1% agarose (Molecular Biology Certified; Biorad) in 0.5 X TBE buffer (section 2.9) was poured into the casting stand (standard casting stand, Biorad) and allowed to set for 1 hour at room temperature. Sample plugs were loaded while the gel remained in the casting stand. Care was taken so that samples were less than 90% of the height of the well. They were then sealed in place by filling each sample well with 1% low melting agarose (Biorad). The agarose was allowed to solidify for 15 minutes.

2.5.3.2 Gel electrophoresis

The electrophoresis chamber was rinsed briefly with two litres of pre-cooled distilled water. Two litres of pre-cooled 0.5 X TBE electrophoresis buffer was poured into the chamber and allowed to circulate briefly through the recirculating pump. The gel was placed in the centre of the chamber and the buffer flow through the pump was adjusted to 1L/hour so that the gel was not disturbed during electrophoresis. The following electrophoresis parameters and voltage were selected to separate DNA in the range 450kb-225kb (accurate operating instructions can be found in the Biorad CHEF-DR II manual):  
a) Pulsewave 760: Initial A Time: 25 seconds, Final A time: 25 seconds (note: this pulse mode is called ramping). b) Model 200/2.0 power supply: Voltage : 175V, Run Time: 20 hours. After the first run a) Pulsewave 760 was set again to: Initial A Time: 45 seconds, Final A Time: 45 seconds, b) Voltage: 175V, Run Time: 20 hours. Gels were run in the cold room to ensure the temperature was maintained at approximately 14°C during electrophoresis.

2.5.3.3 Gel staining

After the 40 hour run the pulse field gel was placed into 0.5μg/ml ethidium bromide solution (freshly prepared) and stained for 30 minutes. It was then destained in distilled water for 15 minutes, visualised on a UV transilluminator and photographed.

2.5.4 Isolation of YAC clones from Pulse Field gels

YAC DNA which needed to be separated and isolated from Pulse Field gels was run in 1% low melting point agarose (Molecular Biology Certified; Biorad). Casting, loading and gel electrophoresis was carried out as described in section 2.5.3.1 and section 2.5.3.2 respectively. The required YAC band was then cut out of the ethidium bromide
stained gel under UV light. The gel slice was then put into a small petri-dish containing a little TE buffer pH 8.0 and stored at 4°C for further use.

2.6 The Polymerase Chain Reaction (PCR)

2.6.1 General considerations

Primer design, thermal cycling conditions, dNTP concentration and magnesium concentration were initially based on parameters reviewed by Innis and Gelfand (1990) and Saiki (1990). Different sets of primers were individually optimised, based upon their 3' complementarity and primer melting temperatures (Tm) calculated by using the formula

\[ 4(G + C) + 2(A + T) = Tm \] (melting temperature)

then reducing the annealing temperature by 2°-5°C (for primers up to 20 bases in length), or, more accurately for longer primers by using computer software to analyse primer pairs (see section 2.8.4).

Cycle length and number varied between 30 and 40 cycles, depending upon the template, particular methodology, and the model of thermocycler used. These will be noted where necessary. Once optimised, reactions were generally performed on the same machine, it being one of the following models: a Hybaid Thermal Cycler, a Hybaid OmniGene, or a Perkin Elmer GeneAmp PCR system 9600.

2.6.2 Standard parameters for a typical PCR

Unless otherwise stated, the following constitute a 'standard' PCR reaction in a 50μl volume. PCR amplifications were performed on 150ng of genomic DNA in 1 X Taq reaction buffer (NBL, Bioloine, Promega: 10mM Tris-HCl, 50mM KCl, 1.5mM-2.0mM MgCl2 and 0.1% non-ionic detergent), 0.2mM of each dNTP, 25 pico moles of each primer, and 0.3-0.5 units of Taq polymerase (NBL, Bioloine, Promega). The reaction mixtures were overlaid with one drop of mineral oil to prevent evaporation. After a routine initial denaturation step of 95°C for 3 minutes the cycling conditions were:
denaturation at 94°C for 5-30 seconds, annealing at the appropriate temperature for 30 seconds and extension at 72°C allowing for approximately 1 minute per kilobase extension time.
2. Materials and Methods

2.6.2.1 PCR from a bacteriophage plaque

Heat denaturation of a phage lysate destroys the bacteriophage protein coat releasing DNA into the surrounding media, facilitating the use of PCR to amplify DNA inserts of interest. Isolated plaques were stabbed into separate eppendorf tubes which contained a few ml of SM buffer and a small drop of chloroform and stored at 4°C. An aliquot (3μl) of the phage lysate was used as a template for PCR as described in section 2.6.2.

2.6.2.2 PCR from a bacterial colony

Heat denaturation of a bacterial colony destroys the bacterial cell wall releasing plasmid DNA into the surrounding media. The DNA insert of interest can then be amplified by PCR. Half of the bacterial colony of interest was picked into the PCR reaction tube with a toothpick and PCR was carried out as described in section 2.6.2. The rest of the bacterial colony was stored on L-agar plates at 4°C for future use.

2.6.2.3 PCR from YAC plug DNA

YAC DNA isolated in plugs (see section 2.5.2) was used in PCR by cutting a small piece of the plug DNA using a clean scalpel and transferring it into a 0.5μl microfuge tube. The agarose plug was then melted at 95°C for 10 minutes. An equal volume of water was added and an aliquot of 3μl was used as template DNA in the PCR (section 2.6.2).

2.7 Analysis of microsatellites

Microsatellite markers were selected from the proximal short arm of the X chromosome in and around the RP2 and RP3 critical regions (section 3.2.2) for haplotype analysis in XLRP families (section 3.1.4). Primers for PCR amplification of microsatellites were obtained from either the HGMP resource centre or Cruachem.

2.7.1 5'-end labelling of primers

Primer (10pmoles) was labelled using [γ-^32]ATP in a final volume of 10μl by adding the following; 1μl of T4 polynucleotide kinase (5-10 units; Pharmacia), 2μl of [γ-^32]ATP (specific activity 6000 Ci/mmole; Amersham) and 1μl of One-Phor-All buffer (Pharmacia). The reaction mix was incubated at 37°C for 45 min.
The end-labelled primer was used in PCR reactions carried out in individual volumes of 25μl. The reaction mix comprised: 3μl diluted genomic DNA (~ 150ng), 1 X BioTaq buffer (Bioline), 0.2mM dNTP's (final concentration), 0.2μM (final concentration) of each of the forward and reverse primers (one of which being end-labelled) and 0.3 - 0.5 units of BioTaq polymerase (Bioline). PCR amplification was carried out as described in section 2.6.2 in an Omnigene Thermal Cycler.

After PCR amplification, 10μl of formamide loading buffer (section 2.9) was added to each reaction. Samples were denatured at 95°C for 4 minutes immediately prior to running on 6% denaturing acrylamide gels. Gels were run at 100 Watts using a BioRad 48X60 cm sequencing apparatus (Section 2.7.2) for the required length of time to obtain good resolution of each microsatellite marker. Gels were processed and autoradiographed as described in section 2.7.2.

### 2.7.2 Polyacrylamide denaturing gel electrophoresis

Apparatus used for running denaturing gels was purchased from BioRad. The gel plates were washed extensively, rinsed with ethanol and air dried. The back plate (holding the buffer reservoir) was silanised with SigmaCote (Sigma) before assembly (according to manufacturer's instructions). Acrylamide concentrate (19:1 acrylamide:bisacrylamide, Sequagel, National Diagnostics) was mixed with diluent (8.3M Urea; Sequagel, National Diagnostics) and 10 X TBE buffer to give a 6% denaturing gel solution.

For 200mls of 6% polyacrylamide gel the following were mixed thoroughly: 132mls of 8.3M urea solution, 20ml of 10 X TBE and 48mls of 25% acrylamide concentrate. 50mls was mixed with 300μl of 25% ammonium persulphate and 300μl TEMED and immediately poured in the casting tray to form a sealing gel at the base of the plates. The remaining 150ml of the acrylamide solution was mixed with 600μl of 25% ammonium persulphate and 60μl TEMED and slowly poured into the plates along one side at a 45°angle. The angle of the plates and the rate of flow were adjusted so that the acrylamide ran in smoothly without bubbles. A comb was inserted carefully in the top end of the plates and the gel was allowed to polymerise for approximately 45 minutes.

After the gel had polymerised, the apparatus was removed from the casting tray and placed in the electrophoresis tank. The lower and upper reservoirs were filled with 1 litre of 1 X TBE buffer. The gel was pre-run at a constant power of 100 Watts until the temperature of the gel had reached 50°C, then the comb was removed. The samples were then denatured at 90°C for 5 minutes. Excess urea was flushed from the wells and an aliquot (2-3μl) of the labelled PCR product (section 2.7.1) was loaded into the wells. The gel was subjected to electrophoresis at a constant power of 100W for the required length of time in order to obtain good resolution of each of the microsatellite markers.
After electrophoresis, the two plates were carefully separated. The gel adhered to the front plate, which was then placed in a tray and gently covered with a fixative solution (10% methanol, 10% glacial acetic acid) for 5-10 minutes. Using a piece of 3MM Whatman paper the gel was gently peeled away from the plate, covered with plastic wrap, placed between two more pieces of Whatman paper and vacuum dried at 80°C for 1 hour. The gel was then exposed to Kodak X-OMAT film at room temperature without an intensifying screen or at -70°C with an intensifying screen.

2.8 Sequencing double stranded DNA

2.8.1 Overview of dideoxy chain termination sequencing

Sequencing reactions were carried out using T7 polymerase (Pharmacia) which is a genetically engineered T7 DNA polymerase lacking 3' to 5' exonuclease activity. The dideoxy protocols described are modified from Sanger et al., (1977).

The ddNTPs lack a hydroxyl residue at the 3' position of the deoxyribose, which prevents formation of a phosphodiester bond when incorporated into the elongating DNA molecule and leads to chain termination. By using four different ddNTPs in four separate reactions, populations of oligonucleotides are generated that terminate at positions occupied by every A, C, G or T in the template strand. Sequencing reactions were carried out on double stranded plasmid DNA prepared as described in the section 2.8.2.1 using SP6, T7 and T3 universal primers initially, followed by insert specific oligonucleotides as required.

2.8.2 Dideoxy sequencing from a cloned product

2.8.2.1 DNA Purification and annealing of primer to double stranded template

Double stranded plasmid DNA (approximately 1-2μg) in a volume of 32μl was alkali denatured in 8μl of 2M sodium hydroxide at room temperature for 10 minutes. The DNA was then precipitated using 120μl of absolute alcohol and 7μl of 3M sodium acetate (pH 4.8) at -20°C for 30 minutes, and recovered by centrifugation at 13,000 rpm in a microcentrifuge for 20 minutes. The resultant pellet was washed in 70% alcohol, dried and resuspended in 10μl water.
2. Materials and Methods

DNA sequencing was carried out using a Pharmacia sequencing kit as follows: 2μl of primer (10μM) and 2μl of annealing buffer was then added to the resuspended template, mixed, spun briefly and incubated at 65°C for 5 minutes. The tube was then transferred to a 37°C waterbath and incubated for 10 minutes before being placed at room temperature for 5 minutes, centrifuged briefly and then used in the sequencing reactions as described below (Section 2.8.2.2).

2.8.2.2 Sequencing using T7 (Pharmacia) kits with αS^{35} incorporation labelling of primer

2μl of T7 DNA Polymerase (diluted 1:4 in cold enzyme dilution buffer), 3μl of labelling mix-dATP diluted 1:4 in 330mM NaCl and 1μl of [αS^{35}]-dATP were added to the annealed template. The reaction was incubated at room temperature for 5 minutes. During this time, four tubes with 2.5μl of termination mix (A, C, G and T) were prepared and prewarmed at 37°C for at least 5 minutes. Following labelling, 4.5μl aliquots of the labelled reaction mix were transferred into each of the four sequencing mixes and the samples incubated at 37°C for 4 minutes. The reactions were terminated with 5μl of stop solution (Pharmacia sequencing kit, U.K.). The samples were analysed on 6% polyacrylamide denaturing gels (Section 2.7.2) after denaturation at 80°C for 2 minutes. The remaining material was stored at -20°C and heated just prior to use.

2.8.3 Direct genomic sequencing of PCR products

2.8.3.1 DNA purification

The sequence of interest was amplified from genomic DNA or cDNA with specific primers synthesised from cDNA clone sequence. A 50μl-100μl of the PCR product was then purified through a Sephacryl S-400 spin column (Pharmacia, UK) following manufacturer’s instructions.

2.8.3.2 γP^{32} 5'-end labelling of primer

Primer (10pmoles) was labelled using [γP^{32}] dATP in a final volume of 10μl by adding the following; 1μl of T4 polynucleotide kinase (5-10 units; Pharmacia, UK), 2μl of [γP^{32}] dATP (3000Ci/mM) and 1μl of One-Phor-All buffer (Pharmacia, UK). The reaction mix was incubated at 37°C for 1 hour.
2.8.3.3 Sequencing using T7 (Pharmacia) kits

An aliquot of DNA (10μl) was mixed with 2μl (2.5pmol) of labelled primer (Section 2.8.3.2) and heated to 95°C for 2 minutes and then cooled on ice. The following were then added on ice:- 2μl of annealing buffer (Pharmacia, UK) and 2μl of T7 Polymerase (1:4 dilution in enzyme dilution buffer). The reagents were then mixed gently before 3.5μl-4.0μl aliquots were removed to be mixed with 2.5μl aliquots of each dideoxy nucleotide termination mix. The samples were then incubated at 37°C for 5 minutes before the reaction was terminated by the addition of 4μl of stop solution (Pharmacia, UK). The samples were then analysed on a 6% polyacrylamide denaturing gels (Section 2.7.2).

2.8.4 Database usage and computational analysis of sequence information

Access to a variety of databases required during this project was made easier through the UK Human Genome Mapping Project Resource Centre (UK HGMP-RC). The HGMP-RC computer facilities provide registered users with a simple way to access a wide range of genetics and molecular biology software packages and databases.

The Genome Data Base (GDB) stores gene mapping information including primer sequences and allele frequencies for a large set of human microsatellite mapping markers (Dib et al., 1996; section 1.2.1). Information in GDB can be cross-referenced to OMIM (On-line Mendelian Inheritance in Man), using the direct searching link, via chromosome location and MIM numbers which appear in both GDB and OMIM databases. OMIM stores text information about inherited diseases and is an on-line version of the book "Mendelian Inheritance in Man" by Victor McKusick plus the McKusick Human Gene Map and Molecular Defects List. From the HGMP-RC main menu the option Genome Data gave access to the combined databases GDB/OMIM.

GenBank which is the NIH's database of all known nucleotide and protein sequences including supporting bibliographic and biological information was used for sequence analysis of cDNAs (section 5.5.2). The GenBank sequence database has undergone an expansion in data coverage, annotation content and the development of new services for the scientific community as described in section 1.5.10. In addition to nucleotide sequences, data from major protein sequence and structural databases and from U.S. and European patents is now included in the integrated system (Benson et al., 1993). MEDLINE abstracts from published articles describing the sequences also provide an important new source of biological annotation for sequence entries. Entries include a concise description of the sequence, scientific name and taxonomy of the source
organism, and a table of features specifying coding regions and other sites of biological significance.

High-speed heuristic methods, such as the hash-coding (k-tuple) algorithm employed by FASTA (Pearson and Lipman, 1988; Pearson, 1990) and the approximate word match algorithm employed by BLAST (Altschul et al., 1994) as described in section 1.5.10 were the commonly used sequence database search programs in this study (sections 5.5 & 5.6). BLASTN compares a DNA sequence against DNA sequence databases, and uses a positive score for matching residues and a penalty for mismatches. For an initial quick search BLASTN was used. These programs produced a list of the sequence identifiers (eg. locus names and accession numbers) and title lines of statistically significant matches followed by a display of alignments of the query with each of the matched sequences as shown in section 5.6. The "Analysis and Manipulation of Sequences" option from the HGMP-RC main menu provided programs that allowed for the acquisition, analysis and manipulation of nucleic acid sequences. The BLASTN program which was used initially for sequence analysis in this study (sections 5.5 & 5.6) only found un-gapped locally optimal sequences. The newly developed Gapped BLAST algorithm (Altschul et al., 1997, section 1.5.10) which allowed for gaps to be introduced into the alignments that are returned was used for further sequence analysis. The worldwide web provided easy access to the Gapped BLAST search tool provided by the BLAST server 2.0.

2.9 Reagents and buffers used in this research

2.9.1 Stock reagents in general use

2M Tris HCl (trizma hydrochloride pH 7.5, 7.8 or 8.0): adjust pH to 7.5 or 8.0 with 5M NaOH, autoclaved before use.

0.5M EDTA (ethylenediaminetetraacetic acid pH 8.0): adjust pH to 8.0 by addition of 5M NaOH, autoclaved before use.

TE buffer: 10mM Tris-HCl (pH 7.5), 1mM EDTA (pH 7.5-8.0), autoclaved before use.

5M NaOH (sodium hydroxide): prepared with sterile distilled water (SDW).

10% SDS (sodium dodecyl sulphate): prepared with SDW.

20XSSC: 3M NaCl, 0.3M tri-sodium-citrate.

3M sodium acetate: prepared in sterile distilled water and autoclaved.
2. Materials and Methods

2.9.2 Preparation of DNA

Human genomic DNA

10X lysis buffer: 50mM Tris-HCl (pH 7.5), 25mM MgCl₂.6H₂O, 0.6M sucrose, 5% (v/v) Triton X100

10X suspension buffer: 0.1M Tris-HCl (pH 7.5), 0.1M NaCl, 0.1M sodium EDTA (pH 8.0).

0.6M sodium perchlorate (prepared with SDW, autoclaved before use)

Reagent A (5X lysis buffer): 320mM sucrose, 10mM Tris HCl (pH 7.5), 5mM MgCl₂ (adjusted pH to 8.0 with 5M NaOH), autoclaved then added 1% Triton X-100.

Reagent B: 400mM Tris-HCl (pH 7.5), 60mM EDTA. 150mM NaCl (adjusted pH to 8.0 with 5M NaOH), autoclaved then added 1% SDS.

Plasmid DNA

LB broth: 1% (w/v) bactotryptone, 0.5% (w/v) yeast extract, 1% (w/v) NaCl, autoclaved before use.

LB agar: LB broth containing 1.5% (w/v) agar, autoclaved before use.

Ampicillin: 50mg/ml stock in SDW; plasmids-1μl per 1ml media (f.c. 50μg/ml, NB: heat-sensitive)

Tetracycline: 10mg/ml stock in ethanol; plasmids-2.5μl per ml of media (f.c. 25μg/ml, NB: light-sensitive)

Stock solutions of antibiotics dissolved in SDW were sterilised by filtration through a 0.22-micron filter. Antibiotics (tetracycline) dissolved in ethanol need not be sterilised, but are stored in light-tight containers.

Solutions used in conventional plasmid DNA miniprep method

Solution I: 50mM glucose, 25mM Tris-HCl (pH 8.0), 10mM EDTA, prepared from 0.5M EDTA and 2M Tris-HCl stocks, autoclaved before use.

Solution II: 0.2M NaOH, 1% (w/v) SDS (prepared from 5M NaOH and 10% SDS stocks).

Solution III: 3M potassium acetate 2M acetic acid, autoclaved before use.

Solutions used in plasmid DNA Magic/Wizard miniprep method

(Promega.USA)

Cell resuspension solution: 50mM Tris-HCl, pH 7.5, 10mM EDTA and 100μg/ml RNase A.

Cell lysis solution: 0.2M NaOH and 1% SDS.

Neutralisation solution: 1.32M potassium acetate, pH 4.8.

Column wash solution (concentrations prior to ethanol addition): 200mM NaCl, 20mM Tris-HCl, pH 7.5 and 5mM EDTA. Diluted with 95% ethanol, final ethanol concentrations will be approximately 55%.
2. Materials and Methods

Yeast DNA

AHC broth: 6.7g yeast nitrogen base, 10g casein hydrolysate, 20g D-glucose, 20mg adenine hemisulphate, made up to 1 litre with water. Autoclaved before use.

AHC agar: AHC broth containing 1.5% (w/v) agar, autoclaved before use.

SCE buffer: 1M sorbitol, 0.1M NaCl, 10mM EDTA, autoclaved before use.

SCEM buffer: SCE with 30mM β-mercaptoethanol, freshly prepared.

1M sorbitol: autoclaved before use.

Proteinase K/sarcosyl solution: 1% sodium lauryl sarcosine (SLS), 0.5M EDTA pH 8.0, 2mg/ml proteinase K.

10mM stock PMSF (phenylmethylsulphonylfluoride): 17.4mg PMSF in 10ml isopropanol (i.e. 1.74mg/ml). 1.14ml of stock solution were added to 50ml TE buffer to give a final concentration of 0.04mg/ml.

2.9.3 Gel electrophoresis

10XTAE buffer (pH 8.0): 0.4M tris-acetate, 10mM EDTA.

10XTBE buffer: 1M trisma base, 0.83M boric acid, 10mM EDTA.

10X loading buffers- agarose gels: 25% (w/v) ficoll, 0.25% (w/v) orange G, 10mM EDTA.

polyacrylamide gels: 95% formamide, 0.3% bromophenol blue, 0.3% xylene cyanol, 10mM EDTA.

Ethidium bromide (10mg/ml): 1g ethidium bromide/100ml SDW.

2.9.4 Southern blot reagents

Depurinatoir: 0.25M HCl

Denaturer: 0.5M NaOH, 1.5M NaCl.

0.4M NaOH: prepared with SDW.

Hybridisation mix: 0.3M NaPi (pH 7.2; 23.22ml 1M Na2HPO4 plus 6.78ml 1M NaH2PO4 per litre of sterile distilled water) 1mM EDTA, 7% SDS, 1%

Bovine serum albumin (BSA, fraction V; Sigma)
2.9.5 **Transformation of cloned DNA using competent *E. coli* host cells**

**0.1M Calcium chloride**: prepared with sterile distilled water (SDW) and autoclaved before use.

**SOC media**: To 950ml of distilled water added: 20g bacto-tryptone, 5g bacto-yeast extract, 0.5g NaCl, 10ml of 250mM solution of KCl and pH adjusted to 7.0 with 5N NaOH (~0.2ml). The volume of the solution was then adjusted to 1 litre with distilled water and autoclaved. The solution was allowed to cool to 60°C or less and then 20ml of a sterile 1M solution of glucose (filter sterilised) was added.

**10% maltose**: prepared with SDW and filter sterilised.

**10mM MgSO$_4$H$_2$O**: prepared with SDW and autoclaved before use.

**SM buffer**: 5.8g NaCl, 2g MgSO$_4$.7H$_2$O, 50ml 1M Tris.Cl (pH 7.5), 2% gelatin solution and made up to 1 litre with distilled water. The buffer was autoclaved before use.
CHAPTER THREE

Characterisation of XLRP pedigrees using
haplotype analysis
CHAPTER THREE

CHARACTERISATION OF XLRP PEDIGREES USING HAPLOTYPE ANALYSIS

3.1 Introduction

3.1.1 Clinical and pathological characteristics of XLRP

Retinitis pigmentosa (RP) is a set of retinal diseases in which photoreceptor cells degenerate and is one of the most common forms of hereditary blindness. Clinically it is characterised by night blindness, a gradual loss of peripheral visual field, and eventually loss of central vision in most cases. Fundus examination late in the disease typically shows a pale optic nerve head, attenuated retinal vessels, and typical "bone-spicule" pigmentedary deposits within the outer retina (Bird, 1975). Atrophy of the choriocapillaris is generally a late sign, in contrast to choroideremia which is characterised by progressive dystrophy of the choroid, retinal pigment epithelium and retina from early adulthood (section 1.8.6; Heckenlively and Bird, 1988). A progressive deterioration of retinal function is revealed by electroretinographic analysis, in which the electrical signals generated by the retina in response to flashes of light are reduced in amplitude and delayed in response times (Berson et al., 1969). In typical cases rod photoreceptors are more severely affected early in the disease, hence the symptom of night blindness. Later in the disease cone photoreceptors also degenerate.

The primary effects are evident at the level of photoreceptor and/or retinal pigment epithelial cell function although the precise cellular locus of abnormality is poorly understood. The disease is genetically heterogeneous and can be inherited as an autosomal dominant, autosomal recessive, X-linked dominant and X-linked recessive trait.

Unlike other forms of RP the progress and appearance of the X-linked (XLRP) form of the disease is usually different in males and females; males being more severely affected and the expression of the disease being variable in females. In males the onset of the disease is in the first decade with loss of dark adaptation, reduction of visual fields in the second decade, loss of visual acuity by 20 years of age and severe visual loss by 40. Female carriers show a variable but generally mild degree of visual loss, which occurs much later than in males and is associated with characteristically mild ophthalmological changes in the fundus and abnormalities on visual field and electrophysiological testing.
3.1.2 Clinical heterogeneity of XLRP

Clinical differences have been reported between families in which the RP2 and RP3 forms of XLRP (section 1.8.7) are segregating, which is a less compelling line of evidence for the existence of these two loci. There is a small body of evidence that suggests the possibility that RP2 is a true rod-cone dystrophy, with the major functional deficit in the rod system (Wright, 1990) while in RP3 there may be a more equal involvement of rod and cone photoreceptors, perhaps implying that the primary defect lies in another cell type such as the retinal pigment epithelium or Muller cells (Berson et al., 1980; Heckenlively, 1988). Although there have been many studies of the phenotype of XLRP hemizygotes and heterozygotes with unknown genotypes (Keith et al., 1991; Kaplan et al., 1992; Musarella et al., 1989), the phenotypic expression of the retinal disease in XLRP patients known to have the RP2 and RP3 genotype is not well understood.

In one report, an XLRP family linked to the RP3 locus showed diverse fundus abnormalities in the hemizygotes (Keith et al., 1991). In another study Kaplan et al., (1992) reported that RP2 and RP3 could be distinguished clinically in 9 French families. Patients segregating the RP2 form of XLRP were reported to have early onset night blindness and severe myopia whereas RP3 patients have later onset night blindness and little if any myopia. Other groups believe that the wide clinical variation within X-linked RP makes genetic locus identification on the basis of clinical features problematic (Wright et al., 1991). Several authors suggested that XLRP with tapetal reflex (described in detail in section 3.1.3), in one or more carriers was a distinct genetic form mapping to the distal (RP3) location (Nussbaum et al., 1985; Denton et al., 1988; Curtis and Blank, 1989; Musarella et al., 1989), although one out of 14 carrier females in a family segregating the proximal locus (RP2) has been reported to show a tapetal reflex (Friedrich et al., 1985). Musarella et al., (1988) failed to establish any relationship between the genotype and the retinal phenotype of carriers, and electroretinogram studies were not found to be discriminative either (Arden et al., 1983, Berson et al., 1979). In a recent study Flaxell, et al., (1997) found no clear phenotypic differences between the RP2 and RP3 forms in 17 British XLRP families with respect to myopia, onset of night blindness and the presence of primary cone dysfunction. The tapetal reflex was found to exist in carriers of both RP2 and RP3. Hence they concluded that there is a continuum incorporating phenotypic features of both forms of disease with inter-familial variability prevalent leading to the wide range of clinical presentations observed.

Heterogeneity can be detected with or without classification of families into two or more groups on clinical grounds. If the clinical demarcation corresponds closely to the genetic one, then the analysis is very powerful. However if there are no such reliable clinical distinction as presently seems to be the case for XLRP, then establishing
heterogeneity depends on factors such as the number of available families, the relative proportion of each type and the genetic distance between the two loci assuming only two loci exist in this interval.

### 3.1.3 Obligate carriers and variable expressivity

The clinical characteristics of carrier females vary from a complete absence of symptoms to severe abnormality similar to hemizygotes (Bird, 1975). Most if not all adult heterozygous females have detectable degenerative changes in the ocular fundus. The ocular changes in heterozygous females are most easily detected by fundus examination, visual field testing and dark adaptation measurements (Fishman et al., 1986). The single most frequent abnormality is peripheral retinal pigment epithelial atrophy, which is found in all adult heterozygous females (Bird, 1975).

The great majority of carriers show a highly attenuated form of retinal degeneration, with late onset of symptoms, mild visual loss and minimal clinical signs. The latter consist of varying degrees of pigmented degeneration, thinning of the pigment epithelial cell layer and in some individuals, an abnormal fundus reflex, referred to as a "tapetal reflex" due to its resemblance to the iridescent choroidal layer (tapetum lucidum) present in some mammals (Falls and Cotterman, 1948; Fishman et al., 1986). This is a golden metallic sheen which is occasionally also seen in young hemizygous males (Heckenlively, 1988). However, its detection can be subjective and the reflex diminishes with age, so it is not always a helpful indicator of carrier status (Bird, 1975). Carriers may have clinical signs at any age but tend not to become symptomatic until middle to late life (Bird, 1975). Severely affected carriers may be night blind, have reduced visual field and sometimes loss of visual acuity. Like affected males, myopia and astigmatism are common in heterozygous females and carriers may show myopic degeneration of the fundus (Fishman et al., 1986).

The great variability in functional loss from one heterozygote to another has been claimed to support Lyon's hypothesis (Lyon, 1961). This hypothesis explains the existence of carriers with mild as well as very severe disease without having to invoke specific genetic mechanisms such as X-linked intermediate or dominant inheritance. In females who are heterozygous for XLRP, random inactivation of one or other of the two X chromosomes in each cell during early embryogenesis (Lyon, 1961) causes the phenotype to range from asymptomatic to severely affected, depending on the proportion of retinal cells with an active disease-bearing X-chromosome (Bird, 1975; Fishman et al., 1986). It is likely that the expression of the heterozygous state is governed by many factors, including the influences of the normal allele and chromosome inactivation, ambient illumination of the retina, and anatomical and physiological factors affecting the retina.
3. Characterisation of XLRP pedigrees using haplotype analysis

About 10-20% of obligate carriers show a normal or near-normal phenotype when assessed in specialist centres (Fishman et al., 1986; Friedrich et al., 1985). This has important implications for carrier detection in females who are at 50% risk of being heterozygous for XLRP. It cannot be assumed that those who have no detectable carrier phenotype do not carry the gene, and while retinal changes may become apparent in later life, the question of genetic status is most important to those of child-bearing age.

3.1.4 Definition and the use of haplotype analysis in XLRP pedigrees

Once linkage is established and heterogeneity (section 1.8.7.1) has been addressed, it is important to classify XLRP families as segregating the RP2, RP3 or other loci (section 1.8.7) so that best estimate locations for the genes can be determined. Fine mapping of the region of interest is then carried out to narrow the critical region in order to find the specific disease-causative gene that is causing the disease by positional cloning. Haplotype analysis (section 1.3.6) was used in this study in an attempt to classify the XLRP families and in particular to finely map the RP2 locus.

Haplotype analysis using microsatellites from the interval of interest (as described in section 1.3.6) was carried out on each family. Genetic mapping of twelve new XLRP families and one choroideremia family was carried out using highly polymorphic microsatellite markers and new markers isolated and characterised in our lab (Thiselton et al., 1995) in conjunction with markers developed by Weissenbach et al., (1992) and NIH/CEPH consortium (1993). Twenty eight microsatellite markers covering the Xp22.13-Xp11.22 region were used to generate haplotypes for all sampled individuals. Haplotypes were constructed assuming the minimal number of recombination events. The aim was firstly to identify families as RP2 or RP3 and also using informative critical crossovers between markers and the RP2 critical region to further refine the interval containing the RP2 gene.

3.2 Methodology

3.2.1 Subjects and Samples

Twelve XLRP families (nine British, two Italian and one American) and one choroideremia family were analysed. The choroideremia family was of Spanish origin. Members of the British XLRP families who agreed to participate in the study were examined by clinicians at Moorfields Eye Hospital in London. The Italian, Spanish and American patients were examined in their respective countries. All affected individuals
were examined and their disease status was confirmed. A diagnosis of XLRP was based on detailed family history and comprehensive ophthalmological tests including fundus examination, visual field assessment and fluorescein angiography, in order to determine fundus morphology and electroretinography (ERG) measurements. Blood samples were collected from members of the British families including affected individuals, their normal spouses and normal first degree relatives. DNA was extracted as described in section 2.1.1. DNA samples from the Italian, Spanish and American families were sent to this laboratory from clinical collaborators.

3.2.2 Polymorphic markers on Xp used in this study

The Figure 3.1 shows the genetic order of the 28 markers from Xp22.13 to Xp11.22 used in this study. These markers which cover a region of about 20 cM (Nelson et al., 1995), extending distal and proximal to the RP2 and RP3 genetic intervals were used to classify the twelve XLRP families using haplotype analysis (section 3.1.4). Marker DXS989 (Xp22.13) is located within the RP15 locus and the marker DXS1214 (Xp21.2) is located at the RP6 proximal boundary distal to RP3. The markers DXS1242, DXS572, CYBB (Xp21.1) are also distal to the RP3 interval. Markers DXS1110 and DXS6679 (Xp21.1) lie within the RP3 critical interval determined at the onset of this study (5th X-chromosome Report; Willard et al., 1994) and DXS1068, DXS556, DXS574 DXS977, and DXS228 (Xp11.4) are located just proximal to RP3 and distal to the RP2 critical regions. Microsatellite markers DXS7, MAOA, DXS8080, DXS8026, DXS1055, DXS1003, DXS1146, SYN1, DXS426, DXS722, DXS1126, DXS573 and DXS1039 (Xp11.3-Xp11.23) are all located within the RP2 critical interval and were used to generate extensive haplotypes as a prerequisite for refining this region using critical crossover events. Markers DXS1000, DXS988 and DXS1204 (Xp11.22) are located just proximal to the RP2 critical region.

Table 3.1 gives a summary of these markers with their respective chromosomal location, polymerase chain reaction annealing temperatures, allele size ranges and heterozygosity values. Online information regarding these marker loci were obtained using the Genome Data Base (GDB) and/or from relevant literature.

3.2.3 Detection of Microsatellite Polymorphisms

The forward primer for each microsatellite marker was end-labelled with $^{32}$P-$\gamma$ATP (section 2.7.1) and the dinucleotide repeats were then amplified from genomic DNA (section 2.7.1). Alleles were detected by electrophoresing the PCR products on 6% denaturing polyacrylamide gels, followed by exposure to X-ray film (section 2.7.2).
Figure 3.1 Schematic diagram showing the order and approximate location and distance (cM) of the 28 genetic markers from Xp22.13 to Xp11.22 used in the haplotype analysis.

<table>
<thead>
<tr>
<th>Chromosome Site</th>
<th>Marker</th>
<th>Reference 1</th>
<th>Reference 2</th>
<th>Reference 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xp 22.13</td>
<td>DXS989</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xp 22.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xp 22.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xp 21.3</td>
<td>DXS669</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xp 21.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xp 21.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xp 11.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xp 11.3</td>
<td>MAOA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xp 11.23</td>
<td>SYN1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xp 11.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### TABLE 3.1 Microsatellite markers used for genetic characterisation of XLRP families

<table>
<thead>
<tr>
<th>Marker</th>
<th>Chromosomal location</th>
<th>Annealing Temperature</th>
<th>Allele size range (bp)</th>
<th>Heterozygosity</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>DXS989</td>
<td>Xp22.13</td>
<td>55°C</td>
<td>173-199</td>
<td>0.82</td>
<td>Weissenbach et al., 1992</td>
</tr>
<tr>
<td>DXS1214</td>
<td>Xp21.3-21.2</td>
<td>55°C</td>
<td>212-220</td>
<td>0.79</td>
<td>Gyapay et al., 1994</td>
</tr>
<tr>
<td>DXS1242</td>
<td>Xp21.1</td>
<td>55°C</td>
<td>214-228</td>
<td>0.77</td>
<td>Feener et al., 1991</td>
</tr>
<tr>
<td>DXS572</td>
<td>Xp21.1</td>
<td>56°C</td>
<td>156-164</td>
<td>0.29</td>
<td>Lindsay et al., 1991</td>
</tr>
<tr>
<td>CYBB (JL152)</td>
<td>Xp21.1</td>
<td>62°C</td>
<td>146-160</td>
<td>0.76</td>
<td>Hardwick et al., 1993</td>
</tr>
<tr>
<td>DXS1110 (CGD)</td>
<td>Xp21.1</td>
<td>56°C</td>
<td>252-268</td>
<td>0.68</td>
<td>Roux et al., 1993</td>
</tr>
<tr>
<td>DXS6679</td>
<td>Xp21.1</td>
<td>55°C</td>
<td>124-130</td>
<td>0.64</td>
<td>Meitinger, T, 1994</td>
</tr>
<tr>
<td>DXS1068</td>
<td>Xp11.4</td>
<td>56°C</td>
<td>245-259</td>
<td>0.82</td>
<td>Weissenbach et al., 1992</td>
</tr>
<tr>
<td>DXS556</td>
<td>Xp11.4</td>
<td>57°C</td>
<td>176-192</td>
<td>0.76</td>
<td>Thiselton et al., 1993</td>
</tr>
<tr>
<td>DXS574</td>
<td>Xp11.4</td>
<td>50°C</td>
<td>204-212</td>
<td>0.61</td>
<td>Lindsay et al., 1991</td>
</tr>
<tr>
<td>DXS977</td>
<td>Xp11.4</td>
<td>50°C</td>
<td>148-162</td>
<td>0.52</td>
<td>Kamakari et al., 1995</td>
</tr>
<tr>
<td>DXS228 (IaA6)</td>
<td>Xp11.4</td>
<td>56°C</td>
<td>140-150</td>
<td>0.53</td>
<td>Coleman et al., 1991</td>
</tr>
<tr>
<td>DXS7 (L1.28)</td>
<td>Xp11.3</td>
<td>62°C</td>
<td>157-162</td>
<td>0.49</td>
<td>Moore et al., 1992</td>
</tr>
<tr>
<td>MAOA</td>
<td>Xp11.3</td>
<td>56°C</td>
<td>112-126</td>
<td>0.72</td>
<td>Black et al., 1991</td>
</tr>
<tr>
<td>DXS8080</td>
<td>Xp11.3</td>
<td>58°C</td>
<td>177-191</td>
<td>0.68</td>
<td>Dib et al., 1996</td>
</tr>
<tr>
<td>DXS8026</td>
<td>Xp11.3</td>
<td>58°C</td>
<td>222-226</td>
<td>0.60</td>
<td>Dib et al., 1996</td>
</tr>
<tr>
<td>DXS1055</td>
<td>Xp11.3-p11.23</td>
<td>58°C</td>
<td>81-93</td>
<td>0.72</td>
<td>Gyapay et al., 1994</td>
</tr>
<tr>
<td>DXS1003</td>
<td>Xp11.3-p11.23</td>
<td>57°C</td>
<td>169-195</td>
<td>0.80</td>
<td>Gyapay et al., 1994</td>
</tr>
<tr>
<td>DXS1146</td>
<td>Xp11.23</td>
<td>56°C</td>
<td>130-142</td>
<td>0.45</td>
<td>Hong et al., 1993</td>
</tr>
<tr>
<td>SYN1</td>
<td>Xp11.23</td>
<td>60°C</td>
<td>240-260</td>
<td>0.84</td>
<td>Kirchgessner et al., 1991</td>
</tr>
<tr>
<td>DXS426</td>
<td>Xp11.23</td>
<td>52°C</td>
<td>290-310</td>
<td>0.52</td>
<td>Luty et al., 1990</td>
</tr>
<tr>
<td>DXS722</td>
<td>Xp11.23</td>
<td>57°C</td>
<td>188-192</td>
<td>0.47</td>
<td>Lindsay et al., 1991</td>
</tr>
<tr>
<td>DXS1126</td>
<td>Xp11.23</td>
<td>60°C</td>
<td>230-252</td>
<td>0.68</td>
<td>Donnelly et al., 1994</td>
</tr>
<tr>
<td>DXS573</td>
<td>Xp11.23</td>
<td>56°C</td>
<td>137-145</td>
<td>0.72</td>
<td>Lindsay et al., 1991</td>
</tr>
<tr>
<td>DXS1039</td>
<td>Xp11.23-p11.22</td>
<td>56°C</td>
<td>93-103</td>
<td>0.61</td>
<td>Gyapay et al., 1994</td>
</tr>
<tr>
<td>DXS1000</td>
<td>Xp11.22</td>
<td>55°C</td>
<td>230-236</td>
<td>0.35</td>
<td>Gyapay et al., 1994</td>
</tr>
<tr>
<td>DXS988</td>
<td>Xp11.22</td>
<td>55°C</td>
<td>134-144</td>
<td>0.62</td>
<td>Gyapay et al., 1994</td>
</tr>
<tr>
<td>DXS1204</td>
<td>Xp11.22</td>
<td>65°C</td>
<td>243-247</td>
<td>0.81</td>
<td>Gyapay et al., 1994</td>
</tr>
</tbody>
</table>
3.3 Results of clinical and haplotype analysis

3.3.1 Family RP85

British family XLRP 85 (Figure 3.2) comprises 4 generations with DNA obtained from 2 affected males, 2 unaffected males, 2 obligate carrier females, 1 clinically diagnosed carrier heterozygote and one female of unknown carrier status.

3.3.1.1 Clinical diagnosis of affected males and carrier females

The two affected males did not show early (i.e. before 10 years of age) onset of night blindness and had low myopia. There was also evidence of cone dystrophy in the affected males. Tapetal reflex was observed in some of the carrier females studied, (Flaxell et al., 1997).

3.3.1.2 Haplotype analysis of XLRP 85

In family RP85 (Figure 3.2) 18 markers were used for haplotype analysis. Affected males III-3 and III-4, unaffected males III-2 and IV-1 and obligate carrier females II-1 and II-2 did not show any cross-over events throughout the 15 cM region studied. The phase of the paternal alleles was established from analysis of 2 obligate carrier sibs, (II-1 & II-2) who share their affected father's disease-associated haplotype. To facilitate convenience of haplotype reading the pedigree XLRP85 (Figure 3.2) only shows 8 of the markers studied and the Table 3.2 gives all the 18 markers studied.

Since there were no recombination events within the RP2/RP3 critical regions on proximal Xp, the RP85 family could not be defined as either RP2 or RP3. The distal and proximal boundary of the disease could not be defined with the markers used and the possibility of disease in this family segregating with the RP6 or RP15 loci cannot be excluded.
3. Characterisation of XLRP pedigrees using haplotype analysis

Family XLRP 85

I

II

III

IV

DXS572 1 1 1 1
DXS1109 1 1 1 1
DXS877 2 2 1 1
DXS7 3 2 1 1
MAOA 5 3 2 2
DXS426 4 3 4 4
DXS722 2 1 2 2
DXS977 1 1 1 1

DXS572 1 1 1 1
DXS1109 1 1 1 1
DXS877 2 2 1 1
DXS7 3 2 1 1
MAOA 5 3 2 2
DXS426 4 3 4 4
DXS722 2 1 2 2
DXS977 1 1 1 1

DXS572 1 1 1 1
DXS1109 1 1 1 1
DXS877 2 2 1 1
DXS7 3 2 1 1
MAOA 5 3 2 2
DXS426 4 3 4 4
DXS722 2 1 2 2
DXS977 1 1 1 1

FIGURE 3.2
Pedigree of XLRP 85 showing haplotypes constructed with the markers listed in order from distal to proximal.

This family could not be defined as either RP2 or RP3 due to lack of recombinants.
The possibility of it being a RP6 or RP15 family cannot be excluded.
The boxed genotype denotes alleles segregating with the disease associated haplotype and
the unboxed genotype alleles segregating with the non-disease haplotype.

□ Normal males and females
■ Affected males
○ Obligate carrier females
♀ Clinically described as
heterozygote carrier female
♀ Female of unknown carrier status
Crossed symbols represent deceased family members

Non-paternity

159
TABLE 3.2  Haplotype data for 18 microsatellite markers used in the study of XLRP 85.

<table>
<thead>
<tr>
<th>Individual of Generation</th>
<th>II 1</th>
<th>II 1a</th>
<th>II 2</th>
<th>II 2a</th>
<th>III 1</th>
<th>III 1a</th>
<th>III 2</th>
<th>III 3</th>
<th>III 4</th>
<th>IV 1</th>
<th>IV 2*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenotype</td>
<td>Carrier</td>
<td>Normal</td>
<td>Carrier</td>
<td>Normal</td>
<td>Carrier</td>
<td>Normal</td>
<td>Normal</td>
<td>Affected</td>
<td>Affected</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>Microsatellite Marker</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
</tr>
<tr>
<td>DXS572</td>
<td>1,1</td>
<td>1</td>
<td>1,1</td>
<td>1</td>
<td>1,1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1,1</td>
</tr>
<tr>
<td>DXS1110</td>
<td>1,1</td>
<td>1</td>
<td>1,1</td>
<td>2</td>
<td>1,1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1,1</td>
</tr>
<tr>
<td>DXS5556</td>
<td>2,2</td>
<td>2</td>
<td>2,3</td>
<td>4</td>
<td>2,2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2,2</td>
</tr>
<tr>
<td>DXS574</td>
<td>2,2</td>
<td>2</td>
<td>2,1</td>
<td>2</td>
<td>2,2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1,2</td>
</tr>
<tr>
<td>DXS977</td>
<td>2,2</td>
<td>1</td>
<td>2,2</td>
<td>3</td>
<td>2,1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2,1</td>
</tr>
<tr>
<td>DXS228</td>
<td>1,3</td>
<td>3</td>
<td>1,2</td>
<td>4</td>
<td>1,3</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3,3</td>
</tr>
<tr>
<td>DXS57</td>
<td>3,2</td>
<td>1</td>
<td>3,1</td>
<td>2</td>
<td>3,1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3,1</td>
</tr>
<tr>
<td>MAOA</td>
<td>5,3</td>
<td>2</td>
<td>5,1</td>
<td>3</td>
<td>5,2</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>4,2</td>
</tr>
<tr>
<td>DXS1055</td>
<td>3,1</td>
<td>1</td>
<td>3,1</td>
<td>2</td>
<td>3,1</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1,1</td>
</tr>
<tr>
<td>DXS1003</td>
<td>2,4</td>
<td>4</td>
<td>2,4</td>
<td>3</td>
<td>2,4</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4,4</td>
</tr>
<tr>
<td>SYN1</td>
<td>1,3</td>
<td>1</td>
<td>1,3</td>
<td>3</td>
<td>1,1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>3,1</td>
</tr>
<tr>
<td>DXS426</td>
<td>4,3</td>
<td>4</td>
<td>4,3</td>
<td>2</td>
<td>4,4</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4,4</td>
</tr>
<tr>
<td>DXS722</td>
<td>2,1</td>
<td>2</td>
<td>2,1</td>
<td>2</td>
<td>2,2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2,2</td>
</tr>
<tr>
<td>DXS1126</td>
<td>3,1</td>
<td>3</td>
<td>3,1</td>
<td>2</td>
<td>3,2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4,3</td>
</tr>
<tr>
<td>DXS573</td>
<td>1,1</td>
<td>3</td>
<td>1,1</td>
<td>1</td>
<td>1,3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1,3</td>
</tr>
<tr>
<td>DXS1039</td>
<td>2,1</td>
<td>3</td>
<td>2,1</td>
<td>2</td>
<td>2,3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3,3</td>
</tr>
<tr>
<td>DXS988</td>
<td>3,3</td>
<td>2</td>
<td>3,3</td>
<td>4</td>
<td>3,2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1,2</td>
</tr>
<tr>
<td>DXS1204</td>
<td>1,2</td>
<td>1</td>
<td>1,2</td>
<td>3</td>
<td>1,1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3,1</td>
</tr>
</tbody>
</table>

*Alleles segregating on one of the X-chromosomes for individual IV-2 shows non-paternity
3.3.2 Family RP87

British family XLRP87 (Figure 3.3) comprises 3 generations with DNA obtained for 3 obligate carriers, 2 affected males, 2 unaffected males and one female of unknown carrier status.

3.3.2.1 Clinical diagnosis of affected males and carrier females

The affected males in this family did not show early (before 10 years of age) onset of night blindness and did not have any form of myopia. There was no evidence of cone dystrophy in the affected males. Tapetal reflex was not observed in the carrier females. Individual VI-7 has a normal fundus though she is an obligate carrier female with an affected son VII-2. This shows a lack of phenotypic expression for the XLRP disease in this obligate carrier. It cannot be assumed therefore that those who have no detectable carrier phenotype do not carry the disease-associated mutation in the gene.

3.3.2.2 Haplotype analysis of XLRP 87

In family RP87 (Figure 3.3 & Table 3.3) 18 markers were used for haplotype analysis. Individual VII-2, an affected male is recombinant with respect to his obligate carrier mother VI-7 below the marker SYN1, as he has inherited the disease bearing chromosome intact from DXS572 (Xp21.1) to SYN1 (Xp11.23). He could therefore be affected due to either the RP2 or RP3 disease genes. His half sister VII-1 has a cross-over below the marker DXS572 from the normal haplotype to the affected haplotype and below MAOA (hatched box denoting uninformative markers DXS7 and MAOA) to the normal haplotype again. However, because of her unknown clinical status this information does not at present contribute to disease gene localisation in this family. Non-paternity is also detected in this individual since the haplotype data on her paternal X chromosome is not the same as VI-7a.

The phase of the maternal alleles for both VII-1 and VII-2 in VI-7 is firmly established from analysis of 2 obligate carriers (V-1 & V-2) and her affected male cousin (VI-2) who share the mother's disease-associated haplotype. Therefore since the disease-associated haplotype extends from DXS572 (Xp21.1) to SYN1 (Xp11.23) in this family, it cannot be classified as either RP3 or RP2. Since markers distal to DXS572 have not been studied the possibility of disease segregating with the RP6 or RP15 loci cannot be excluded in this family. The pedigree of XLRP87 (Figure 3.3) presents data for 10 of the markers analysed but Table 3.3 presents the haplotype data for all the 18 markers studied.
Family XLRP 87

V

DXS872 1 1 1
DXS1110 1 1 1
DXS228 1 2 2
DXS7 2 1 2
MAOA 3 1 1
DXS1003 2 4 2
SYN1 3 2 1
DXS426 4 2 3
DXS1126 1 3 2
DXS573 2 4 3

VI

DXS872 1 2 2
DXS1110 1 1 1
DXS228 1 2 2
DXS7 3 1 1
MAOA 3 1 1
DXS1003 2 4 4
SYN1 3 2 2
DXS426 4 2 2
DXS1126 1 3 3
DXS573 2 4 4

VII

Pedigree of XLRP 87 showing haplotypes constructed with the markers listed. This family could not be defined as either RP2 or RP3 due to lack of recombinants.

The possibility of disease segregating with the RP6 or RP15 loci cannot be excluded.

The boxed genotype indicates alleles segregating with the disease associated haplotype and the unboxed genotype shows alleles segregating with the non-disease haplotype. The genotype in the hatched box of individual VII-1 indicates uninformative markers DXS7 and MAOA.

Non-paternity

□ Normal males and females

■ Affected males

● Obligate carrier females

♀ Female of unknown carrier status

FIGURE 3.3
TABLE 3.3 Haplotype data for 18 microsatellite markers used in the study of XLRP 87.

<table>
<thead>
<tr>
<th>Individual of Generation</th>
<th>V 1 Carrier female</th>
<th>V 1a Normal male</th>
<th>V 2 Carrier female</th>
<th>V 2a Normal male</th>
<th>VI 2 Affected male</th>
<th>VI 3 Normal male</th>
<th>VI 4 Normal male</th>
<th>VI 7 Carrier female</th>
<th>VI 7a Normal male</th>
<th>VII 1 Carrier female</th>
<th>VII 2 Normal male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenotype</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Microsatellite Marker</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
</tr>
<tr>
<td>DXS572</td>
<td>1,2</td>
<td>2</td>
<td>1,2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1,2</td>
<td>2</td>
<td>2</td>
<td>2,2</td>
<td>1</td>
</tr>
<tr>
<td>DXS1110</td>
<td>1,1</td>
<td>1</td>
<td>1,1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1,2</td>
<td>2</td>
<td>1,1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>DXS556</td>
<td>3,2</td>
<td>1</td>
<td>3,1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3,3</td>
<td>3</td>
<td>3,2</td>
<td>3</td>
</tr>
<tr>
<td>DXS574</td>
<td>1,2</td>
<td>2</td>
<td>1,2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1,1</td>
<td>1</td>
<td>1,2</td>
<td>1</td>
</tr>
<tr>
<td>DXS977</td>
<td>3,3</td>
<td>1</td>
<td>3,2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3,3</td>
<td>3</td>
<td>3,2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>DXS228</td>
<td>1,2</td>
<td>3</td>
<td>1,4</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1,4</td>
<td>4</td>
<td>1,4</td>
<td>1</td>
</tr>
<tr>
<td>DXS57</td>
<td>3,1</td>
<td>2</td>
<td>3,5</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3,3</td>
<td>3</td>
<td>3,3</td>
<td>3</td>
</tr>
<tr>
<td>MAOA</td>
<td>3,1</td>
<td>3</td>
<td>3,2</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3,3</td>
<td>3</td>
<td>3,2</td>
<td>3</td>
</tr>
<tr>
<td>DXS1055</td>
<td>2,2</td>
<td>3</td>
<td>2,2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2,1</td>
<td>1</td>
<td>1,3</td>
<td>2</td>
</tr>
<tr>
<td>DXS1003</td>
<td>2,4</td>
<td>2</td>
<td>2,3</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>2,1</td>
<td>1</td>
<td>1,2</td>
<td>2</td>
</tr>
<tr>
<td>SYN1</td>
<td>3,2</td>
<td>1</td>
<td>3,2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3,2</td>
<td>2</td>
<td>2,1</td>
<td>3</td>
</tr>
<tr>
<td>DXS426</td>
<td>4,2</td>
<td>3</td>
<td>4,4</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>4,1</td>
<td>1</td>
<td>1,1</td>
<td>1</td>
</tr>
<tr>
<td>DXS722</td>
<td>2,2</td>
<td>2</td>
<td>2,2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2,2</td>
<td>2</td>
<td>2,1</td>
<td>2</td>
</tr>
<tr>
<td>DXS1126</td>
<td>1,3</td>
<td>2</td>
<td>1,1</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1,3</td>
<td>3</td>
<td>3,3</td>
<td>3</td>
</tr>
<tr>
<td>DXS573</td>
<td>2,4</td>
<td>3</td>
<td>2,1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>2,1</td>
<td>1</td>
<td>1,4</td>
<td>1</td>
</tr>
<tr>
<td>DXS1039</td>
<td>3,2</td>
<td>1</td>
<td>3,3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3,3</td>
<td>3</td>
<td>3,3</td>
<td>3</td>
</tr>
<tr>
<td>DXS988</td>
<td>3,1</td>
<td>2</td>
<td>3,2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3,2</td>
<td>2</td>
<td>2,2</td>
<td>2</td>
</tr>
<tr>
<td>DXS1204</td>
<td>2,1</td>
<td>1</td>
<td>2,3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2,2</td>
<td>2</td>
<td>2,1</td>
<td>2</td>
</tr>
</tbody>
</table>
3. Characterisation of XLRP pedigrees using haplotype analysis

3.3.3 Family RP 88

XLRP 88 (Figure 3.4) is a small British family of two generations with DNA obtained for 2 affected males, 1 unaffected male, 1 obligate carrier female and 2 females of unknown carrier status.

3.3.3.1 Clinical diagnosis of affected males and carrier females

The two affected males have low myopia with early onset of night blindness before the age of 10 years. There was some evidence of cone dystrophy in the affected males. Tapetal reflex was not evident on fundus examination of obligate carrier female II-1.

3.3.3.2 Haplotype analysis of XLRP 88

Haplotype analysis was carried out using 17 markers (Figure 3.4 & Table 3.4). Affected males III-2 and III-5 have inherited the complete affected haplotype from their obligate carrier mother (II-1) hence segregating for both RP2 and RP3 disease associated haplotypes. Both clinical data and genetic data agree in these individuals except for their sister III-1, a female (30 years old) of unknown clinical status and genetically carrying the same affected haplotype alleles as her affected brother III-2. Since all of these individuals were non-recombinant for markers DXS572 to DXS988 (refer to Table 3.4), the RP 88 family could not be classified as either a RP2 or RP3 family. The possibility of disease segregating for the RP6 or RP15 loci cannot be excluded in this family since markers distal to DXS572 have not been studied. Table 3.4 presents haplotype data for all 17 markers used in this study, while Figure 3.4 presents data for 9 of these markers to facilitate convenience of haplotype readings.

3.3.4 Family RP 89

XLRP 89 (Figure 3.5) is a small British family of 3 generations with DNA available for only 2 affected males and 2 obligate carrier females.

3.3.4.1 Clinical diagnosis of affected males and carrier females

On clinical diagnosis the affected males were found to suffer from low myopia with early onset (before 10 years of age) of night blindness. There was also some evidence of cone dystrophy in the affected males. Tapetal reflex was evident on fundus examination of obligate carrier females.
3. Characterisation of XLRP pedigrees using haplotype analysis

**Family XLRP 88**

**FIGURE 3.4**

Pedigree of XLRP 88 showing haplotypes constructed with the 9 markers listed.

This family could not be defined as either RP2 or RP3 due to lack of recombinants.
The possibility that this is an RP6 or RP15 family cannot be excluded.
The boxed genotype shows alleles segregating with the disease associated haplotype and the unboxed genotype alleles segregating with the non-disease haplotype.

□ □ Normal males and females
■ Affected males
○ Obligate carrier female
○ Females of unknown carrier status
TABLE 3.4  Haplotype data for 17 microsatellite markers used in the study of XLRP 88.

<table>
<thead>
<tr>
<th>Microsatellite Marker</th>
<th>Alleles</th>
<th>Alleles</th>
<th>Alleles</th>
<th>Alleles</th>
<th>Alleles</th>
<th>Alleles</th>
<th>Alleles</th>
</tr>
</thead>
<tbody>
<tr>
<td>DXS572</td>
<td>1, 1</td>
<td>1</td>
<td>1, 1</td>
<td>1</td>
<td>1, 1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>DXS1110</td>
<td>2, 1</td>
<td>3</td>
<td>2, 3</td>
<td>2</td>
<td>1, 3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>DXS556</td>
<td>1, 2</td>
<td>3</td>
<td>1, 3</td>
<td>1</td>
<td>2</td>
<td>2, 3</td>
<td>1</td>
</tr>
<tr>
<td>DXS574</td>
<td>2, 2</td>
<td>1</td>
<td>2, 1</td>
<td>2</td>
<td>2</td>
<td>2, 1</td>
<td>2</td>
</tr>
<tr>
<td>DXS977</td>
<td>1, 2</td>
<td>1</td>
<td>1, 1</td>
<td>1</td>
<td>2</td>
<td>2, 1</td>
<td>1</td>
</tr>
<tr>
<td>DXS228</td>
<td>2, 2</td>
<td>3</td>
<td>2, 3</td>
<td>2</td>
<td>2</td>
<td>2, 3</td>
<td>2</td>
</tr>
<tr>
<td>DXS7</td>
<td>1, 2</td>
<td>1</td>
<td>1, 1</td>
<td>1</td>
<td>2</td>
<td>2, 1</td>
<td>1</td>
</tr>
<tr>
<td>MAOA</td>
<td>1, 3</td>
<td>2</td>
<td>1, 2</td>
<td>1</td>
<td>3</td>
<td>3, 2</td>
<td>1</td>
</tr>
<tr>
<td>DXS1003</td>
<td>2, 1</td>
<td>3</td>
<td>2, 3</td>
<td>2</td>
<td>1</td>
<td>1, 3</td>
<td>2</td>
</tr>
<tr>
<td>SYN1</td>
<td>1, 2</td>
<td>1</td>
<td>1, 1</td>
<td>1</td>
<td>2</td>
<td>2, 1</td>
<td>1</td>
</tr>
<tr>
<td>DXS426</td>
<td>1, 2</td>
<td>3</td>
<td>1, 3</td>
<td>1</td>
<td>2</td>
<td>2, 3</td>
<td>1</td>
</tr>
<tr>
<td>DXS722</td>
<td>2, 1</td>
<td>1</td>
<td>2, 1</td>
<td>2</td>
<td>1</td>
<td>1, 1</td>
<td>2</td>
</tr>
<tr>
<td>DXS1126</td>
<td>2, 2</td>
<td>1</td>
<td>2, 1</td>
<td>2</td>
<td>2</td>
<td>2, 1</td>
<td>2</td>
</tr>
<tr>
<td>DXS573</td>
<td>1, 1</td>
<td>1</td>
<td>1, 1</td>
<td>1</td>
<td>1</td>
<td>1, 1</td>
<td>1</td>
</tr>
<tr>
<td>DXS1039</td>
<td>2, 3</td>
<td>1</td>
<td>2, 1</td>
<td>2</td>
<td>3</td>
<td>3, 1</td>
<td>2</td>
</tr>
<tr>
<td>DXS988</td>
<td>2, 2</td>
<td>1</td>
<td>2, 1</td>
<td>2</td>
<td>2</td>
<td>1, 1</td>
<td>2</td>
</tr>
<tr>
<td>DXS1204</td>
<td>3, 1</td>
<td>2</td>
<td>3, 2</td>
<td>3</td>
<td>1</td>
<td>1, 2</td>
<td>3</td>
</tr>
</tbody>
</table>
3. Characterisation of XLRP pedigrees using haplotype analysis

3.3.4.2 Haplotype analysis of XLRP 89

Nineteen microsatellite markers were used for haplotype analysis in family RP 89 (Figure 3.5 & Table 3.5). The phase of the maternal alleles of V-2 was established from analysis of her affected mother IV-1 and her affected brother V-1. Obligate carrier female V-2 is recombinant with respect to her affected mother IV-1 between markers DXS574 and DXS977, such that she has inherited the disease associated haplotype proximal to DXS574. This excludes RP3 segregation and defines this family as RP2. The distal boundary of the XLRP2 interval could not be further refined from her crossover data.

Individual VI-3, an affected male, is also a recombinant with respect to his carrier mother V-2 proximal to the marker MAOA, hence showing segregation for the RP2 disease-associated interval. This crossover defines the distal boundary of the XLRP2 interval, with the disease associated haplotype located proximal to MAOA, providing firm evidence for RP2 segregating in family XLRP 89. In this family the proximal boundary could not be refined further due to lack of recombinational events at the proximal portion of the RP2 interval.

In summary, key recombination events in family RP 89 are consistent with RP2 segregation and indicate an RP2 critical region with a distal boundary at the MAOA (Xp11.3) locus. Figure 3.5 of pedigree RP 89 show haplotype data for 11 of the markers used in this study indicating the key cross-overs while Table 3.5 shows haplotype data for all 19 markers used.

3.3.5 Family RP 90

British family XLRP 90 (Figure 3.6) comprises 4 generations with DNA available for 2 affected males, 3 obligate carrier females and 2 clinically described carrier females.

3.3.5.1 Clinical diagnosis of affected males and carrier females

Affected males showed absence of myopia and early onset of night blindness before 10 years of age. There was also no evidence of cone dystrophy in these affected males. Tapetal reflex was not evident on fundus examination of carrier females.

3.3.5.2 Haplotype analysis of XLRP 90

Haplotype analysis was studied using 17 markers (Figure 3.6 & Table 3.6). Affected males (cousins), IV-2 and IV-4 have inherited the complete affected haplotype from their respective obligate carrier mothers III-1 and III-2, hence segregating for both RP2 and RP3 disease loci. Their sisters IV-3 and IV-5 (respectively) have also inherited
3. Characterisation of XLRP pedigrees using haplotype analysis

Family XLRP 89

IV

DXS572 1
DXS574 2
DXS977 2
DXS228 2
DXS7 2
MAOA 2
DXS8026 2
DXS1003 2
DXS426 1
DXS573 2

V

DXS572 1
DXS5110 1
DXS574 2
DXS977 2
DXS228 2
DXS7 2
MAOA 2
DXS8026 2
DXS1003 2
DXS426 1
DXS573 2

VI

DXS572 1
DXS5110 2
DXS574 1
DXS977 1
DXS228 1
DXS7 2
MAOA 2
DXS8026 2
DXS1003 1
DXS426 1
DXS573 2

FIGURE 3.5
Pedigree of XLRP 89 showing haplotypes constructed with the 11 markers listed.
Recombinant individuals VI-3 & V-2 define this family as segregating for the RP2 disease. Cross-over data in VI-3 also localises the RP2 gene to the region below MAOA.
The boxed genotype shows alleles segregating with the disease-associated haplotype and the unboxed genotype shows alleles segregating with the non-disease haplotype.

□ ○ Normal males and females
■ ○ Affected males
● ○ Obligate carrier females
### TABLE 3.5 Haplotype data for 19 microsatellites markers used to refine the XLRP2 interval in RP 89.

<table>
<thead>
<tr>
<th>Individual of Generation</th>
<th>IV 1</th>
<th>V 1</th>
<th>V 2</th>
<th>VI 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenotype</td>
<td>Carrier female</td>
<td>Affected male</td>
<td>Carrier female</td>
<td>Affected male</td>
</tr>
<tr>
<td>Microsatellite Marker</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
</tr>
<tr>
<td>DXS572</td>
<td>1, 1</td>
<td>1</td>
<td>1, 1</td>
<td>1</td>
</tr>
<tr>
<td>DXS1110</td>
<td>1, 1</td>
<td>1</td>
<td>1, 2</td>
<td>2</td>
</tr>
<tr>
<td>DXS556</td>
<td>1, 1</td>
<td>1</td>
<td>1, 1</td>
<td>1</td>
</tr>
<tr>
<td>DXS574</td>
<td>3, 2</td>
<td>3</td>
<td>2, 1</td>
<td>1</td>
</tr>
<tr>
<td>DXS977</td>
<td>1, 2</td>
<td>1</td>
<td>1, 1</td>
<td>1</td>
</tr>
<tr>
<td>DXS228</td>
<td>2, 1</td>
<td>2</td>
<td>2, 1</td>
<td>1</td>
</tr>
<tr>
<td>DXS7</td>
<td>1, 2</td>
<td>1</td>
<td>1, 2</td>
<td>2</td>
</tr>
<tr>
<td>MAOA</td>
<td>2, 1</td>
<td>2</td>
<td>2, 3</td>
<td>2</td>
</tr>
<tr>
<td>DXS1055</td>
<td>1, 1</td>
<td>1</td>
<td>1, 1</td>
<td>1</td>
</tr>
<tr>
<td>DXS1003</td>
<td>1, 2</td>
<td>1</td>
<td>1, 2</td>
<td>1</td>
</tr>
<tr>
<td>SYN1</td>
<td>2, 1</td>
<td>2</td>
<td>2, 1</td>
<td>2</td>
</tr>
<tr>
<td>DXS426</td>
<td>1, 1</td>
<td>1</td>
<td>1, 1</td>
<td>1</td>
</tr>
<tr>
<td>DXS722</td>
<td>1, 2</td>
<td>1</td>
<td>1, 2</td>
<td>1</td>
</tr>
<tr>
<td>DXS1126</td>
<td>2, 1</td>
<td>2</td>
<td>2, 1</td>
<td>2</td>
</tr>
<tr>
<td>DXS573</td>
<td>1, 2</td>
<td>1</td>
<td>1, 3</td>
<td>1</td>
</tr>
<tr>
<td>DXS1039</td>
<td>1, 1</td>
<td>1</td>
<td>1, 1</td>
<td>1</td>
</tr>
<tr>
<td>DXS988</td>
<td>2, 1</td>
<td>2</td>
<td>2, 2</td>
<td>2</td>
</tr>
<tr>
<td>DXS1204</td>
<td>1, 2</td>
<td>1</td>
<td>1, 3</td>
<td>1</td>
</tr>
</tbody>
</table>
3. Characterisation of XLRP pedigrees using haplotype analysis

**Family XLRP 90**

**FIGURE 3.6**
Pedigree of XLRP 90 shows haplotypes constructed with the 9 markers listed. This family could not be defined as either RP2 or RP3 due to lack of recombinants. The possibility of disease segregating with the RP6 or RP15 loci cannot be excluded. The boxed genotype shows alleles segregating with the disease associated haplotype and the unboxed genotype segregating with the non-disease haplotype.

- Normal males
- Affected males
- Obligate carrier females
- Clinically described as heterozygote carrier females

Crossed symbols represent deceased family members.
TABLE 3.6 Haplotype data for 17 microsatellite markers used in the study of XLRP 90.

<table>
<thead>
<tr>
<th>Microsatellite Marker</th>
<th>Alleles</th>
<th>Alleles</th>
<th>Alleles</th>
<th>Alleles</th>
<th>Alleles</th>
<th>Alleles</th>
<th>Alleles</th>
<th>Alleles</th>
<th>Alleles</th>
</tr>
</thead>
<tbody>
<tr>
<td>DXS572</td>
<td>1,1</td>
<td>1</td>
<td>1,1</td>
<td>2</td>
<td>1,1</td>
<td>1</td>
<td>1,2</td>
<td>1</td>
<td>1,1</td>
</tr>
<tr>
<td>CYBB</td>
<td>1,2</td>
<td>2</td>
<td>1,2</td>
<td>2</td>
<td>1,2</td>
<td>1</td>
<td>1,2</td>
<td>1</td>
<td>1,1</td>
</tr>
<tr>
<td>DXS1110</td>
<td>1,2</td>
<td>1</td>
<td>1,1</td>
<td>1</td>
<td>1,1</td>
<td>1</td>
<td>1,1</td>
<td>1</td>
<td>1,2</td>
</tr>
<tr>
<td>DXS556</td>
<td>2,2</td>
<td>1</td>
<td>2,1</td>
<td>2</td>
<td>2,1</td>
<td>2</td>
<td>2,2</td>
<td>2</td>
<td>2,3</td>
</tr>
<tr>
<td>DXS977</td>
<td>3,4</td>
<td>1</td>
<td>3,1</td>
<td>2</td>
<td>3,1</td>
<td>3</td>
<td>3,2</td>
<td>3</td>
<td>3,1</td>
</tr>
<tr>
<td>DXS7</td>
<td>2,1</td>
<td>2</td>
<td>2,2</td>
<td>2</td>
<td>2,2</td>
<td>2</td>
<td>2,2</td>
<td>2</td>
<td>2,1</td>
</tr>
<tr>
<td>MAOA</td>
<td>2,3</td>
<td>2</td>
<td>2,2</td>
<td>1</td>
<td>2,2</td>
<td>2</td>
<td>2,1</td>
<td>2</td>
<td>2,2</td>
</tr>
<tr>
<td>DXS1055</td>
<td>1,1</td>
<td>1</td>
<td>1,1</td>
<td>3</td>
<td>1,1</td>
<td>1</td>
<td>1,3</td>
<td>1</td>
<td>1,2</td>
</tr>
<tr>
<td>DXS1003</td>
<td>1,2</td>
<td>1</td>
<td>1,1</td>
<td>1</td>
<td>1,1</td>
<td>1</td>
<td>1,1</td>
<td>1</td>
<td>1,2</td>
</tr>
<tr>
<td>SYN1</td>
<td>2,4</td>
<td>4</td>
<td>2,4</td>
<td>1</td>
<td>1,1</td>
<td>1</td>
<td>1,1</td>
<td>1</td>
<td>1,2</td>
</tr>
<tr>
<td>DXS426</td>
<td>1,1</td>
<td>1</td>
<td>1,1</td>
<td>1</td>
<td>1,1</td>
<td>1</td>
<td>1,1</td>
<td>1</td>
<td>1,1</td>
</tr>
<tr>
<td>DXS722</td>
<td>2,2</td>
<td>1</td>
<td>2,1</td>
<td>2</td>
<td>2,1</td>
<td>2</td>
<td>2,2</td>
<td>2</td>
<td>2,2</td>
</tr>
<tr>
<td>DXS1126</td>
<td>2,1</td>
<td>2</td>
<td>2,2</td>
<td>1</td>
<td>2,2</td>
<td>2</td>
<td>2,1</td>
<td>2</td>
<td>2,1</td>
</tr>
<tr>
<td>DXS573</td>
<td>1,2</td>
<td>3</td>
<td>1,3</td>
<td>2</td>
<td>1,3</td>
<td>1</td>
<td>1,2</td>
<td>1</td>
<td>1,1</td>
</tr>
<tr>
<td>DXS1039</td>
<td>3,1</td>
<td>2</td>
<td>3,2</td>
<td>4</td>
<td>3,2</td>
<td>3</td>
<td>3,4</td>
<td>3</td>
<td>3,3</td>
</tr>
<tr>
<td>DXS988</td>
<td>3,1</td>
<td>2</td>
<td>3,2</td>
<td>3</td>
<td>3,2</td>
<td>3</td>
<td>3,3</td>
<td>3</td>
<td>3,2</td>
</tr>
<tr>
<td>DXS1204</td>
<td>1,1</td>
<td>2</td>
<td>1,2</td>
<td>1</td>
<td>1,2</td>
<td>1</td>
<td>1,1</td>
<td>1</td>
<td>1,1</td>
</tr>
</tbody>
</table>
the complete affected haplotype from their obligate carrier mothers III-1 and III-2. Their affected genotypes agree with the clinical data which indicated that they were carriers.

The phase of the maternal alleles of III-1 and III-2 was consistent with the haplotype of their phase-known obligate carrier mother II-2. Since all the individuals studied were non-recombinants from markers DXS572 to DXS988 (refer to Table 3.6), the RP 90 family could not be classified as either RP2 or a RP3 family. The possibility of disease in this family segregating with the RP6 or RP15 loci cannot be excluded since markers above DXS572, covering the map locations of these disease loci have not been tested. Table 3.6 presents haplotype data for all 17 markers used in this study, while Figure 3.6 presents data for 9 of these markers.

3.3.6 Family RP 91

XLRP 91 (Figure 3.7) is a British family which comprises 3 generations with DNA obtained for 4 affected males, 1 normal male, 2 obligate carrier females and two females of unknown carrier status.

3.3.6.1 Clinical diagnosis of affected males and carrier females

On clinical diagnosis affected males showed early onset (before 10 years of age) of night blindness with no myopia. There was also no evidence of cone dystrophy in the affected males. Tapetal reflex was observed on fundus examination of carrier females. Individual V-1 was found to be a clinically unaffected male (30 years old). Individuals V-3 and V-4 were clinically difficult to diagnose and hence are denoted as females with unknown carrier status.

3.3.6.2 Haplotype analysis of XLRP 91

In family RP 91 (Figure 3.7 & Table 3.7) 12 microsatellite markers were used in haplotype analysis. Individual V-1, a normal male is a recombinant with respect to his obligate carrier mother IV-1, below the marker MAOA. His sister V-2, a obligate carrier female (with an affected son- not shown) is also a recombinant with respect to her carrier mother IV-1, below the marker MAOA, segregating for the disease-associated haplotype distal to the marker DXS1003. The phase of the maternal alleles of IV-1 was confirmed from the haplotype data of her affected brothers IV-2, IV-3, IV-4 and her affected son V-5, all being non-recombinants for markers CYBB to DXS988 (refer to Table 3.7 for haplotype data on this marker).

Individual V-3, a non-recombinant, has the complete affected haplotype of her obligate carrier mother IV-1 but was clinically found to be difficult to diagnose
Family XLRP 91

III

IV

V

FIGURE 3.7
Pedigree of XLRP 91 showing haplotypes constructed with 9 of the markers listed.
This family could not be defined as either RP2 or RP3 due to lack of useful crossover data.
The possibility of disease segregating with the RP6 or RP15 loci cannot be excluded.
The boxed genotype shows alleles segregating with the disease associated haplotype and
the unboxed genotype shows alleles segregating with the non-disease haplotype.

□ = Normal males and females
■ = Affected males
○ = Obligate carrier females
♀ = Females of unknown carrier status
Crossed symbols represent deceased family members

173
Table 3.7: Haplotype data for 12 microsatellite markers used to establish critical crossover events in XLRP 91.

<table>
<thead>
<tr>
<th>Individual of Generation</th>
<th>IV 1a</th>
<th>IV 1</th>
<th>IV 2</th>
<th>IV 3</th>
<th>IV 4</th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenotype</td>
<td>Normal male</td>
<td>Carrier female</td>
<td>Affected male</td>
<td>Affected male</td>
<td>Affected male</td>
<td>Normal male</td>
<td>Carrier female</td>
<td>Normal male</td>
<td>Carrier female</td>
<td>Affected male</td>
</tr>
<tr>
<td>Microsatellite Marker</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
</tr>
<tr>
<td>CYBB</td>
<td>1,2</td>
<td>1,1</td>
<td>2,1</td>
<td>1,2</td>
<td>2,1</td>
<td>1,1</td>
<td>2,1</td>
<td>1,1</td>
<td>1,1</td>
<td>1,1</td>
</tr>
<tr>
<td>DXS1110</td>
<td>1,3</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>1,1</td>
<td>2,1</td>
</tr>
<tr>
<td>DXS556</td>
<td>1,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>1,1</td>
<td>2,1</td>
</tr>
<tr>
<td>DXS977</td>
<td>1,1</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>1,1</td>
<td>2,1</td>
</tr>
<tr>
<td>DXS228</td>
<td>1,1</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>1,1</td>
<td>2,1</td>
</tr>
<tr>
<td>DXS7</td>
<td>1,1</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>1,1</td>
<td>2,1</td>
</tr>
<tr>
<td>MAOA</td>
<td>1,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>1,1</td>
<td>2,1</td>
</tr>
<tr>
<td>DXS1003</td>
<td>1,1</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>1,1</td>
<td>2,1</td>
</tr>
<tr>
<td>DXS426</td>
<td>1,1</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>1,1</td>
<td>2,1</td>
</tr>
<tr>
<td>DXS722</td>
<td>1,1</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>1,1</td>
<td>2,1</td>
</tr>
<tr>
<td>DXS573</td>
<td>1,1</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>1,1</td>
<td>2,1</td>
</tr>
<tr>
<td>DXS988</td>
<td>1,1</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>2,2</td>
<td>1,1</td>
<td>1,1</td>
<td>2,1</td>
</tr>
</tbody>
</table>
(indicated as unknown carrier status). This could hence be an example of random X-
inactivation with an affected genotype but a clinically inconclusive phenotype (section
3.1.3). Her sister V-4 is a recombinant with respect to her obligate carrier mother IV-1,
segregating for the disease associated haplotype below the marker DXS977. Since she
was also found to be of unknown carrier status this crossover data does not contribute
towards the classification of this family.

In summary RP 91 could not be defined as either RP2 or RP3 because of lack of
informative recombination events. Although the clinically normal male V-1 is not
segregating for the disease haplotype above the marker DXS1003, the RP2 disease could
not be excluded since the RP2 disease gene could lie between markers MAOA and
DXS1003. Hence further markers between these two loci need to be tested such as
DXS8080, DXS8026 and DXS1055 (Table 3.1). It could however also be a RP3 family
because of individual V-2 who is segregating for the disease haplotype above DXS1003
or a RP6 or RP15 family since markers at these loci have not been tested. The pedigree of
XLRP 91 (Figure 3.7) shows haplotype data for 9 of the markers studied and the Table 3.7
shows haplotype data for all the 12 markers studied.

3.3.7 Family RP 92

British family XLRP 92 (Figure 3.8) is made up of 3 generations with DNA
available for 2 affected males, 2 obligate carrier females and 1 clinically described carrier
female.

3.3.7.1 Clinical diagnosis of affected males and carrier females

The affected males in this family expressed early onset of night blindness (< 10
years of age) and also a low form of myopia. There was no evidence of cone dystrophy in
the affected males. On fundus examination, tapetal reflex was not observed in the carrier
females. Individual IV-2 has abnormal ERG and fundus appearance, hence she is a
carrier of the disease.

3.3.7.2 Haplotype analysis of XLRP 92

Haplotype analysis was studied in family RP 92 using 18 markers (Figure 3.8 &
Table 3.8). The affected male IV-4 is recombinant with respect to his obligate carrier
mother III-1 with a crossover below DXS572 to the disease-associated haplotype, so the
possibility of the RP6 or RP15 loci being responsible for disease in this family can be
excluded. The affected carrier female IV-2 is also recombinant with respect to her
obligate carrier mother III-1, below the marker DXS572 (since marker DXS1110 is
3. Characterisation of XLRP pedigrees using haplotype analysis

Family XLRP 92

Pedigree of XLRP 92 showing haplotypes constructed with the 11 markers listed.

Pedigree of XLRP 92 showing haplotypes constructed with the 11 markers listed. XLRP 92 is defined as a family segregating for the RP3 form of disease due to data from affected carrier female IV-2.

The boxed genotype shows alleles segregating with the disease associated haplotype and the unboxed genotype indicates alleles segregating with the non-disease haplotype.

The genotype in the hatched box of affected carrier female IV-2 indicates uninformative marker DXS1110.

Normal males and females
Affected males
Obligate carrier females
 Clinically described as heterozygote carrier female

Crossed symbols represent deceased family members

FIGURE 3.8
### TABLE 3.8 Haplotype data for 18 microsatellite markers used in the study of XLRP 92.

<table>
<thead>
<tr>
<th>Individual of Generation</th>
<th>III 1</th>
<th>IV 2</th>
<th>IV 3</th>
<th>IV 4</th>
<th>V 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenotype</td>
<td>Carrier female</td>
<td>Carrier female</td>
<td>Affected male</td>
<td>Affected male</td>
<td>Carrier female</td>
</tr>
<tr>
<td>Microsatellite Marker</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
</tr>
<tr>
<td>DXS572</td>
<td>1, 2</td>
<td>1, 2</td>
<td>1</td>
<td>2</td>
<td>1, 2</td>
</tr>
<tr>
<td>DXS1110</td>
<td>1, 1</td>
<td>1, 3</td>
<td>1</td>
<td>1</td>
<td>1, 2</td>
</tr>
<tr>
<td>DXS556</td>
<td>2, 3</td>
<td>3, 1</td>
<td>2</td>
<td>2</td>
<td>2, 1</td>
</tr>
<tr>
<td>DXS574</td>
<td>3, 2</td>
<td>2, 1</td>
<td>3</td>
<td>3</td>
<td>3, 1</td>
</tr>
<tr>
<td>DXS977</td>
<td>4, 1</td>
<td>1, 3</td>
<td>4</td>
<td>4</td>
<td>4, 2</td>
</tr>
<tr>
<td>DXS228</td>
<td>2, 1</td>
<td>1, 2</td>
<td>2</td>
<td>2</td>
<td>2, 3</td>
</tr>
<tr>
<td>DXS7</td>
<td>2, 2</td>
<td>2, 2</td>
<td>2</td>
<td>2</td>
<td>2, 1</td>
</tr>
<tr>
<td>MAOA</td>
<td>2, 3</td>
<td>3, 2</td>
<td>2</td>
<td>2</td>
<td>2, 1</td>
</tr>
<tr>
<td>DXS1055</td>
<td>1, 1</td>
<td>1, 1</td>
<td>1</td>
<td>1</td>
<td>1, 1</td>
</tr>
<tr>
<td>DXS1003</td>
<td>2, 2</td>
<td>2, 1</td>
<td>2</td>
<td>2</td>
<td>2, 2</td>
</tr>
<tr>
<td>SYN1</td>
<td>2, 1</td>
<td>1, 1</td>
<td>2</td>
<td>2</td>
<td>2, 1</td>
</tr>
<tr>
<td>DXS426</td>
<td>3, 2</td>
<td>2, 1</td>
<td>3</td>
<td>3</td>
<td>3, 4</td>
</tr>
<tr>
<td>DXS722</td>
<td>2, 2</td>
<td>2, 1</td>
<td>2</td>
<td>2</td>
<td>2, 2</td>
</tr>
<tr>
<td>DXS1126</td>
<td>1, 2</td>
<td>2, 2</td>
<td>1</td>
<td>1</td>
<td>1, 1</td>
</tr>
<tr>
<td>DXS573</td>
<td>1, 3</td>
<td>3, 3</td>
<td>1</td>
<td>1</td>
<td>1, 2</td>
</tr>
<tr>
<td>DXS1039</td>
<td>1, 3</td>
<td>3, 4</td>
<td>1</td>
<td>1</td>
<td>1, 2</td>
</tr>
<tr>
<td>DXS988</td>
<td>2, 1</td>
<td>1, 2</td>
<td>2</td>
<td>2</td>
<td>2, 3</td>
</tr>
<tr>
<td>DXS1204</td>
<td>2, 1</td>
<td>1, 1</td>
<td>2</td>
<td>2</td>
<td>2, 3</td>
</tr>
</tbody>
</table>
uninformative). She shows segregation for the RP3 form of disease since she has a crossover below the marker DXS572 from the affected haplotype to the normal haplotype. The phase of the maternal alleles in III-1 was firmly established from analysis of her affected sons IV-3 and IV-4 who share their mother's affected haplotype.

The RP 92 family is therefore classified as segregating for the RP3 disease due to segregation of the disease haplotype in the clinically affected carrier female IV-2. The RP6 and RP15 loci are excluded here because of the data from individual IV-4. Table 3.8 presents haplotype data for all 18 markers used in this study, while Figure 3.8 presents data for 11 of these markers.

3.3.8 Family RP 93

XLRP 93 (Figure 3.9) is a large British family made up of 5 generations with DNA available for 2 affected males, 3 normal males, 6 obligate carrier females, 1 clinically described carrier female and 2 females of unknown carrier status.

3.3.8.1 Clinical diagnosis of affected males and carrier females

On clinical assessment individual III-3 is a carrier female with abnormal fundus changes and ERG readings. Individuals III-6, III-7, IV-3 and IV-4 were not clinically examined but are all obligate carrier females. Individuals III-8 and IV-7 are asymptomatic females and both IV-10 and IV-11, sons of III-8 are clinically unaffected males.

3.3.8.2 Haplotype analysis of XLRP 93

Eight microsatellite markers were used in the haplotype analysis of RP 93 (Figure 3.9). Individual IV-5 is an affected male, recombinant with respect to his grandmother II-2 below the marker DXS977 to the non-disease haplotype showing segregation with either an RP3, RP6 or RP15 disease associated haplotype. He inherited his recombinant chromosome from his mother III-7. His obligate carrier sisters IV-4 and half sister IV-3 are also recombinants inheriting their mother's complete disease-associated haplotype. The phase of the maternal alleles in II-2 is firmly established from analysis of her obligate carrier sister II-1.

Individual III-8 is recombinant with respect to her mother II-2 with disease-associated haplotype above DXS977, however she appears clinically normal. She is hence denoted as of unknown carrier status since her disease associated genotype might not be evident due to X-inactivation (section 3.1.3). Her sons IV-10 and IV-11 were both found to be unaffected, having inherited their mother's non-disease haplotype.
FIGURE 3.9
Pedigree of XLRP 93 showing haplotypes constructed with the 8 markers listed. This family could be segregating for the RP3 disease locus because of the data from recombinants III-7, IV-3, IV-4 and IV-5. The possibility of disease loci RP6 or RP15 being responsible for the disease in this family cannot be excluded.

The boxed genotype indicates alleles segregating with the disease associated haplotype and the unboxed genotype indicates alleles segregating with the non-disease haplotype.

- Normal males and females
- Affected males
- Obligate carrier females
- Clinically described as heterozygote carrier female
- Females of unknown clinical status

Crossed symbols represent deceased family members.
3. Characterisation of XLRP pedigrees using haplotype analysis

Recombinant IV-7 (with respect to her obligate carrier mother III-7) is a clinically normal female, she is also denoted as of unknown carrier status for the same reason as individual III-8. Her X chromosome inherited from her mother harbours a crossover below DXS1110 to the affected haplotype for markers DXS556 and DXS977 and the normal haplotype below the marker DXS977. However no conclusions can be drawn from her crossover data due to the absence of a clear clinical phenotype.

Individuals III-3 (carrier female), III-6 (obligate carrier female) and IV-1 (affected male) are all non-recombinants and hence do not contribute to disease gene localisation. In summary RP 93 could be a RP3 family because of the data from affected male IV-5 and all 3 obligate carrier females III-7, IV-3 and IV-4 which appear to segregate for an RP3 disease-associated haplotype. The possibility of the family mapping to the RP6 or RP15 lod cannot be excluded since the relevant markers above DXS1110 such as DXS572 and DXS1214 (Table 3.1) have not been tested. Figure 3.9 of the pedigree of XLRP 93 shows haplotype data on all the 8 microsatellite markers used in this analysis.

3.9 Family RP 94

RP 94 (Figure 3.10) is a British family of 3 generations with DNA available for 5 affected males, 2 normal males, 5 obligate carrier females and 2 females of unknown carrier status.

3.9.1 Clinical diagnosis of affected males and carrier females

On clinical assessment affected males showed early onset (<10 years of age) of night blindness with no myopia. There was also no evidence of cone dystrophy in the affected males. No clinical data was available for any of the obligate carrier females or females of unknown carrier status.

3.9.2 Haplotype analysis of XLRP 94

Haplotype analysis was carried out using 14 markers (Figure 3.10 & Table 3.9). Individual IV-2, an affected male, is recombinant with respect to his obligate carrier mother III-1, proximal to the marker DXS1110 (since DXS6679 is uninformative). He could therefore either be segregating the RP2 and RP3 disease loci. The phase of the maternal alleles in III-1 was confirmed by analysis of her obligate carrier sibs (III-3, III-4 and III-5) who share their sister’s affected haplotype.

Individuals IV-7 and IV-13 are non-recombinants having inherited their mother’s complete affected genotypes. The affected males IV-11, IV-12 are recombinant below marker DXS989 showing segregation at either the RP2 and RP3 disease loci. Individuals
FIGURE 3.10

Pedigree of XLRP 94 shows haplotypes constructed with 10 of the markers listed. This family could not be defined as either an RP2 or RP3 family due to lack of recombinants. The boxed genotype shows alleles segregating with the disease-associated haplotype and the unboxed genotype alleles segregating with the non-disease haplotype.

□ ○ Normal males and females
■ × Affected males
● ○ Obligate carrier females
○ ○ Females of unknown carrier status

Crossed symbols represent deceased family members.
### TABLE 3.9 Haplotype data for 14 microsatellite markers used in the study of XLRP 94.

<table>
<thead>
<tr>
<th>Individual of Generation</th>
<th>III 1</th>
<th>III 3</th>
<th>III 4</th>
<th>III 5</th>
<th>III 5a</th>
<th>IV 1</th>
<th>IV 2</th>
<th>IV 3</th>
<th>IV 6</th>
<th>IV 7</th>
<th>IV 9</th>
<th>IV 11</th>
<th>IV 12</th>
<th>IV 13</th>
<th>IV 14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenotype</td>
<td>Carrier female</td>
<td>Carrier female</td>
<td>Carrier female</td>
<td>Carrier female</td>
<td>Normal male</td>
<td>Carrier female</td>
<td>Affected male</td>
<td>Normal male</td>
<td>Affected male</td>
<td>Normal male</td>
<td>Affected male</td>
<td>Normal male</td>
<td>Affected male</td>
<td>Affected male</td>
<td>Carrier female</td>
</tr>
<tr>
<td>Microsatellite Marker</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
</tr>
<tr>
<td>DXS989</td>
<td>1,2</td>
<td>1,2</td>
<td>1,2</td>
<td>1,2</td>
<td>4</td>
<td>1,3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2,3</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>DXS1214</td>
<td>1,3</td>
<td>1,1</td>
<td>1,1</td>
<td>1,1</td>
<td>2</td>
<td>1,3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1,2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>DXS1242</td>
<td>3,2</td>
<td>3,1</td>
<td>3,1</td>
<td>3,1</td>
<td>3</td>
<td>3,2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3,1</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>DXS1110</td>
<td>1,2</td>
<td>1,1</td>
<td>1,1</td>
<td>1,1</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>DXS6679</td>
<td>3,3</td>
<td>3,2</td>
<td>3,2</td>
<td>3,3</td>
<td>1</td>
<td>3,1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3,3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>DXS1068</td>
<td>4,2</td>
<td>4,3</td>
<td>4,3</td>
<td>4,2</td>
<td>1</td>
<td>4,2</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4,1</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>DXS556</td>
<td>1,2</td>
<td>1,2</td>
<td>1,2</td>
<td>1,2</td>
<td>1</td>
<td>1,1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1,2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>DXS574</td>
<td>4,2</td>
<td>4,1</td>
<td>4,1</td>
<td>4,2</td>
<td>4</td>
<td>4,4</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4,3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>DXS977</td>
<td>2,3</td>
<td>2,1</td>
<td>2,1</td>
<td>2,3</td>
<td>2</td>
<td>2,3</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>DXS7</td>
<td>2,2</td>
<td>2,2</td>
<td>2,2</td>
<td>2,2</td>
<td>2</td>
<td>2,1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>MAOA</td>
<td>3,2</td>
<td>3,2</td>
<td>3,2</td>
<td>3,2</td>
<td>4</td>
<td>3,1</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>SYN1</td>
<td>4,2</td>
<td>4,2</td>
<td>4,2</td>
<td>4,2</td>
<td>2</td>
<td>4,2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>DXS426</td>
<td>2,2</td>
<td>2,2</td>
<td>2,2</td>
<td>2,2</td>
<td>2</td>
<td>2,1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>DXS1126</td>
<td>1,1</td>
<td>1,1</td>
<td>1,2</td>
<td>1,2</td>
<td>1</td>
<td>1,3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1,3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
3. Characterisation of XLRP pedigrees using haplotype analysis

IV-3 and IV-6 have inherited their mother's (III-1 and III-3 respectively) non-disease haplotype. Individual IV-9 is denoted as an obligate carrier female since she has an affected son (not shown) who was not included in the analysis. Individual IV-14, (female of unknown carrier status) has inherited the disease-associated haplotype below marker DXS989, but was found to be myopic with no other signs or symptoms of RP.

From the haplotype data for individual IV-2 (affected male) and also all the other individuals studied who are non-recombinant for markers between DXS989 to DXS1126, the RP 94 family could not be defined as one segregating for the RP2 or the RP3 disease loci. The RP6 and RP15 disease loci can be excluded due to crossover data below DXS989 from the normal to the affected haplotype in individuals IV-9, IV-11 and IV-12. Table 3.9 presents haplotype data for all 14 markers used in this study, while Figure 3.10 presents data for 10 of these markers to facilitate convenience of haplotype readings.

3.3.10 Family RP 1120

American family XLRP 1120 (Figure 3.11) is a large family made up of 5 generations with DNA available for 4 affected males, 2 normal males, 6 obligate carrier females and 2 females of unknown carrier status.

3.3.10.1 Clinical diagnosis of affected males and carrier females

On clinical assessment males III-2, IV-3, IV-8 and IV-9 were affected with abnormal fundus and ERG readings. They all suffered from night blindness. Females IV-1 and IV-2 are asymptomatic but clinically of unknown carrier status.

3.3.10.2 Haplotype analysis of XLRP 1120

Haplotype analysis was studied in family RP 1120 using 11 microsatellite markers (Figure 3.11). One or both obligate carrier females II-1 and II-4 could be recombinant with respect to their obligate carrier mother I-1. Since their mother had passed away her DNA was not available for analysis and hence her exact affected haplotype could not be established. In any case both carrier females II-1 and II-4 have passed on their disease-associated haplotype (segregating just proximal to CYBB for marker DXS1110 and just distal to DXS228) to their affected children. Affected male III-2 is a recombinant with respect to his mother II-1 but retains the affected haplotype above the marker DXS228. Individuals III-2 and III-1 who are the affected son and daughter respectively of II-1 have a common genotype for the 3 markers DXS1110, DXS556 and DXS977. Their unaffected sibs III-3 and III-4 have inherited the normal haplotype for these markers from their mother (II-1).
Family XLRP 1120

I

II

III

IV

Figure 3.11
Pedigree of XLRP 1120 showing haplotypes constructed with the 11 markers listed. This family is classified as an RP3 family.
The boxed genotype indicates alleles segregating with the disease associated haplotype and the unboxed genotype alleles segregating with the non-disease haplotype.

Normal males and females
Affected males
Obligate carrier females
Females of unknown carrier status
Crossed symbols represent deceased family members.

Inferred genotype

3. Characterisation of XLRP pedigrees using haplotype analysis
Both affected sibs III-1 and III-2 have passed on their disease-associated haplotype from markers DXS1110 to DXS977 to their children IV-3 and IV-4 respectively. Affected son IV-3 is a recombinant with respect to his mother III-1 having inherited the disease-associated haplotype for markers DXS1110, DXS556 and DXS977, due to a crossover just proximal to CYBB. His unaffected sisters are segregating for the normal haplotype for the same markers. Obligate carrier females III-7 and III-8 also segregate for the disease-associated alleles at markers DXS1110, DXS556 and DXS977, having inherited their mother's affected haplotype for these markers. Obligate carrier female III-8 has passed on her disease haplotype to her affected sons IV-8 and IV-9. In summary the XLRP 1120 family can be classified as an RP3 family as all the affected males and obligate carrier females studied share the affected genotype at markers DXS1110, DXS556 and DXS977 which span the RP3 critical interval. Figure 3.11 of the pedigree of XLRP 1120 shows haplotype data on all the 11 microsatellite markers used in this analysis.

3.3.11 Family RP 88/10

Italian family XLRP 88/10 (Figure 3.12) comprises 3 generations with DNA available for 5 normal males and 5 clinically described heterozygote carrier females.

3.3.11.1 Clinical diagnosis of affected males and carrier females

On clinical diagnosis males III-2, III-4, IV-2, IV-3 and IV-5 were found to be definitely asymptomatic. Individuals II-1, III-1, III-3, IV-1 and IV-6 were ascertained clinically symptomatic for RP hence being described as carrier females.

3.3.11.2 Haplotype analysis of XLRP 88/10

In family RP 88/10 (Figure 3.12 & Table 3.10) 14 markers were used in haplotype analysis. Individual III-4 a normal male (45 years old) is recombinant with respect to his clinically affected mother II-1, inheriting the affected haplotype below the marker DXS556. This suggests segregation of the RP3 locus in this family. The phase of the maternal alleles of II-1 was confirmed from the haplotype data of her normal son III-2 and her obligate carrier daughter III-1. Carrier females III-1, III-3, IV-1 and IV-6 are all non-recombinant as are the normal males III-2, IV-2, IV-3 and IV-5. In summary, the recombinant event observed in normal male III-4 classified this family as an RP3 family. Figure 3.12 of pedigree XLRP 88/10 shows haplotype data for 10 microsatellite markers used in this study while Table 3.10 indicates haplotype data for all 14 markers used.
**Family XLRP 88/10**

**II**

- CYBB: 2 2 4
- DXS1110: 2 2 1
- DXS556: 3 2 2
- DXS977: 3 1 1
- DXS7: 3 2 1
- MAOA: 3 3 1
- DXS1003: 4 1 3
- DXS426: 3 3 1
- DXS1126: 1 1 3
- DXS573: 2 3 4

**III**

- CYBB: 2 4 3 2 2 4 2 4
- DXS1110: 2 1 1 2 2 1 2 2
- DXS556: 3 2 1 2 3 2 2 2
- DXS977: 3 1 3 1 3 1 3 1
- DXS7: 3 1 1 2 3 1 3 2
- MAOA: 3 1 1 3 3 1 3 3
- DXS1003: 4 3 1 1 4 3 3 4 1 3
- DXS426: 3 1 3 3 3 1 3 3 1
- DXS1126: 1 3 2 1 1 3 3 1 1 3
- DXS573: 2 4 1 3 2 4 4 2 3 4

**IV**

- CYBB: 2 3 4 1 4 2 4
- DXS1110: 2 1 1 1 1 2 1
- DXS556: 3 1 2 3 2 3 2
- DXS977: 3 3 1 2 1 3 3 3
- DXS7: 3 1 1 3 1 3 3 3
- MAOA: 3 1 1 4 1 3 2
- DXS1003: 4 1 3 2 3 4 1
- DXS426: 3 3 1 2 1 3 3 3
- DXS1126: 1 2 3 2 3 1 1
- DXS573: 2 1 4 4 4 2 2

**FIGURE 3.12**

Pedigree of XLRP 88/10 showing haplotypes constructed with the 10 markers listed. Recombinant normal male III-4 defines this family as an RP3 family.

The boxed genotype shows alleles segregating with the disease associated haplotype and the unboxed genotype for alleles segregating with the non-disease haplotype.

- □ O Normal males and females
- O Clinically described as heterozygote carrier female
- ■ Affected males

Crossed symbols represent deceased family members
**TABLE 3.10**  Haplotype data for 14 microsatellite markers used in the study of XLRP 88/10.

<table>
<thead>
<tr>
<th>Individual of Generation</th>
<th>Phenoype</th>
<th>CYBB</th>
<th>DXS1110</th>
<th>DXS556</th>
<th>DXS9977</th>
<th>DXS228</th>
<th>DXS7</th>
<th>MAOA</th>
<th>DXS1003</th>
<th>SYN1</th>
<th>DXS426</th>
<th>DXS722</th>
<th>DXS1126</th>
<th>DXS573</th>
<th>DXS988</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Carrier female</td>
<td>Normal male</td>
<td>Carrier female</td>
<td>Normal male</td>
<td>Carrier female</td>
<td>Normal male</td>
<td>Carrier female</td>
<td>Normal male</td>
<td>Carrier female</td>
<td>Normal male</td>
<td>Carrier female</td>
<td>Normal male</td>
<td>Carrier female</td>
<td>Normal male</td>
<td>Carrier female</td>
</tr>
<tr>
<td>II 1</td>
<td>2,2</td>
<td>4</td>
<td>2,4</td>
<td>3</td>
<td>2</td>
<td>2,4</td>
<td>4</td>
<td>2</td>
<td>2,4</td>
<td>3</td>
<td>2</td>
<td>2,4</td>
<td>4</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>II 1a</td>
<td>2,2</td>
<td>1</td>
<td>2,1</td>
<td>1</td>
<td>2</td>
<td>2,1</td>
<td>1</td>
<td>2</td>
<td>2,1</td>
<td>1</td>
<td>2</td>
<td>2,1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>II 1a</td>
<td>3,2</td>
<td>2</td>
<td>3,2</td>
<td>1</td>
<td>2</td>
<td>3,2</td>
<td>2</td>
<td>2</td>
<td>2,2</td>
<td>3,1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>III 1</td>
<td>3,1</td>
<td>1</td>
<td>3,1</td>
<td>3</td>
<td>1</td>
<td>3,1</td>
<td>1</td>
<td>3</td>
<td>1,1</td>
<td>3,3</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>III 1a</td>
<td>3,2</td>
<td>3</td>
<td>3,3</td>
<td>2</td>
<td>2</td>
<td>3,3</td>
<td>3</td>
<td>3</td>
<td>2,3</td>
<td>3,2</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>III 3</td>
<td>3,2</td>
<td>1</td>
<td>3,1</td>
<td>1</td>
<td>2</td>
<td>3,1</td>
<td>1</td>
<td>3</td>
<td>2,1</td>
<td>3,1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3,3</td>
</tr>
<tr>
<td>III 3a</td>
<td>3,3</td>
<td>1</td>
<td>3,1</td>
<td>1</td>
<td>3</td>
<td>3,1</td>
<td>1</td>
<td>3</td>
<td>3,1</td>
<td>3,1</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>3,2</td>
</tr>
<tr>
<td>IV 1</td>
<td>4,1</td>
<td>3</td>
<td>4,3</td>
<td>1</td>
<td>1</td>
<td>4,3</td>
<td>3</td>
<td>4</td>
<td>1,3</td>
<td>3,3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4,1</td>
</tr>
<tr>
<td>IV 2</td>
<td>4,1</td>
<td>3</td>
<td>4,3</td>
<td>1</td>
<td>1</td>
<td>4,3</td>
<td>3</td>
<td>4</td>
<td>1,3</td>
<td>3,3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4,1</td>
</tr>
<tr>
<td>IV 3</td>
<td>3,2</td>
<td>2</td>
<td>3,2</td>
<td>1</td>
<td>2</td>
<td>3,2</td>
<td>2</td>
<td>3</td>
<td>2,2</td>
<td>3,1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>3,3</td>
</tr>
<tr>
<td>IV 4</td>
<td>3,3</td>
<td>1</td>
<td>3,1</td>
<td>1</td>
<td>3</td>
<td>3,1</td>
<td>1</td>
<td>3</td>
<td>3,1</td>
<td>3,3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3,3</td>
<td></td>
</tr>
<tr>
<td>IV 5</td>
<td>2,2</td>
<td>2</td>
<td>2,2</td>
<td>1</td>
<td>2</td>
<td>2,2</td>
<td>2</td>
<td>2</td>
<td>2,2</td>
<td>2,1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2,1</td>
</tr>
<tr>
<td>IV 6</td>
<td>1,1</td>
<td>3</td>
<td>1,3</td>
<td>2</td>
<td>1</td>
<td>1,3</td>
<td>3</td>
<td>1</td>
<td>1,3</td>
<td>1,2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1,1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Characterisation of XLRP pedigrees using haplotype analysis.
3.3.12 Family RP 88/75

XLRP 88/75 (Figure 3.13) is an Italian family which comprises 4 generations with DNA available for 4 affected males, 3 normal males, 4 obligate carrier females, 1 clinically described carrier female and 1 female of unknown clinical status.

3.3.12.1 Clinical diagnosis of affected males and carrier females

Affected males II-1, IV-1, IV-2 and V-1 were found to be clinically symptomatic. Individual IV-1 is an affected male (26 years old) with attenuated retinal vessels and diffuse pigmented changes. His visual acuity is also poor with abnormal ERG readings. His cousin of 35 years old, IV-4 is clinically asymptomatic. Female IV-7 has abnormal fundus and ERG readings, hence is clinically described as a carrier female.

3.3.12.2 Haplotype analysis of XLRP 88/75

Haplotype analysis was performed using 15 markers (Figure 3.13 & Table 3.11). Individual IV-1, an affected male is recombinant with respect to his obligate carrier mother III-1 and has a crossover proximal to the marker DXS977 from the normal haplotype to the affected, which proves this is an RP2 family. The phase of the maternal alleles of III-1 is confirmed from the haplotype data of her affected father II-1 and her affected carrier sibs III-2 and III-3, all being non-recombinants for markers CYBB to DXS988 (refer to Table 3.11).

Clinically normal male IV-4 (35 years old) is recombinant with respect to his carrier mother III-2 below the marker DXS228 (markers DXS7 and MAOA are uninformative) to the non-disease haplotype. This crossover defines the distal boundary of the XLRP2 interval, with the disease associated haplotype located proximal to DXS228. The affected son V-1 of obligate carrier female IV-5 is recombinant above DXS1110, placing the disease below this marker.

Individuals IV-2, IV-5 and IV-7 with the affected haplotype and normal individuals IV-3 and IV-6 are all non-recombinants and hence do not provide any additional locus refinement data. The fact that IV-7, a carrier female, carries the disease-associated haplotype agrees with her clinically established phenotype. In summary the two essential crossovers in IV-1 and IV-4 provide evidence that family RP88/75 is an RP2 family. Figure 3.13 shows haplotype data for 10 of the markers used in the study indicating the essential crossovers while Table 3.11 shows haplotype data for all 15 markers used to help refine the XLRP2 interval.
3. Characterisation of XLRP pedigrees using haplotype analysis

Family XLRP 88/75

**Figure 3.13**
Pedigree of XLRP 88/75 showing haplotypes constructed with the 10 markers listed.
Recombinant individuals IV-1 & IV-4 define this family as an RP2 family.
The boxed genotype shows alleles segregating with the disease-associated haplotype and the unbound genotype shows alleles segregating with the non-disease haplotype.
The genotype in the hatched box of clinically normal male IV-4 indicates uninformative markers DXS7 and MAOA.

- Normal males and females
-Affected males
-Obligate carrier females
-Clinically described as heterozygote carrier female
-Female of unknown carrier status

189
TABLE 3.11  Haplotype data for 15 microsatellite markers used in the study of XLRP 88/75.

<table>
<thead>
<tr>
<th>Individual of Generation</th>
<th>II 1</th>
<th>II 1a</th>
<th>III 1</th>
<th>III 2</th>
<th>IV 1</th>
<th>IV 2</th>
<th>IV 3</th>
<th>IV 4</th>
<th>IV 5</th>
<th>IV 6</th>
<th>IV 7</th>
<th>V 1</th>
<th>V 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenotype</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Affected male</td>
<td>Normal female</td>
<td>Carrier female</td>
<td>Carrier female</td>
<td>Affected male</td>
<td>Affected male</td>
<td>Normal male</td>
<td>Normal male</td>
<td>Carrier female</td>
<td>Normal male</td>
<td>Carrier female</td>
<td>Affected male</td>
<td>Normal female</td>
</tr>
<tr>
<td>Microsatellite Marker</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
</tr>
<tr>
<td>CYBB</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>DXS1110</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>DXS556</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>DXS977</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>DXS228</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>DXS7</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>MAOA</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>DXS8080</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>DXS8026</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>DXS1055</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>DXS1003</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>DXS426</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>DXS722</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>DXS573</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>DXS988</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

3. Characterisation of XLRP pedigrees using haplotype analysis
3. Characterisation of XLRP pedigrees using haplotype analysis

3.3.13 Choroideremia family 22 (CHM 22)
(Spanish family misdiagnosed as XLRP)

Spanish family CHM 22 (Figure 3.14) comprises 3 generations with DNA obtained for 2 affected males, 2 unaffected males, 2 obligate carrier females and 3 clinically described carrier females.

3.3.13.1 Clinical diagnosis of affected males and carrier females

Clinical diagnosis was carried out by clinicians in Italy. The individual IV-1 of 14 years old is totally asymptomatic suffering neither night blindness nor other symptoms of RP. He showed normal visual field, visual acuity and normal ERG readings. Hence he was concluded to be unaffected. The individual III-6 (carrier female of 44 years old) is asymptomatic and did not suffer from either night blindness or constriction of the visual field although she had asymmetric ERG. Her fundus examination showed normal macula and disc in both eyes and no tapetal reflex. There was pigmentary migration but not in the typical 'bone spicule fashion' of RP but with a granular pigmentary appearance in the periphery. Her daughter IV-2, of 13 years of age is also asymptomatic with an abnormal ERG like her mother’s. No tapetal reflex was seen but there was grey granular appearance in the pigmented epithelium. Since the slides and photographs sent to us showed more of the macula than the peripheral retina, the clinical data was inconclusive.

Choroideremia (section 1.8.6) another X-linked eye disease on Xq has similar peripheral disease as RP. Affected males in this disease show areas of total choroidal vascular atrophy with macular sparing and progressive dystrophy of choroid, retinal pigmented epithelium (RPE) and retina. It produces visual field constriction and results in blindness by the 4th to 5th decade of life. Heterozygous females show no visual defects but striking funduscopic changes such as 'salt and pepper pigmentation' (peripheral RPE granularity) and atrophy around the optic disc. Abnormal retinal function tests are also evident.

As the available clinical data did not unambiguously indicate XLRP, further information was requested for affected male III-5 and his carrier sister III-3. The clinical data obtained subsequently from these individuals confirmed that the disease in this family was indeed choroideremia. Carrier female III-3 is asymptomatic having normal visual field and acuity and normal ERG readings. On fundus examination, her retinal pigmented epithelium (RPE) appeared delustrated and with granularity, not in the typical 'bone spicule' deposits of pigment as in RP. They were yellow areas 'flecks like' around the maculae in the whole posterior pole and around large vessels. This
3. Characterisation of XLRP pedigrees using haplotype analysis

Family CHM 22 haplotyped using Xp21.1-Xp11.23 markers

**FIGURE 3.14**

Pedigree of CHM 22 shows haplotypes constructed with the 10 markers listed in the Xp21.1-Xp11.23 region. This family was clinically misdiagnosed as RP initially but was then later clinically diagnosed as being affected with Choroideremia. The boxed genotype shows alleles segregating with the disease-associated haplotype and the unbound genotype shows alleles segregating with the non-disease haplotype. The genotype in the hatched box of carrier female III-6 indicates uninformative marker DXS722.

- **CYBB**
- **DXS1110**
- **DXS977**
- **DXS228**
- **DXS7**
- **MAOA**
- **DXS426**
- **DXS722**
- **DXS1126**
- **DXS573**

Legend:
- Normal males and females
- Obligate carrier females
- Clinically described as heterozygote carrier females

192
TABLE 3.12 Haplotype data for 15 microsatellite markers in the Xp21.1-Xp11.23 region used in the study of CHM 22 family.

<table>
<thead>
<tr>
<th>Individual of Generation</th>
<th>II 1</th>
<th>II 2</th>
<th>II 2a</th>
<th>III 2</th>
<th>III 3</th>
<th>III 4</th>
<th>III 5</th>
<th>III 5a</th>
<th>IV 1</th>
<th>IV 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenotype</td>
<td>Carrier female</td>
<td>Carrier female</td>
<td>Normal male</td>
<td>Affected male</td>
<td>Normal male</td>
<td>Affected male</td>
<td>Carrier female</td>
<td>Normal male</td>
<td>Carrier female</td>
<td></td>
</tr>
<tr>
<td>Microsatellite Marker</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
<td>Alleles</td>
</tr>
<tr>
<td>CYBB</td>
<td>1,2</td>
<td>1,2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2,3</td>
<td>3</td>
<td>2</td>
<td>2,3</td>
</tr>
<tr>
<td>DXS1110</td>
<td>1,1</td>
<td>1,1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1,1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>DXS556</td>
<td>3,2</td>
<td>3,2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2,1</td>
<td>3</td>
<td>2</td>
<td>2,3</td>
</tr>
<tr>
<td>DXS977</td>
<td>3,2</td>
<td>3,2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2,1</td>
<td>3</td>
<td>2</td>
<td>2,3</td>
</tr>
<tr>
<td>DXS228</td>
<td>3,2</td>
<td>3,2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2,1</td>
<td>1</td>
<td>2</td>
<td>2,1</td>
</tr>
<tr>
<td>DXS7</td>
<td>1,1</td>
<td>1,1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1,2</td>
<td>2</td>
<td>1</td>
<td>1,2</td>
</tr>
<tr>
<td>MAOA</td>
<td>1,2</td>
<td>1,2</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2,4</td>
<td>3</td>
<td>2</td>
<td>2,3</td>
</tr>
<tr>
<td>DXS1003</td>
<td>3,2</td>
<td>3,2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2,1</td>
<td>3</td>
<td>2</td>
<td>2,3</td>
</tr>
<tr>
<td>SYN1</td>
<td>2,1</td>
<td>2,1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1,1</td>
<td>3</td>
<td>1</td>
<td>1,3</td>
</tr>
<tr>
<td>DXS426</td>
<td>3,1</td>
<td>3,1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1,2</td>
<td>1</td>
<td>1</td>
<td>1,1</td>
</tr>
<tr>
<td>DXS722</td>
<td>2,2</td>
<td>2,2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2,2</td>
<td>1</td>
<td>2</td>
<td>2,1</td>
</tr>
<tr>
<td>DXS1126</td>
<td>2,1</td>
<td>2,1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2,2</td>
<td>1</td>
<td>2</td>
<td>2,1</td>
</tr>
<tr>
<td>DXS573</td>
<td>3,4</td>
<td>3,4</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3,2</td>
<td>1</td>
<td>3</td>
<td>3,1</td>
</tr>
<tr>
<td>DXS1000</td>
<td>1,2</td>
<td>1,2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1,2</td>
<td>2</td>
<td>1</td>
<td>1,2</td>
</tr>
<tr>
<td>DXS988</td>
<td>1,3</td>
<td>1,3</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1,2</td>
<td>1</td>
<td>1</td>
<td>1,1</td>
</tr>
</tbody>
</table>
phenotype is typical of choroideremia patients and not RP patients. The affected male III-5 of 43 years old confirmed this since he had night blindness and visual field constriction. He complained of poor visual acuity at 14 years old. He had generalised and severe choroidal atrophy with areas of bare sclera with some preserved zones of normal RPE - more typical of choroideremia males.

3.3.13.2 Haplotype analysis of CHM 22 using markers in the Xp21.1-Xp11.23 region

In family CHM 22 (Figure 3.14 & Table 3.12) 15 markers in the Xp21.1-Xp11.23 region were initially used in haplotype analysis after receiving a diagnosis of XLRP. Based on genetic studies two crossovers were observed, to give evidence that the disease locus was RP2 and not RP3. Normal male III-4 is recombinant with respect to his obligate carrier mother II-2 with a crossover proximal to marker DXS977 from the affected haplotype to the normal. Since the RP3 disease locus has been genetically placed distal to OTC (this marker lies between markers DXS1110 and DXS977) from previous reports (Ott et al., 1990) and since this individual has inherited the affected haplotype of his mother above DXS977 but is phenotypically normal, the disease locus in this family CHM 22 was concluded to be RP2 and not RP3. The phase of the maternal alleles of II-2 was firmly established from analysis of her obligate carrier sib II-1 and her affected son III-5 who all share the same affected haplotype.

The second crucial crossover (Figure 3.14) was found in the carrier female III-6 who is recombinant with respect to her obligate carrier mother II-2 and has a crossover proximal to the marker DXS426 (DXS722 is an uninformative marker here) from the normal haplotype to the affected. This was very interesting data at that time since no crossovers proximal and so close to DXS426 had been reported in other XLRP family studies or in any of our own XLRP families. This would have narrowed the region of search for the RP2 gene from 13 cM to about 7 cM, a significant refinement. Affected males III-2 and III-5 both non-recombinants, had inherited the whole affected haplotype from markers CYBB to DXS988 (Table 3.12) from their respective mothers. They hence did not contribute to further crossover or refinement data.

Another interesting observation was that the son IV-1 of the carrier female III-6 had inherited the complete affected haplotype of his mother below the marker DXS426 to DXS988 (Table 3.12) but was reported by evidence of clinical data from our Spanish collaborators to be asymptomatic. Three possible explanations could account for this.

(i) There was a double crossover between DXS426 and DXS573 such that he had inherited the normal portion of the chromosome bearing the disease gene from his mother. This could have occurred by recombination below DXS426 to the normal
3. Characterisation of XLRP pedigrees using haplotype analysis

haplotype of this mother and below DXS1126 to the DXS573 allele which was of the affected haplotype. This is a very unlikely event since for a double crossover to occur in a region of 5 cM (distance between marker DXS426 and DXS255; 5th X-chromosome report, 1994) would be a probability of 1 in 400. This figure is obtained since loci which are 5cM apart on a genetic map will show about 5% recombination, hence for a double recombination 5/100 X 5/100 percentage of recombination will be expected.

(ii) The clinical data on this boy was not accurate and he could actually be affected.

(iii) Another alternative explanation would be a more proximal locus for RP2 between DXS988 and the centomere in this family.

His clinically asymptomatic sister (carrier female IV-2) was reported to have abnormal fundus and ERG readings. She had also inherited the same affected haplotype as her brother IV-1 below the marker DXS426 to DXS988. It was hence unusual to find abnormal fundus and ERG readings in the carrier female IV-2 but not in the male IV-1 though both carry the same affected genotype. Figure 3.14 shows haplotype data for 10 of the Xp21.1-Xp11.23 markers used in this study indicating the crossovers while Table 3.12 shows haplotype data for all 15 Xp markers used.

Since this family and its crossovers were of great significance to the genetic mapping of RP2 and the physical mapping strategy of the rest of the group, it was essential to confirm the clinical phenotype from our Spanish collaborators and hence clinical data was obtained both of ERG tests and fundus examinations for the carrier mother (III-6), normal son (IV-1) and carrier daughter (IV-2) as described in Section 3.3.13.1. Since the clinical data sent was inconclusive other clinical data was requested for carrier female III-3 and her affected brother III-5. From subsequent clinical data obtained for these individuals, it was confirmed that the disease involved was choroideremia. Choroideremia (CHM) is an X-linked chorioretinal degeneration and maps to the Xq21.1-Xq21.2 region, hence markers linked to this disease locus were investigated in this family.

3.3.13.3 Haplotype analysis of CHM 22 using markers in the Xq21.1-Xq21.2 region of the Choroideremia disease interval.

Six markers mapping to Xq21.1-Xq21.2 (Figure 3.15) and closely linked to CHM were used to study the haplotypes in the same individuals described previously. The individuals III-3 (carrier female), III-5 (affected male) and III-6 (carrier female) have inherited the complete affected haplotype for markers DXS986 to DXS990 from their obligate carrier mother II-2. The phase of the maternal alleles of II-2 was firmly
3. Characterisation of XLRP pedigrees using haplotype analysis

**Family CHM 22 haplotyped using Xq21.1-Xq21.2 markers**

**FIGURE 3.15**

Pedigree of CHM 22 showing haplotypes constructed with the 6 markers listed in the Xq21.1-Xq21.2 region.

Based on both clinical and genetic evidence this family is a choroideremia disease.

The boxed genotype shows alleles segregating with the disease-associated haplotype and the unboxed genotype shows alleles segregating with the non-disease haplotype.

- **□**: Normal males and females
- **■**: Affected males
- **●**: Obligate carrier females
- **○**: Clinically described as heterozygote carrier females

196
established from analysis of her obligate carrier sib II-1 and her affected son III-5. The normal male IV-1 had inherited the complete normal haplotype of his mother III-6 and hence had no disease. His carrier sister IV-2 had inherited the affected haplotype from her mother for markers DX986 to DXS1169 and hence had the disease. In summary based on genetic evidence and subsequent clinical evidence this family can be defined as a choroideremia family. Figure 3.15 of pedigree CHM 22 presents haplotype data on all the six Xq21.1-Xq21.2 microsatellite markers used in this analysis.

3.3.14 Summary of Families

Table 3.13 gives a brief summary of the 12 XLRP families and the 1 choroideremia family studied through haplotype analysis. 40% of the total XLRP families studied could be classified as either RP2 or RP3. Out of these classified families 60% were found to segregate for the RP3 disease locus and 40% for the RP2 disease locus. Seven of the XLRP families could not be assigned to any one of the RP loci on the X chromosome as they lack recombinants necessary to make a clear distinction. From this study the relative frequency of the RP2 and RP3 forms of disease in the British population could not be determined since there were too few non-British families to make a comparison.

**TABLE 3.13 SUMMARY OF FAMILIES**

<table>
<thead>
<tr>
<th>XLRP Family No</th>
<th>Country of Origin</th>
<th>Disease Locus</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHM 22</td>
<td>Spain</td>
<td>Choroideremia (Xq21.1-Xq21.2) (misdiagnosed as XLRP)</td>
</tr>
<tr>
<td>XLRP 1120</td>
<td>America</td>
<td>RP3</td>
</tr>
<tr>
<td>XLRP 92</td>
<td>Britain</td>
<td>RP3</td>
</tr>
<tr>
<td>XLRP 93</td>
<td>Britain</td>
<td>RP3 (confirmed by mutation in RPGR gene; Wright AF., per com.)</td>
</tr>
<tr>
<td>XLRP 88/75</td>
<td>Italy</td>
<td>RP2</td>
</tr>
<tr>
<td>XLRP 89</td>
<td>Britain</td>
<td>RP2</td>
</tr>
<tr>
<td>XLRP 94</td>
<td>Britain</td>
<td>Not determined</td>
</tr>
<tr>
<td>XLRP 88/10</td>
<td>Italy</td>
<td>RP2 or RP3</td>
</tr>
<tr>
<td>XLRP 85</td>
<td>Britain</td>
<td>RP3/RP6 or RP15</td>
</tr>
<tr>
<td>XLRP 87</td>
<td>Britain</td>
<td>RP2/RP3/RP6 or RP15</td>
</tr>
<tr>
<td>XLRP 88</td>
<td>Britain</td>
<td>RP2/RP3/RP6 or RP15</td>
</tr>
<tr>
<td>XLRP 90</td>
<td>Britain</td>
<td>RP2/RP3/RP6 or RP15</td>
</tr>
<tr>
<td>XLRP 91</td>
<td>Britain</td>
<td>RP2/RP3/RP6 or RP15</td>
</tr>
</tbody>
</table>
3.3A Further genetic analysis of the 12 XLRP families

The cloning of the RPGR (retinitis pigmentosa GTPase regulator) gene found within the RP3 genomic interval (Meindl et al., 1996; Roepman et al., 1996) led to mutation screening of 29 XLRP families (Zito et al., 1999) from this laboratory for which the disease segregated for the RP3 locus. Three of the reported XLRP families (RP87, RP90 and RP1120; section 3.3) in this study were found from this screening effort to have new and different mutations as shown in Table 3.14. The RP93 family was found to have a mutation in exon 6 contributing to a stop codon (Wright A, per. comm). The recent identification of the RP2 gene by Schwahn et al., (1998) has also allowed for mutational screening of XLRP families used in this study, however no RP2 gene mutations were found in any of the other eight XLRP families analysed (Hardcastle et al., 1999). Hence in order to further genetically analyse the remaining eight XLRP families for other independent mutations likely to be present at the RP2 or RP3 locus, linkage analysis (section 3.3A.1) and the possibility of founder effect(s) (section 3.3A.2) linked to these XLRP disease loci was investigated.

<table>
<thead>
<tr>
<th>Family</th>
<th>Intron/Exon</th>
<th>DNA Position/Change</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>RP87</td>
<td>Exon 11</td>
<td>1435-1436 2-bp insertion</td>
<td>Frameshift</td>
</tr>
<tr>
<td>RP90</td>
<td>Exon 10</td>
<td>1297 2-bp insertion</td>
<td>Frameshift</td>
</tr>
<tr>
<td>RP93</td>
<td>Exon 6</td>
<td>640 G — &gt; A</td>
<td>Stop codon</td>
</tr>
<tr>
<td>RP1120</td>
<td>Intron 1</td>
<td>86 plus 5 G — &gt; A</td>
<td>Putative splicing</td>
</tr>
</tbody>
</table>

3.3A.1 Linkage analysis of eight European XLRP families

Linkage analysis was carried out on eight European families for markers in Xp21.1-Xp11.22 mapping at the XLRP2 and XLRP3 loci. Standard two-point linkage analysis was performed using the Cyrillic program package version 2.1 (Chapman C, et al., 1990). For the analysis presented here, all women who were obligate carriers and/or had symptoms of the disease were defined as affected and marker allele frequencies were assumed to be equal for all alleles.
3. Characterisation of XLRP pedigrees using haplotype analysis

Table 3.15
Two-point lod scores between Xp21.1-Xp11.22 markers (shown in order from proximal to distal) and XLRP disease phenotype in eight European XLRP families described in section 3.3.

(1) XLRP85

<table>
<thead>
<tr>
<th>Marker</th>
<th>θ=0</th>
<th>0.05</th>
<th>0.10</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>DXS556</td>
<td>0.60</td>
<td>0.56</td>
<td>0.51</td>
<td>0.41</td>
<td>0.29</td>
<td>0.16</td>
</tr>
<tr>
<td>DXS977</td>
<td>0.13</td>
<td>0.11</td>
<td>0.10</td>
<td>0.08</td>
<td>0.05</td>
<td>0.03</td>
</tr>
<tr>
<td>DXS7</td>
<td>1.03</td>
<td>0.94</td>
<td>0.85</td>
<td>0.66</td>
<td>0.46</td>
<td>0.24</td>
</tr>
<tr>
<td>MAOA</td>
<td>1.03</td>
<td>0.94</td>
<td>0.85</td>
<td>0.66</td>
<td>0.46</td>
<td>0.24</td>
</tr>
<tr>
<td>DXS1003</td>
<td>0.73</td>
<td>0.64</td>
<td>0.55</td>
<td>0.37</td>
<td>0.20</td>
<td>0.07</td>
</tr>
<tr>
<td>DXS426</td>
<td>0.43</td>
<td>0.38</td>
<td>0.33</td>
<td>0.22</td>
<td>0.12</td>
<td>0.03</td>
</tr>
<tr>
<td>DXS722</td>
<td>0.43</td>
<td>0.38</td>
<td>0.33</td>
<td>0.22</td>
<td>0.12</td>
<td>0.03</td>
</tr>
<tr>
<td>DXS573</td>
<td>0.13</td>
<td>0.11</td>
<td>0.10</td>
<td>0.08</td>
<td>0.05</td>
<td>0.03</td>
</tr>
</tbody>
</table>

(2) XLRP88

<table>
<thead>
<tr>
<th>Marker</th>
<th>θ=0</th>
<th>0.05</th>
<th>0.10</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>DXS1110</td>
<td>0.60</td>
<td>0.54</td>
<td>0.47</td>
<td>0.32</td>
<td>0.17</td>
<td>0.05</td>
</tr>
<tr>
<td>DXS556</td>
<td>0.60</td>
<td>0.54</td>
<td>0.47</td>
<td>0.32</td>
<td>0.17</td>
<td>0.05</td>
</tr>
<tr>
<td>DXS7</td>
<td>0.60</td>
<td>0.54</td>
<td>0.47</td>
<td>0.32</td>
<td>0.17</td>
<td>0.05</td>
</tr>
<tr>
<td>MAOA</td>
<td>0.60</td>
<td>0.54</td>
<td>0.47</td>
<td>0.32</td>
<td>0.17</td>
<td>0.05</td>
</tr>
<tr>
<td>DXS1003</td>
<td>0.60</td>
<td>0.54</td>
<td>0.47</td>
<td>0.32</td>
<td>0.17</td>
<td>0.05</td>
</tr>
<tr>
<td>DXS426</td>
<td>0.60</td>
<td>0.54</td>
<td>0.47</td>
<td>0.32</td>
<td>0.17</td>
<td>0.05</td>
</tr>
<tr>
<td>DXS1039</td>
<td>0.60</td>
<td>0.54</td>
<td>0.47</td>
<td>0.32</td>
<td>0.17</td>
<td>0.05</td>
</tr>
<tr>
<td>DXS1204</td>
<td>0.60</td>
<td>0.54</td>
<td>0.47</td>
<td>0.32</td>
<td>0.17</td>
<td>0.05</td>
</tr>
</tbody>
</table>

(3) XLRP89

<table>
<thead>
<tr>
<th>Marker</th>
<th>θ=0</th>
<th>0.05</th>
<th>0.10</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>DXS1110</td>
<td>-∞</td>
<td>-1.00</td>
<td>-0.70</td>
<td>-0.40</td>
<td>-0.22</td>
<td>-0.10</td>
</tr>
<tr>
<td>DXS574</td>
<td>-∞</td>
<td>-1.72</td>
<td>-1.14</td>
<td>-0.59</td>
<td>-0.30</td>
<td>-0.12</td>
</tr>
<tr>
<td>DXS228</td>
<td>-∞</td>
<td>-0.74</td>
<td>-0.48</td>
<td>-0.26</td>
<td>-0.16</td>
<td>-0.08</td>
</tr>
<tr>
<td>DXS7</td>
<td>-∞</td>
<td>-0.74</td>
<td>-0.48</td>
<td>-0.26</td>
<td>-0.16</td>
<td>-0.08</td>
</tr>
<tr>
<td>MAOA</td>
<td>-∞</td>
<td>-0.74</td>
<td>-0.48</td>
<td>-0.26</td>
<td>-0.16</td>
<td>-0.08</td>
</tr>
<tr>
<td>DXS8026</td>
<td>0.60</td>
<td>0.54</td>
<td>0.47</td>
<td>0.34</td>
<td>0.21</td>
<td>0.10</td>
</tr>
<tr>
<td>DXS1003</td>
<td>0.60</td>
<td>0.54</td>
<td>0.47</td>
<td>0.34</td>
<td>0.21</td>
<td>0.10</td>
</tr>
<tr>
<td>DXS573</td>
<td>0.60</td>
<td>0.54</td>
<td>0.47</td>
<td>0.34</td>
<td>0.21</td>
<td>0.10</td>
</tr>
</tbody>
</table>
### 3. Characterisation of XLRP pedigrees using haplotype analysis

#### (4) XLRP91

<table>
<thead>
<tr>
<th>Marker</th>
<th>θ=0</th>
<th>0.05</th>
<th>0.10</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYBB</td>
<td>0.81</td>
<td>0.72</td>
<td>0.63</td>
<td>0.46</td>
<td>0.30</td>
<td>0.14</td>
</tr>
<tr>
<td>DXS1110</td>
<td>0.94</td>
<td>0.84</td>
<td>0.74</td>
<td>0.53</td>
<td>0.34</td>
<td>0.16</td>
</tr>
<tr>
<td>DXS977</td>
<td>0.81</td>
<td>0.72</td>
<td>0.63</td>
<td>0.46</td>
<td>0.30</td>
<td>0.14</td>
</tr>
<tr>
<td>DXS228</td>
<td>0.81</td>
<td>0.72</td>
<td>0.63</td>
<td>0.46</td>
<td>0.30</td>
<td>0.14</td>
</tr>
<tr>
<td>MAOA</td>
<td>0.81</td>
<td>0.72</td>
<td>0.63</td>
<td>0.46</td>
<td>0.30</td>
<td>0.14</td>
</tr>
<tr>
<td>DXS1003</td>
<td>-∞</td>
<td>-0.43</td>
<td>-0.20</td>
<td>-0.05</td>
<td>-0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>DXS426</td>
<td>-∞</td>
<td>-0.43</td>
<td>-0.20</td>
<td>-0.05</td>
<td>-0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>DXS573</td>
<td>-∞</td>
<td>-0.55</td>
<td>-0.30</td>
<td>-0.11</td>
<td>-0.04</td>
<td>0.00</td>
</tr>
</tbody>
</table>

#### (5) XLRP92

<table>
<thead>
<tr>
<th>Marker</th>
<th>θ=0</th>
<th>0.05</th>
<th>0.10</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>DXS572</td>
<td>-∞</td>
<td>-0.72</td>
<td>-0.44</td>
<td>-0.20</td>
<td>-0.08</td>
<td>-0.02</td>
</tr>
<tr>
<td>DXS556</td>
<td>0.30</td>
<td>0.26</td>
<td>0.22</td>
<td>0.13</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td>DXS977</td>
<td>0.30</td>
<td>0.26</td>
<td>0.22</td>
<td>0.13</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td>DXS228</td>
<td>0.30</td>
<td>0.26</td>
<td>0.22</td>
<td>0.13</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td>MAOA</td>
<td>0.30</td>
<td>0.26</td>
<td>0.22</td>
<td>0.13</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td>DXS426</td>
<td>0.30</td>
<td>0.26</td>
<td>0.22</td>
<td>0.13</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td>DXS573</td>
<td>0.30</td>
<td>0.26</td>
<td>0.22</td>
<td>0.13</td>
<td>0.06</td>
<td>0.02</td>
</tr>
</tbody>
</table>

#### (6) XLRP94

<table>
<thead>
<tr>
<th>Marker</th>
<th>θ=0</th>
<th>0.05</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>DXS989</td>
<td>-∞</td>
<td>-2.16</td>
<td>-1.33</td>
<td>-0.58</td>
<td>-0.23</td>
<td>-0.05</td>
</tr>
<tr>
<td>DXS1242</td>
<td>-∞</td>
<td>0.67</td>
<td>0.83</td>
<td>0.83</td>
<td>0.66</td>
<td>0.38</td>
</tr>
<tr>
<td>DXS1110</td>
<td>-∞</td>
<td>-0.72</td>
<td>-0.44</td>
<td>-0.19</td>
<td>-0.08</td>
<td>-0.02</td>
</tr>
<tr>
<td>DXS6679</td>
<td>0.60</td>
<td>0.56</td>
<td>0.51</td>
<td>0.41</td>
<td>0.29</td>
<td>0.16</td>
</tr>
<tr>
<td>DXS1068</td>
<td><strong>2.11</strong></td>
<td>1.95</td>
<td>1.79</td>
<td>1.43</td>
<td>1.02</td>
<td>0.55</td>
</tr>
<tr>
<td>DXS977</td>
<td><strong>2.11</strong></td>
<td>1.95</td>
<td>1.79</td>
<td>1.43</td>
<td>1.02</td>
<td>0.55</td>
</tr>
<tr>
<td>MAOA</td>
<td>1.81</td>
<td>1.65</td>
<td>1.49</td>
<td>1.13</td>
<td>0.72</td>
<td>0.28</td>
</tr>
<tr>
<td>SYN1</td>
<td>1.81</td>
<td>1.65</td>
<td>1.49</td>
<td>1.13</td>
<td>0.72</td>
<td>0.28</td>
</tr>
</tbody>
</table>
Characterisation of XLRP pedigrees using haplotype analysis

The lod scores obtained in most of the families studied showed statistically insignificant results in families RP88, RP89, RP91, RP92 and RP88/10 of zero and negative values. This is due to the small pedigree size, and lack of meioses in these families. Families XLRP85 show slightly higher lod scores of 1.03 (θ=0) at markers DXS7 and MAOA, XLRP84 of 2.11 (θ=0) at markers DXS1068 and DXS977 and XLRP88/75 lod scores of 2.11 (θ=0) for marker DXS1003; 1.81 (θ=0) at markers DXS8080 and DXS1055. It is of interest to note that marker DXS1003 which has been found to be closely linked to the RP2 gene from haplotype analysis studies gives a lod score of 2.11 at θ=0. However this lod score could be due to the informativeness of this marker in this family. In order to establish linkage of the disease phenotype in these families more informative individuals need to be ascertained so that the number of meiotic events could be increased.
3. Characterisation of XLRP pedigrees using haplotype analysis

3.3A.2 Founder effect analysis of eight European XLRP families

Data associating particular alleles of marker loci with disease phenotype can be used to construct haplotypes of genomic regions inclusive of a disease locus. The similarity of such chromosomal haplotypes reflecting the ancestor disease alleles and detectable by genotyping a set of markers surrounding the disease locus provides the possibility of identifying disease loci by simply searching for shared haplotypes in affected individuals. This has led to the identification of founder effects, with the detection of linkage disequilibrium (LD) being used to identify genes/regions in other diseases and/or to narrow down the location of other disease genes (Snarey et al., 1994; Votruba et al., 1998).

Eight different European families with X-linked retinitis pigmentosa (Table 3.16), unrelated on the basis of genealogy were used in the analysis to establish any relationship with the identification of a shared haplotype. Extended haplotypes were constructed manually where affected individuals were genotyped by PCR using 16 highly polymorphic CA(n) repeat microsatellites (Table 3.16) spanning a region of ~26 cM covering the XLRP3 and XLRP2 loci (Figure 3.1). One affected male bearing the full disease haplotype (section 3.3) from each of the eight families was selected for analysis, except in family RP88/10 where two carrier females were included since no affected males were available in this family. All families were informative for all marker loci tested. Haplotypes segregating with the XLRP disease were constructed for all eight families to identify a possible ancestral haplotype.

Table 3.16 X-linked retinitis pigmentosa-associated haplotypes in 8 European pedigrees

<table>
<thead>
<tr>
<th>Pedigree Marker loci</th>
<th>RP85</th>
<th>RP88</th>
<th>RP89</th>
<th>RP91</th>
<th>RP92</th>
<th>RP94</th>
<th>RP88/75</th>
<th>RP88/10</th>
</tr>
</thead>
<tbody>
<tr>
<td>DXS1242</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>DXS1110</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>DXS8439</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>DXS1068</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>DXS556</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>DXS574</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>DXS228</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>MAOA</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>DXS8080</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>DXS1055</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>SYNi</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>DXS426</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>DXS1126</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>DXS573</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>DXS1039</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>DXS1204</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>
Table 3.16 indicates that the four markers (DXS1242, DXS1110, DXS8439 and DXS1068) spanning 2.8 cM (Figure 3.1) covering the RP3 disease-associated region showed no identification of a shared haplotype among any of the eight families. No common haplotype was identified around the RP2 loci between MAOA and DXS426 for the five markers (MAOA, DXS8080, DXS1055, SYN1 and DXS426) spanning a distance of 10 cM (Figure 3.1). Markers (DXS556, DXS574 & DXS228) between the RP3 and RP2 disease loci and markers distal (DXS1126, DXS573, DXS1039 & DXS1204) to the RP2 disease-associated loci also revealed no shared genotype over these genetic intervals in any of the eight families studied. Hence no specific allelic association or founder effect(s) was found from Xp21.1 to Xp11.22 for markers DXS1242 to DXS1204. Through this analysis no other independent mutations could be determined at the RP2 or RP3 loci for XLRP disease in these eight families, although these could be determined on mutational screening once the unknown gene(s) responsible for XLRP disease has been established in these families.
3.4 Discussion

X-linked retinitis pigmentosa is genetically heterogeneous with at least two loci, RP2 and RP3 on the short arm of the X-chromosome. In this study haplotype analysis was performed in twelve XLRP families with between 8-19 polymorphic microsatellite markers spanning the RP2 and RP3 critical regions. Haplotype analysis of these families has enabled classification of the the families studied as RP2 or RP3 in some cases, and in some of the RP2 families new distal boundaries for the RP2 critical region have been defined. The ability to distinguish between RP2 and RP3 is dependent upon the detection of crossovers dissecting the target region by genetic linkage/haplotype analysis due to the lack of reliable clinical differences between the 2 disease entities (Flaxell et al., 1997; Wright et al., 1991).

The prevalence of the RP2 and RP3 forms of disease within XLRP families has been reported by different groups. In the multicentre analysis of Ott et al., (1990), 75% of families were RP3 type with confidence limits extending from 45% to 90%. The highest proportion of RP2 (36%) was found in 21 XLRP families studied by Wright et al., (1987, 1991) and Aldred et al., 1994a, whereas Musarella et al., (1988 & 1990) found no evidence of the RP2 locus amongst their 20 families. From the study presented here, 40% of XLRP families segregated for the RP2 disease locus and 60% for the RP3 disease locus. The proportion of RP2 families found was similar to that from the Scottish studies (Wright et al., 1991, Aldred et al., 1994a), but the proportion of RP3 families was found to be less than that reported by Ott et al., 1990. This is probably not significant considering the smaller number of 12 families studied here. About forty percent of XLRP families were found to be segregating either for the RP2 or RP3 loci from this study. Due to lack of essential crossover data the remaining 60% could not be defined as either RP2 or RP3. It is not possible to compare the proportion of RP2 in Britain and elsewhere using this study due to the small number of non-British XLRP families, though so far no evidence has been found that RP2 is more common in Britain than the rest of Europe and the world (Aldred et al., 1994a).

In this study two RP2 families XLRP 89 (Figure 3.5, section 3.3.4) and XLRP 88/75 (Figure 3.13, section 3.3.12) harboured key crossovers which enabled refinement of the distal boundary, localising the RP2 gene proximal to MAOA (Xp11.3). This further confirms previous linkage data of the RP2 gene to between markers DXS7 and DXS426 (Bhattacharya et al., 1984, Wright et al., 1987, Coleman et al., 1990 and Ott et al., 1990). For families RP89 and RP88/75 this boundary can be further refined by genotyping with new microsatellite markers in the region between MAOA and DXS8026 and between DXS228 and DXS8080 respectively. The proximal boundary could not be refined further in these RP2 families due to a lack of crossovers at the region of DXS225 (Meitinger et al.,...
3. Characterisation of XLRP pedigrees using haplotype analysis

1989 & Fraser et al., 1989) and distal to it at DXS573. It is not necessary to test any further proximal markers below DXS255 since the proximal flanking marker for RP2 is DXS6616 (Xp11.23) (Thiselton et al., 1996), which lies ~2.4cM distal to DXS255. Hence the XLRP2 critical region could not be narrowed further from this interval and remained at 13cM.

Families RP1120 (Figure 3.11, section 3.3.10) and RP92 (Figure 3.8, section 3.3.7) were clearly found to be segregating the RP3 disease locus and not the RP2, RP6 or RP15 disease loci. Since the RP3 disease-causative gene has already been found (Meindl et al., 1996), no further refinement of the RP3 disease loci is necessary using new microsatellite markers. However the affected individuals in these families can be screened for mutations in the RPGR gene.

The family RP93 (Figure 3.9, section 3.3.8) based on an affected recombinant was clearly found not to be segregating the RP2 disease locus, but could however be segregating the RP3 disease gene. The possibility of the RP6 or RP15 disease loci being responsible was initially considered based on haplotype analysis. However this was excluded on mutation screening by the Edinburgh group for the RPGR gene. A nucleotide change in exon 6 of the RPGR gene (Meindl et al., 1996) was found for affected male IV-5 in family RP93 (A.Wright pers.comm.). This mutation positively confirms this family as an RP3 family.

Haplotype analysis of XLRP 88/10 (Figure 3.12, section 3.3.11) was able to clearly exclude the RP2 disease loci. However crossover data from a normal male could classify this family as segregating for either the RP3, RP6 or RP15 disease loci. The RP6 and RP15 disease loci could be excluded by genotyping for more markers distal to DXS572. With mutation screening for the RPGR gene it may be possible to either exclude or include RP3 as the disease-causative locus.

Six other XLRP families studied, RP85, RP87, RP88, RP90, RP91 and RP94 could not be positively categorised as to the segregation for a particular XLRP locus. It would be useful to extend these families and genotype more affected and normal males, in order to detect crossover or recombination events which may define them as either RP2 or RP3 and even narrow the region for the RP2 disease locus. By screening affected individuals for mutations in the RPGR gene, the RP3 gene may be either excluded or included as disease-causative, though not all affected members of RP3 families were found to have mutations in the RPGR gene (Meindl et al., 1996).

Further genetic analysis of XLRP families has enabled two other XLRP loci, RP6 and RP15 which are located more distally on Xp to be described (Musarella et al., 1990, Ott et al., 1990, Teague et al., 1994, McGuire et al., 1995). The presence of further XLRP loci can be implied since the disease in some families maps to none of these locations (Aldred et al., 1994). For 6 of the 7 uncategorised families the RP6 and RP15 disease loci could not be excluded, hence genotyping more markers distal to DXS572 may define these families as either RP6 or RP15. However, the probability of these families being
RP6 or RP15 is small as to date the existence of the RP6 locus is only statistical (Ott et al., 1990). Although Musarella et al., (1988) had initially proposed a RP6 locus between DXS28 and DMD, only weak evidence for this locus was found in later studies on 20 XLRP families (Musarella et al., 1990). The RP15 locus has been established in only one family, which has been reported as a cone-rod degeneration (McGuire et al., 1995). Hence the majority of these XLRP families will fall into the categories of RP2 or RP3.

In a recent paper (Thiselton et al., 1996) by the XLRP group in this laboratory, two non-British XLRP families have enabled us to define new proximal and distal boundaries for the RP2 gene critical interval and significantly reduce its size. The new flanking markers in American family NRP are MAOB (Xp11.3) and DXS6616 (Xp11.23) while those in family F72 are DXS8083 (Xp11.3-11.23) and DXS6941 (Xp11.23). Taken together, assuming the defective gene is the same in both families, this data suggests a refined RP2 interval flanked by DXS8083 and DXS6616 on proximal Xp, spanning approximately 4-5cM.

Due to a lack of recombinants, small families and classification difficulties, the resolution of the genetic localisation is unlikely to improve unless deletions in patients or cytogenetic abnormalities are identified. Hence further mapping studies will increasingly rely on physical mapping techniques, such as the development of a YAC contig spanning the critical region which is presently being undertaken by this laboratory. However detailed physical mapping of a 5 Mb region is a major task. An alternative approach is to identify expressed sequences from the region of interest as potential candidate genes for the disorder (eg. using cDNA selection method, Chapter 4).

In some of the XLRP families studied (eg. XLRP 88, 91, 93 and 94) some heterozygote carrier females were ascertained to be clinically asymptomatic though they had inherited the affected haplotype. These were probably carriers affected by the random nature of X-inactivation with absence of phenotypic expression of the RP disease but with disease associated genotype (sections 1.7.4.1 & 3.1.3). Some other carrier females though were found to be clinically symptomatic with the affected genotype eg. in families XLRP 90, 88/10 and 88/75. Lyon's hypothesis supports this great variability in functional loss from one heterozygote to another. Random inactivation of one or other of the two X-chromosomes in each cell during early embryogenesis causes the phenotype to range from asymptomatic to severely affected, depending on the proportion of retinal cells with an active disease bearing X-chromosome.

An emerging pattern in the study of molecular genetics of retinal degenerations is the recognition of genetic and clinical heterogeneity in these diseases. Recent discoveries in this area have revealed previously unrecognised associations between different clinical entities that share common gene mutations (gene sharing), as well as distinctly different molecular alterations within the spectrum of what traditionally was believed to be the same disease (locus heterogeneity). This genetic phenomenon also
3. Characterisation of XLRP pedigrees using haplotype analysis

Known as non-allelic heterogeneity reflects the limited repertoire of responses of the retina to a variety of genetic lesions.

Overlapping map locations in Xp11.3-Xp11.22 have been found for other inherited X-linked retinal disorders such as Congenital Stationary Night Blindness X (CSNBX; Aldred et al., 1992; Bech Hansen et al., 1992; Berger et al., 1995), Aland Island Eye Disease (AIED; Alitalo et al., 1991; Schwartz et al., 1991; Glass et al., 1993) and X-linked progressive cone dystrophy (Bergen et al., 1993; Meire et al., 1994; Hong et al., 1994). Two forms of autosomal CSNB have been shown to be allelic to RP: both CSNB and RP can result from mutations in the rhodopsin gene (Dryja et al., 1993, Rao et al., 1994) and the gene encoding the β-subunit of the rod cGMP phosphodiesterase (Gal et al., 1994). One could then speculate that the above diseases may be allelic to XLRP2. If so, the refinement of the RP2 critical region may also contribute significantly towards the localisation of these diseases.

Genotype-phenotype correlation studies in XLRP families show evidence of clinical heterogeneity. A tapetal reflex, when reliably observed in XLRP carriers, seems to correlate with an RP3 localisation (Nussbaum et al., 1985; Denton et al., 1988; Curtis and Blank, 1989; Musarella et al., 1989), but the absence of a tapetal reflex does not exclude this locus. The presence of moderate to high myopia was found to be a characteristic feature of the RP2-type disease in the affected males of some families (Wright et al., 1991). These observations were also confirmed by Kaplan et al., (1992), who found that affected males in families showing night blindness as the initial symptom at about age 10 years appeared to show linkage to the RP3 region, while families with myopia presenting at the age of 3 years appeared to show linkage to the RP2 region.

These results were contradictory to those of Flaxell et al., 1997 who examined affected males and obligate carrier females from 17 XLRP families assigned as either RP2 or RP3 by haplotype analysis (some families from this study) and/or by heterogeneity analysis. On examination no clear phenotypic differences were found between RP2 and RP3 with respect to myopia, onset of night blindness and the presence of primary cone dysfunction. The tapetal reflex was found to exist in carriers of both RP2 and RP3. Narrowing of the RP2 region by haplotype analysis studies will improve reliability of carrier detection and may facilitate characterisation of the gene and its mutations and will help resolve the issue of clinical and genetic heterogeneity.
3.5 Conclusion

Thirteen new XLRP families were analysed using twenty eight microsatellites spanning Xp21.1-Xp11.22. From this haplotype analysis, three families were found to be RP3, two families RP2, one choroideremia and seven could not be unambiguously assigned. The RP2 families show recombination events below MAOA, further localising the RP2 gene proximal to MAOA and confirming the previous linkage data. The distal boundary of the XLRP2 interval could not be further refined from the RP2 families analysed.

The four different and new mutations in the RPGR gene for four of the XLRP3 families described in section 3.3A suggest a high new mutation rate for the RPGR gene and no founder effect which has been confirmed by the founder effect analysis (section 3.3A.2) at the RP3 loci. This supports the assertion of Fujita et al., (1996) that most RPGR mutations are of independent origin. The statistically insignificant lod scores obtained for the eight families from linkage analysis could be improved by collecting and analysing further informative individuals in order to give more conclusive results. From both the linkage and founder effect analysis no other independent mutations could be determined at the RP2 and RP3 loci in the eight European families.
CHAPTER FOUR

cDNA selection for enriched transcripts

from genomic clones
CHAPTER FOUR

cDNA SELECTION FOR ENRICHED TRANSCRIPTS FROM GENOMIC CLONES

4.1 Introduction

To combine the information on the genetic and physical maps of a genomic region with the localisation and characterisation of genes located in the area, transcription maps have to be constructed. The isolation and analysis of genes from a large genomic region, is an inherent part of positional cloning strategies for the discovery of genes responsible for hereditary diseases such as XLRP2. In disorders with no gross rearrangements of the genetic material (eg. XLRP2), which would allow very precise localisation of the gene, generally all genes from the large candidate genomic region must be isolated and analysed for their causative role in the particular disorder. It will often be possible to compare information on the transcripts with those expected from the features of the disease and to identify appropriate candidate genes. In addition to their application in the search for disease genes, the transcription maps will provide an important insight into the organisation of the human genome, the distribution of genes, their order and its possible functional relevance. Since transcripts are often evolutionarily conserved, they will also serve as anchor points in comparative mapping efforts.

Generally transcription maps can either be constructed by a global approach, via the identification and localisation of most or all transcripts from a given tissue, or by a regional approach, in which the genomic DNA from a particular region is used to identify the genes from a region by a variety of techniques. These techniques include the analysis of evolutionary conservation, CpG islands, exon trapping, direct screening and cDNA selection (section 1.5). As a part of the analysis of the human XLRP2 critical region at Xp11.3-11.22, the cDNA selection method (Tagle et al., 1993) was successfully established as described in this chapter. This then offered a highly efficient route to the isolation of expressed sequences, and was used to isolate a number of cDNA clones from a 500 kb region within the XLRP2 locus. Development of a transcriptional map is equally applicable to any other portion of the XLRP2 locus and will greatly assist in the search for this disease gene and other candidate retinal genes. Futhermore, once the disease gene is cloned, the development of a detailed transcription map of a particular region
will allow further assessment of the possible regulatory relationships between genes in
this region.

4.1.1 General description of the cDNA selection method.

The cDNA selection method described here is a modification of the cDNA
enrichment strategy (section 1.5.6) which involves hybridising biotinylated genomic
dNA to cDNA probes in solution and subsequently capturing the hybridised complexes
using streptavidin-coated magnetic beads.

The general scheme for magnetic bead capture of expressed cDNAs from YACs is
shown in Figure 4.1. Pools of pulsed-field gel purified YAC genomic DNA or purified
cosmid DNA are digested with four-base cutters and ligated to linkers. In place of
restriction enzyme digestion, genomic DNA can also be sonicated and blunt-ended. The
linkered genomic segments are PCR-amplified using a 5'-biotinylated primer whose
sequence matches one of the linker arms. cDNA library inserts are PCR-amplified using
vector primers and hybridised at high stringency to the biotinylated genomic fragment
after blocking of repetitive sequences. The biotinylated genomic-cDNA complexes are
captured using streptavidin-coated paramagnetic beads and the unbound non-specific
cDNAs are washed off. The captured cDNAs are eluted and PCR-amplified. To increase
selectivity, the primary-selected cDNAs are cycled through a second hybridisation
selection step to yield secondary selected cDNAs. The PCR products are digested with a
suitable enzyme and subcloned as a region-specific sublibrary. The analysis of region-
specific cDNA sublibraries represents a simple, rapid and efficient tool for the generation
of a regional transcription map.

4.1.2 Key aspects of the methodology

Magnetic bead capture utilises biotin-streptavidin magnetic bead technology to
isolate cDNAs rapidly from large genomic intervals, giving several thousand-fold
enrichment of the selected cDNAs. The technique can allow parallel analysis of several
large genomic segments of varying complexities and can be applied to the isolation of
expressed sequences from various tissue sources. cDNAs or cDNA libraries from
multiple tissues can easily be mixed or used in parallel, thus increasing the probability of
detection of tissue-specific transcripts.

As both genomic clones and cDNA libraries are amplified by the polymerase
chain reaction (PCR), the method provides a substantial advantage in that only a small
quantity of starting material is required. Furthermore, no specific demands such as
evolutionary conservation, presence of a CpG island or favourable distribution of
exon/intron boundaries need to be fulfilled to allow the isolation of the gene. The
4. cDNA selection for enriched transcripts from genomic clones

**Figure 4.1** General scheme describing the direct cDNA selection method for magnetic bead capture of expressed cDNAs from YACs.
4. cDNA selection for enriched transcripts from genomic clones

analysis of the sub-library is very rapid as many identical sublibrary filters can be spotted and hybridised in parallel with many probes and the resulting short and medium size cDNA clones are good substrates for DNA sequencing.

4.1.3 Technical considerations

The modification of biotinylating genomic DNA coupled with very high affinity binding with streptavidin coated paramagnetic beads provides finer control of solution hybridisation conditions as compared to the kinetics of filter hybridisation. The strength and stability of the biotin-streptavidin coupling allows DNA manipulations, such as thermal denaturation and elution of annealed cDNAs or relative ease in changing buffers and wash solutions. This ensures a flexible system where pre-blocking, hybridisation and washing can be performed easily. In addition, the beads are of equal size (monodispersed) and thus they follow uniform kinetics when subjected to a magnetic field.

The genomic DNA used for capturing cDNAs can be from any genomic source. However enrichment with cosmids DNA seems to work more efficiently than with YACs as during the gel-purification of the latter, YACs often co-migrate with degraded, higher molecular weight yeast chromosomes that harbour tandem ribosomal sequences. This is especially so for yeast chromosome XII, which carries >100 copies of the yeast ribosomal genes. This has led to interesting enrichment artefacts (Lovett et al., 1991), where 30-60% of the clones from the enriched sublibrary from YACs contain ribosomal sequences. This is most likely due to cross-species hybridisation of highly conserved ribosomal sequences between human and yeast and not from yeast contamination of the cDNA libraries, as the same phenomenon is observed when cDNA is extracted and used directly from cell lines (Morgan et al., 1992). Fortunately, these sequences are easily blocked and/or screened out after selection.

A significant number of cDNA libraries contain repetitive sequences (Adams et al., 1992) in the form of expressed repeats or unspliced messages and some cDNA clones contain repeats, particularly in their 3' untranslated region. Artefactually enriching these repeat-containing cDNAs can be partly avoided by suppressing repetitive sequences in the genomic target DNA with sheared human placental DNA. Intermediate repetitive sequences, such as Alu repeats, can be blocked efficiently and are not usually a major problem. However, when an Alu-containing cDNA is isolated, it can be difficult to ascertain whether it is a spurious contaminant or was selected by hybridisation to a stretch of single-copy sequence. Randomly primed cDNA libraries can help to distinguish these possibilities, since they will usually contain another part of the cDNA that does not include a repeat. Low-copy repeats are never adequately quenched, and can contribute up to 10 per cent of the enriched sublibrary. This is especially
important as some chromosomal regions have relatively higher frequency of low-copy repeat sequences than others. Such repeat sequences are sometimes difficult to detect by standard methods (such as hybridisation of a cDNA to a contig of cosmids), as they can map very specifically. A whole-genome Southern blot hybridised with such a cDNA probe can resolve this difficulty.

4.1.4 Potential limitations and possible solutions.

Several potential limitations, as well as possible solutions are associated with the direct selection method. First, the resulting small insert sub-library necessitates screening a full-length cDNA library to obtain full-length transcripts. The time involved in re-screening, however, can be cut substantially by screening random-primed cDNA libraries. Alternatively, the insert size of the enriched cDNAs can be preserved by direct cloning of the PCR products.

Second, genomic clones that contain region-specific low-copy repeats can cause a significant number of enriched cDNAs to contain these repeat sequences. Once identified, however, a running catalogue of these repeats can be used to pre-screen the sub-library or can be incorporated as blocking agents in future magnetic bead capture experiments. Likewise artefactually selected ribosomal cDNA clones can be prescreened with ribosomal DNA probes.

Third, since the methodology is based on hybridisation by sequence homology, pseudogenes and members of multigene families may also be enriched, in contrast to the CpG island approach (non transcribed pseudogenes loose their CpG islands) or exon trapping approach (processed pseudogenes are not isolated). However, these can be sorted through by further mapping or sequencing of the enriched cDNAs.

Lastly, in order to capture all encoded cDNAs from a given region, it would be ideal to use a cDNA library representing all possible transcribed human sequences. In this regard, it is possible to use pooled and normalised libraries from multiple tissue sources and from various developmental stages to provide an equivalent complete cDNA library.

4.1.5 Application of the cDNA selection technique to the RP2 region.

Systematic and reliable identification of coding regions within extensive genomic regions is difficult as genes are irregularly dispersed and may contain many exons. Selective expression of a particular gene with respect to type of tissue or stage of development may also complicate retrieval of cDNAs.
The magnetic bead capture of cDNAs as described above, provides an effective and straightforward technique for isolation of expressed sequences from large genomic tracts (section 1.5.6), such as those cloned into a contig of YACs, cosmids and/or phage clones. This method is especially useful for isolating candidate genes where the defined candidate interval has been delineated by linkage analysis. This method was particularly suitable for the RP2 region, which spanned a genetic distance of approximately 11 cM at the start of this research (Schlessinger et al., 1993), so one does not necessarily want to isolate all the genes in this interval but to restrict the search to sequences expressed in the retina. Additionally the selective power of this method enables enrichment of rare transcripts in specific tissues, such as the retina.

Support for the use of this technique was also derived from reports on successful applications in the isolation of genes involved in X-linked agammaglobulinemia (Vetrie et al., 1993) the cri-du-chat 2 Mb critical region (Simmons et al., 1995), and cDNAs encoded within a 1.2 Mb area of the Down syndrome critical region on chromosome 21 (Peterson et al., 1994), the 1 Mb region flanking BRCA1 on human chromosome 17q21 (Osborne-Lawrence et al., 1995) and a 1 Mb area within the Huntington disease gene region (Rommens et al., 1993). In comparison with other gene-isolation procedures cDNA selection proved to be the more effective method over direct cDNA library screening and genomic sequencing used in the search for candidate genes over a 600 kb genomic region (Harshman et al., 1995).

4.1.6 Elements of the YAC contig used

At the time of this research a 4 Mb YAC based physical contig of the RP2 critical region (Xp11.3-Xp11.22) was under construction in our laboratory (Thiselton et al., 1995a). Heterogeneity analysis suggested the RP2 gene probably lies within 2cM distal to DXS426 (Teague et al., 1994), and this overlapping contig spanned the most likely location for the RP2 gene. On this basis, in order to initiate a transcriptional map in the proximity of DXS426, four well characterised YACs from the contig around this locus, covering approximately 500kb (Figure 4.2), were utilised.

YAC 33CA11 of 290 Kb is non-chimeric and contains the ELK1 and properdin genes.

YAC 4HG2 of 250 Kb is non-chimeric and contains the properdin (PFC) gene.

YAC A1220 of ~390 Kb is chimeric and contains both the TIMP-1 and properdin genes.

YAC C1228 of 320 Kb is chimeric and contains both the TIMP-1 and properdin genes.

YACs 33CA11 and 4HG2 were found to be non-chimeric based on end-clone analysis and YACs A1220 and C1228 found to be chimeric based on end-clone analysis and FISH analysis. The sizes of the various YACs were determined by pulse field gel electrophoretic separation and Southern blot hybridisation analysis as described in section 4.2.2.1.
4. cDNA selection for enriched transcripts from genomic clones

Figure 4.2: Section of YAC Contig on Xp11.23 showing the four overlapping YACs used for cDNA selection with the additional C1022 YAC used to test YAC end clones for chimerism

- end clone of chimeric YAC
- end clone of non-chimeric YAC

4.1.6.1 Positive controls used in selection procedure

The positive control should be a known gene, preferably lying within the target contig, whose corresponding cDNA is present in the pool of starting cDNAs. The known gene may then be used as a positive control for specific enrichment of the control cDNA. If the genomic contig does not contain a known gene, or if the starting cDNA does not contain the corresponding cDNA, then the cDNA can be spiked with a known sequence, or the genomic DNA spiked with a sequence that is represented in the starting cDNA source. Finding a negative control is much easier, a good example is a house-keeping gene that lies outside the contig but is represented in the starting cDNAs. Positive and negative controls were always used in published successful cDNA selection procedures (Lovett et al., 1991; Morgan et al., 1992).

The abundantly expressed retinal gene rhodopsin was used to isolate a YAC of 260 Kb for use as a good external positive control for the technique. The A1220 and C1228 YACs act as good internal positive controls for the region, since the retinal cDNA libraries used were found to contain both properdin (PFC) and TIMP-1 genes (section 4.2.1.2) which map to these YACs. If the selection procedure works well, the retinally
expressed TIMP-1, PFC and rhodopsin genes found in the respective YACs should be
enriched and selected for. The presence of good positive control reporter genes, both
internal (TIMP-1 and PFC; rarely expressed) and external (rhodopsin, abundantly
expressed), is essential to monitor enrichment and efficiency. The negative control gene
in this study was β-actin, which does not lie in the region studied. It is a house-keeping
gene, and is represented in both the retinal cDNA libraries used.

4.1.7 cDNA libraries used

The two adult retinal cDNA libraries used were the λZAPII adult retina cDNA
library (Stratagene, STRATAGENE Ltd., Cambridge, UK) and an adult retina cDNA
library in λgt10 (gift from J. Nathans). Two important limitations are to be understood
when using cDNA libraries. Firstly, the relative abundance of the transcript of interest.
Highly abundant transcripts may constitute 10% of the total mRNA. However, low-
abundance transcripts may be as low as 1 in 10^6. Since the cDNA libraries consist of ~10^6
recombinant clones; this was adequate sequence representation for transcripts expressed
in one tissue type, such as the retina. However, if the cDNA library is derived from a
complex tissue such as total brain, and if the transcript of interest is expressed at a low
level in a small subregion of the brain, then 10^6 clones will probably not contain the
cDNA of interest. Therefore, it is important to sample a sufficiently large set of cDNAs
from a complex tissue.

Another limitation of most cDNA libraries is that they are constructed to yield
full-length or near-full-length copies of the starting mRNA. When these sets of cDNAs
are subjected to multiple PCR amplification cycles, the larger cDNAs are selected against
and the selection is biased toward identifying shorter cDNAs. There are two ways to
avoid this bias. One way is to cut the cDNAs down to a shorter average size (by random
shearing or restriction enzyme digestion), the second is by constructing the primary
cDNA by random priming, so as to build a collection of overlapping molecules of
various sizes.

The isolation of the 3' end of cDNA transcripts is achieved more easily from an
oligo-dT primed cDNA library than a random primed. Several factors might influence
the abundance of inserts in an oligo(dT)-primed library that are complementary to any
short-fragment cDNA insert from a random-primed library. If a fragment is derived
from the 5' end of an mRNA, especially a long mRNA, one with a high GC content, or
one with an unusually stable hairpin structure (Choong et al., 1990), complementary
DNA sequences might be difficult or impossible to obtain with an oligo(dT)-primer.

The ideal initial cDNA library, should be random primed, short insert and
normalised. Among these three factors, the random priming and size selection in favour
of short inserts (200-600 bp) seem to be of major importance (Gecz et al., 1993). This is

211
because a short-fragment random primed library is least affected by any bias introduced by PCR as a result of the size of cDNAs or GC-rich regions of a particular cDNA. At least one, if not more, cDNA fragments can be expected to be recovered even in situations where GC-rich or repetitive sequences are contained within a part of the cDNA. The drawback of a short fragment cDNA library however is that one has to screen a full-length cDNA library with short fragment probes in order to isolate full length cDNAs.

High quality pools of cDNAs containing mixtures of random-primed and oligo-dT primed cDNAs from two different sources were used. These were an adult retinal cDNA library in λZAPII which is available from Stratagene. This library is oligo-dT and random primed, enabling both near full length and overlapping cDNAs to be selected for. The other adult retinal cDNA library was obtained from Jeremy Nathans. This library, in λgt10 was oligo-dT primed and was prepared in two size ranges of cDNAs, of ≥ 1 Kb and ≤ 1 Kb. Both libraries were used separately as starting cDNAs in the selection procedure. They were also assessed for the presence of the positive controls; the rhodopsin, TIMP-1 and PFC genes as discussed in section 4.2.1.2.

4.1.8 Pre-blocking agents used

As described in section 4.1.3, the presence of repetitive elements within the population of cDNAs is one of the central problems in searching for cDNAs using large genomic clones. Therefore one critical parameter is the adequate blocking of these repeats in either the cDNA inserts or the genomic DNA used before the hybridisation. Cot-1 DNA is a useful blocking reagent consisting of the repetitive elements that rapidly re-anneal within the human DNA (primarily satellite sequences and Alu repeats). However, blocking the higher frequency repeats tends to reveal other classes of lower frequency repeats such as long interspersed repeats (LINE 1) that do not get blocked by Cot-1 DNA, thereby complicating the final interpretation.

Yeast contamination in cDNA libraries can also lead to selection artifacts (section 4.1.3), particularly when total yeast DNA is used as the target DNA, rather than the isolated YAC. Even with "purified" YAC DNA, a low concentration of ribosomal contamination in the YAC selects the ribosomal sequences that are present in most cDNA libraries. Ribosomal sequences can hence comprise 70% of the selected material after one cycle of selection if no ribosomal blocking is incorporated (Tagle et al., 1993). These parameters were taken into consideration when the quenching agents were chosen. These are described below:

1. Sheared human placental DNA was used to compete against Alu repeats.
2. pL1A - LINE 1 (Long interspersed elements); a 5.8 kb Acc1 insert cloned in pGEM, which blocks against repeat sequences of 500bp to as large as 7000bp.
4. cDNA selection for enriched transcripts from genomic clones

3. pR7.3 and pR5.8 - Human ribosomal RNA 45S precursor region EcoRI fragments of 7.3 kb and 5.8 kb cloned in the vector pBR 322. These were used to compete against human ribosomal genes.

4. pK515 - A 7 kb *E. coli* ribosomal RNA operon BamHI fragment in pBR 322. This was used as the cDNA inserts for hybridisation were generated by PCR from a crude phage DNA preparation rather than from a purified phage preparation. It was therefore necessary to compete against *E. coli* ribosomal genes found to be conserved in humans.

5. RibH7 and RibH15 - Yeast ribosomal RNA clones of 2.6 kb and 7 kb HindIII fragments in pBR322.

6. Total yeast DNA - Yeast DNA prepared from the YAC host yeast strain AB 1380 was obtained commercially in solution (Gibco, BRL). It is used to quench any co-selection of cDNAs by contaminating yeast DNA in a YAC preparation, hence blocking against the selection of yeast genes conserved in humans.

4.2 Establishment and application of the cDNA selection method to YACs mapping in the critical region of the RP2 locus

4.2.1 Preparation of cDNA

cDNA inserts ready for the direct selection hybridisation procedure were prepared as described below. cDNA libraries λZAPII and λgt10 were initially amplified. The amplified libraries were then titred to determine the plaque forming units (pfu) per μl and then subsequently tested for the presence of positive control genes (section 4.1.6.1). Both amplified cDNA libraries were also independently amplified by PCR and the PCR amplified cDNA inserts were then purified and prepared ready for the direct selection hybridisation procedure.

4.2.1.1 Amplification and titration of adult retinal cDNA libraries

Since only small aliquots were available of both adult retinal cDNA libraries. They were amplified using a plate lysate method as described in section 2.4.4.3. The λZAPII adult retinal cDNA library (Stratagene) had already been amplified by a colleague in the laboratory (Hardcastle, 1993; PhD thesis). The λgt10 (Jeremy Nathans) adult retinal cDNA library, available in the two size ranges of ≥ 1kb and ≤ 1 kb was amplified for both size ranges separately.
4. cDNA selection for enriched transcripts from genomic clones

The titre of the original unamplified λgt10 adult retinal cDNA libraries was determined so that a representation of $10^5$ pfu was present per plate during the plate lysate amplification method (section 2.4.4.3). The titres of both amplified adult retinal cDNA libraries were also determined by the same serial dilution method described in section 2.4.4.2. The titres of the unamplified and amplified λgt10 adult retinal cDNA libraries were as follows:

Unamplified λgt10 adult retinal cDNA library of size $\geq 1$kb $5 \times 10^6$ pfu/µl and
\[ \leq 1$kb $4 \times 10^6$ pfu/µl \\
Amplified λgt10 adult retinal cDNA library of size $\geq 1$kb $4 \times 10^6$ pfu/µl and
\[ \leq 1$kb $9 \times 10^6$ pfu/µl \\
The titre of the λZAPII adult retinal cDNA library was $1 \times 10^5$ pfu per µl, having reduced by 10 fold during storage at 4°C over a period of three years.

4.2.1.2 Testing for known retinal genes in cDNA libraries

The presence or absence of positive control genes rhodopsin, TIMP-1 and PFC found within some of the YACs used were tested for in the λZAPII and λgt10 adult retinal cDNA libraries. Specific primers spanning exon 5 of the rhodopsin gene, and portions of the cDNAs of TIMP-1 and properdin were designed and tested on phage lysates from both retinal cDNA libraries used. Tenfold serial dilutions of bacteriophage stocks of both the λZAPII and λgt10 cDNA libraries (combined equal volumes of $\geq 1$kb and $\leq 1$kb sized cDNAs) were prepared in sterile water. 3µl of undiluted phage and 3 µl of each of the dilutions from $10^1$ to $10^6$ were used in separate PCR reactions as described in section 2.6.2. Genomic DNA and a total X-chromosome somatic cell hybrid DNA (Thy B-X) was included in the control panel. The annealing temperatures and expected sizes of PCR products of the various genes are as follows:

- Rhodopsin cDNA; Tm- 63°C and 220 bp PCR product
- TIMP-1 cDNA; Tm- 68°C and ~ 610 bp PCR product
- Properdin cDNA; Tm- 59°C and 280 bp PCR product.

A 220 bp PCR product was amplified using rhodopsin primers from both the undiluted (neat) and $10^1$ diluted PCR reactions of the λgt10 retinal cDNA library (Figure 4.3A, Lanes 1&2). Similarly a ~610 bp size fragment for TIMP-1 (Figure 4.3B, Lanes 1&2) and a 280 bp fragment for properdin (data not shown) were amplified from the neat and $10^1$ dilution λgt10 cDNA library using the respective gene specific primers. Similar results from each of the rhodopsin, TIMP-1 and properdin specific primers were obtained using the λZAPII retinal cDNA library (data not shown). Since the TIMP-1 and properdin genes are expressed in the retina, probably at a low level, they should be selected for during the procedure.
PCR amplification of rhodopsin and TIMP-1 retinal genes from the amplified λgt10 adult retinal cDNA library. Specific primers for exon 5 of rhodopsin and a portion of the TIMP-1 cDNA were used in PCR reactions on phage lysate DNA of the λgt10 cDNA library (combination of equal volumes of the ≥ 1kb and ≤ 1kb sized cDNAs). Undiluted phage stock DNA (Lanes 1) and tenfold serial diluted (10^1 to 10^9) phage DNAs (Lanes 2-7) were used in separate PCR reactions for each of the gene specific primers. Genomic DNA (Figure A, Lanes 8-10) and (Figure B, Lanes 9-11) and DNA from a total X chromosome somatic cell hybrid (Figure A, Lane 11 & Figure B, Lane 12) were included in the control panel. A negative result was seen in the no DNA track (Figure A, Lane 12 & Figure B, Lane 8) indicating absence of contaminating DNAs in the PCR reaction mix used.

Figures (A) and (B) show ethidium bromide stained gels of PCR amplified products from the test DNAs using the rhodopsin and the TIMP-1 gene specific primers respectively, fractionated on a 1.2% agarose gel with the φX174/HaeIII molecular weight marker (M.W.M) as a size comparison (Lanes 13).

(A) 220 bp PCR fragments were amplified using the rhodopsin exon 5 primers (B) ~610 bp PCR products were obtained using the TIMP-1 cDNA specific primers from both the undiluted phage (Lanes 1) and the 10^9 diluted phage (Lanes 2) DNAs for each retinal gene.

Un - Undiluted phage DNA
C 1, C 2 & C 3 - Genomic DNA (positive control)
Total X DNA - Total X chromosome somatic cell hybrid DNA
4.2.1.3 PCR amplification and preparation of cDNA inserts using vector primers of λZAPII and λgt10

Phage lysates of the λZAPII and λgt10 libraries corresponding to $2 \times 10^6$ viable clones (pfu) were independently amplified in a 100 μl PCR reaction. PCR reactions were set up as described in section 2.6.2. The DNA sequences of the primers used were as follows:

λ-ZAPII T3 primer: 5' - GCT CGA AAT TAA CGC TCA CIA AAG - 3'
T7 primer: 5'- GAA TIG TAA TAG GAG TCA GTA TAG - 3'
λgt10 F primer: 5' - TTG AGG AAG TTC AGG GTG GTT AAG - 3'
(Set A) R primer: 5' - CIT ATG AGT ATT TGT TGG AGG GTA - 3'

Thermal cycling was performed for 30 cycles, where each cycle consisted of denaturation at 94°C for 1 min, annealing at 58°C for 1 min and extension at 72°C for 1 min 30 sec plus an additional 5 seconds extension per cycle using the automatic segment extension option. The reaction products of the amplified cDNAs were analysed on a 1.2% agarose gel (section 2.1.3) using ethidium bromide visualisation.

The results are shown in Figure 4.4A, 4.4B and 4.4C. The smear of PCR products obtained from the λZAPII cDNAs (Fig. 4.4A) fall in a size range of 1.0 kbp to 230 bp. The intensity of the smear seems uniform throughout except for some intense bands at 600 bp and 300 bp (arrows). The λgt10 ≥ 1kb amplified cDNAs fall in a size range of 3 kbp to 600 bp (Fig. 4.4B); the smear of DNA being uniform but of greater intensity at 800 bp to 1 kbp. A faint smear in the size range 800 bp to 200 bp is present for the λgt10 ≤ 1kb amplified cDNAs (Fig. 4.4C) with bands in the region of 600-500bp appearing more intense.

When different sized cDNA inserts present in the starting cDNA library are amplified, smears of PCR products are produced over different size ranges for the various cDNA libraries. On amplification, the mixed random and oligo-dT primed λZAPII cDNA library produced inserts is in an intermediate size range (1 kbp - 230 bp) as was expected from a random primed library. The size selected λgt10 ≥ 1kb and ≤ 1 kb oligo-dT primed cDNA libraries amplified cDNA inserts in larger (3 kp - 600 bp) and smaller (800 bp - 200 bp) size ranges respectively. Full length cDNAs from the selected size ranges can be obtained from λgt10 oligo-dT primed cDNA libraries; this accounted for the larger sized cDNA inserts from the λgt10 ≥ 1kb library. The intense band sizes observed are probably due to the most abundantly sized cDNAs present in each library.
Figure 4.4 Unpurified cDNA inserts from PCR amplified starting cDNA libraries using λZAPII and λgt10 vector primers.

The figures (A), (B) and (C) show ethidium bromide stained gels of PCR amplified cDNA products from phage lysates corresponding to 2 × 10^6 viable clones (pfu) of the retinal cDNA libraries in λZAPII, λgt10 ≥ 1kb and λgt10 ≤ 1kb respectively. The ØX174/HaeIII and the λ/HindIII standard size molecular weight markers (M.W.M) were used as size comparisons.

In figure (A) the smear of PCR products obtained from the λZAPII cDNAs fall in a size range of 1 kbp to 230 bp. The intensity of the smear is uniform throughout except for some intense bands at 600bp and 300bp (arrows). The λgt10≥1kbp amplified cDNAs fall in a size range of 3 kb to 600bp as indicated in figure (B), the smear of DNA being uniform but of greater intensity at 800bp to 1 kbp. A faint smear of size range of 800bp to 200bp is present for the λgt10≤1kbp amplified cDNAs as shown in figure (C), bands in the region of 600-500bp appearing more intense.
4. cDNA selection for enriched transcripts from genomic clones

4.2.1.4 Wizard PCR Preps, DNA purification and precipitation of cDNA inserts

The Wizard PCR Preps System (Promega) was used in the purification of the double stranded PCR amplified cDNA inserts. 200μl of pooled amplified cDNA inserts were purified using this kit according to manufacturer's instructions. The cDNAs were recovered in 100μl sterile water. An aliquot of the purified cDNAs was run on a 1.2% agarose gel (section 2.1.3), to determine recovery and purity of DNAs. The results for the λZAPII cDNAs and the λgt10 ≥ 1kb cDNAs are shown in Figures 4.5A and 4.5B respectively. The recovery of the λgt10 ≤ 1kb purified cDNAs was not so evident on fractionation as seen on a ethidium bromide stained gel although an equal volume of product was run (hence data not shown). Since the purified λgt10 ≤ 1kb cDNAs could not be seen, the purity of the products were assumed.

The different purified cDNAs were then precipitated in 1/10th volume of 3M sodium acetate pH 6.0 and 2 X volume of absolute ethanol, washed in 70% ethanol, air-dried and the precipitate resuspended in sterile water. The concentration of the cDNAs was determined (data not shown) from fractionated products on 1.2% agarose gels (section 2.1.3) by comparing with the intensity and known concentration of the ØX174/HaeIII marker bands. The volume of the resuspended cDNAs were then adjusted to a concentration of 1μg/μl. An equal amount of the λgt10 ≥1kb and λgt10 ≤1kb precipitated cDNA inserts were pooled together to a final concentration of 1μg/μl for use in the hybridisation procedure as described in sections 4.2.4 and 4.2.5.

4.2.2 Preparation of genomic DNA

Genomic DNA from four YACs in the region of interest and the rhodopsin YAC were prepared ready for the direct selection hybridisation procedure as described below. The YACs were separated away from contaminating yeast chromosomes by pulsed field gel electrophoresis. The purified YACs were then digested with two different enzymes and oligonucleotide linkers/adaptors were ligated to each end, followed by PCR amplification employing 5' biotinylated primers matched to the linker sequences, in order to generate large amounts of DNA and to uniformly incorporate biotin into the products.

4.2.2.1 Electrophoretic separation of YACs by pulsed field gel electrophoresis

The YAC clones used (section 4.1.6) were isolated from the ICRF and ICI human YAC libraries (Larin et al., 1991) as described in section 2.5.1, using PCR-based screening. YAC clone DNA in agarose plugs was prepared according to Vollrath and Davies (1987),
4. cDNA selection for enriched transcripts from genomic clones

Figure 4.5 Purified cDNA inserts from PCR amplified λZAPII and λgt10 starting cDNAs.

Purified cDNA inserts of the λZAPII and the λgt10 ≥1kb PCR amplified cDNAs are shown on ethidium bromide stained gels in figures (A) and (B) respectively. The ØX174/HaeIII and the λ/HindIII standard size molecular weight markers (M.W.M) were used as size comparisons.

(A) The λZAPII purified cDNAs show a good recovery and purity since cDNAs below 300bp and primer dimers have been removed, in comparison to those present in the λZAPII unpurified cDNAs (Figure 4.4A).

(B) The λgt10≥1kb purified cDNAs also gave a good recovery, although evidence of purification is less easily seen since the amplified cDNA insert size range lies from 3kb to 600bp in both purified (Figure 4.5B) and unpurified cDNAs (Figure 4.4B). The elimination of excessive primers and primer dimers has been assumed.
with slight modifications as described in section 2.5.2. The YAC DNAs were fractionated on 1% low melting point (LMP) agarose gels using contour-clamped homogeneous electric field (CHEF) gel electrophoresis under conditions that separate the YACs from the rest of the endogenous yeast chromosomal bands. The experiments were conducted on the CHEF-DRII apparatus (Biorad), as described in section 2.5.3.

The following electrophoresis parameters and voltage were selected to separate DNA in the size range 450Kb-200Kb. The gel was run in 0.5X TBE at 14°C for 20 hours at 170V/cm with a switching interval of Initial A Time: 25 seconds, Final A time: 25 seconds using a LKB Pulsaphor apparatus. After the first run the gel was run for a further 20 hours at a switching interval of Initial A Time: 45 seconds, Final A Time: 45 seconds. The gel was then stained with ethidium bromide, destained and visualised under long wave UV light.

When YACs such as 33CA11 and 4HG2 co-migrated with native yeast chromosomes and were indistinguishable on the gel, they were identified by hybridisation of the Southern blot of the gel (section 2.2.1) to human placental DNA, (which when labelled can provide an effective hybridisation probe to check for the presence of human DNA) and subsequent autoradiography (section 2.3). After identification of the YAC, the band of interest was cut out with a clean scalpel and the agarose piece containing the YAC, stored in a small petri dish in TE buffer pH 8.0 at 4°C until further use.

Separated YAC clones on pulsed field gels are as seen in Figures 4.6A and 4.6B. The rhodopsin YAC (Figure 4.6A, Lane 1) and the YAC identified as C1228 (Figure 4.6A, Lane 4 & 5) of 260 Kb and 320 Kb respectively were well separated from the other yeast chromosomal bands and were excised easily from the LMP agarose gel. However, the ~390 Kb A1220 YAC clone (Figure 4.6A, Lane 2 & 3) sits just above the S. cerevisiae yeast chromosomal DNA band III of 365 Kb (Figure 4.6A, Lane 6) and was excised with greater difficulty. Two other YAC clones of 4HG2 of ~250 Kb (Figure 4.6B, Lanes 2 & 3) co-migrated with the smallest S. cerevisiae yeast chromosomal band I of 225 Kb (Fig. 4.6B, Lane 1) and 33CA11 of ~290 Kb (Figure 4.6B, Lanes 4 & 5) co-migrated with the yeast chromosomal band VI of 285 Kb (Fig. 4.6B, Lane 1). This was determined by hybridisation of the Southern blotted gel in Fig. 4.6B to human placental DNA, which showed distinct YAC bands of 250 Kb and 290 Kb in YAC 4HG2 (Figure C, Lane 3) and YAC 33CA11 (Figure C, Lane 5) respectively. The YAC clones of 4HG2 and 33CA11 which could not be separated any further were excised from the gel together with the background yeast chromosomes as indicated in lanes 2 & 4 respectively of Figure 4.3C.
4. cDNA selection for enriched transcripts from genomic clones

![Image of an electrophoretic separation of YAC genomic DNAs using pulse field gels.](image)

Figure 4.6 Electrophoretic separation of YAC genomic DNAs using pulse field gels.

Separation of five different YAC clone DNAs from co-migrating yeast chromosomal bands by CHEF gel electrophoresis. All gels were 1% low melting point agarose in 0.5 X TBE, run at 170V/cm and at a switch interval which separated the DNAs in the range of 450Kb-200Kb.

Figures (A) and (B) show ethidium bromide stained gels with separated YACs of (A) Rhodopsin - 260Kb (Lanes 1), A1220 - ~390Kb (Lanes 2&3) and C1228 - 320Kb (Lanes 4&5) & (B) 4HG2 - ~250Kb (Lanes 2&3) and 33CA11 - 290Kb (Lanes 4&5) which co-migrated with S.cerevisiae yeast chromosomal DNA bands of I and VI of 225Kb and 285Kb respectively (Lane 1). This co-migration was determined by hybridisation results of Southern blotted gel (B) to human placental DNA which showed distinct signals at 250Kb for YAC 4HG2 (Lane 3) and 290Kb for YAC 33CA11 (Lane 5) as indicated in (C). The filter was washed to a stringency of 0.2 X SSC, 0.1% SDS at 65°C. The YACs of 4HG2 and 33CA11 which could not be further separated were excised from the gel together with the contaminating yeast chromosomes as shown in Figure C, lanes 2 & 4 respectively. The S.cerevisiae yeast chromosomal DNA (strain YNN295) was used as a standard size marker as shown in Lanes 6 and 1 of figures (A) and (B) respectively.

221
4.2.2.2 Digestion and ligation of YAC genomic DNA to linkers

The YAC DNA was independently digested with MboI and AluI restriction enzymes. MboI and AluI digested DNA produced sticky ended and blunt ended restriction fragments respectively. The YAC DNA slice in LMP agarose (~ 0.5-1 g) was pre-equilibrated with 500μl 1X restriction enzyme buffer for 30 minutes at room temperature. The buffer was removed and the agarose slice was melted at 65°C for 5 minutes. Each YAC was aliquoted into eppendorf tubes with 34μl per tube. For each YAC digest, 4μl 10X restriction enzyme buffer and 20 units of restriction enzyme, (2μl) of either MboI or AluI enzyme was added into the 34μl YAC genomic DNA aliquot to make a final volume of 40μl and incubated at 37°C for 4-6 hours.

The digested DNA was labelled with biotin, using a 5'-biotinylated PCR primer to amplify the purified fragmented YAC. This procedure requires an oligonucleotide adaptor ligated to the YAC. This adaptor/linker is the template on which the biotinylated oligonucleotide primes the PCR. MboI and AluI linkers were used as adaptors to MboI digested and AluI digested YAC DNA respectively. Top and bottom strands of the MboI linkers were annealed before they were used as linkers in the ligation reaction. Equal amounts of both strands were aliquoted into an eppendorf tube, heated to 80°C for 5 minutes and then allowed to anneal at room temperature for 30 minutes. The same procedure was carried out for the AluI linkers. The sequences of the top and bottom strands of the MboI and AluI linkers used are given below:

**MboI Linkers**

24 - mer Top strand  5' GATCTCGACGAATTCGTGAGACCA 3'
20 - mer Bottom strand  3' AGCTGCTTAAGCACTCTGGT 5'

**AluI Linkers**

24 - mer Top strand  5' GATCTCGACGAATTCGTGAGACCA 3'
20 - mer Bottom strand  3' CTAGAGCTGCTTAAGCACTC 5'

The ligation reaction was set up as follows:
5μl of 10X ligase buffer (Gibco BRL), 5 μl of T4 DNA ligase (1 unit / μl; Gibco BRL) and 1μl of either MboI or AluI linkers (300ng/μl), (or a 50 fold molar excess of linkers,) were added to the digested YAC in a final volume of 51μl. Ligation reactions were incubated at room temperature for approximately 16 hours.
4.2.2.3 5' Biotinylated PCR of tinkered genomic DNA

Biotinylation of DNA oligonucleotides is essential for using Dynabeads M-280 Streptavidin (Dynal, Norway). The high affinity of the streptavidin/biotin interaction and the stability of the magnetic bead-streptavidin interaction, allow specific DNA manipulations such as melting, elution and hybridisation. The biotin can be introduced into the genomic clones by photobiotinylation (Korn et al., 1992) or nick-translation (Korn et al., 1992); in both cases, biotin is incorporated randomly along the length of the DNA. An alternative method (Morgan et al., 1992) is to use a biotinylated primer during an in vitro PCR amplification reaction. The advantage of this method is that large amounts of the YAC DNA can be amplified, and the biotin moiety is at the end of each small fragment rather than present many times within each fragment. This results in less steric hindrance from the large biotin moiety during PCR or during genomic DNA/cDNA hybridisation.

Selector DNA was generated by PCR amplification of the purified ligation, using biotin-labelled MboI and AluI linker primers complementary to the linker sequences as shown:

MboI 5' biotinylated linker primer
21 - mer 5' Biotin-TGG TCT CAC GAA TTC GTC GA 3'

AluI 5' biotinylated linker primer
21 - mer 5' Biotin- CTC ACG AAT TCG TCG AGA TC 3'

The ligation reactions were melted at 65°C for 5 minutes and 5µl was added to a 100µl PCR reaction. PCR reactions were set up as described in section 2.6.2 using dNTPs at a final concentration of 0.2mM, and 200 pmoles of biotinylated primers, with 1 X Taq buffer (Bioline) and 0.05 units BioTaq enzyme (Bioline). The amount of biotinylated primer used allows for quantitation of the amount of strepavidin-coated magnetic beads to use in the binding reaction: this is harder to estimate if the genomic DNA is photobiotinylated or nick-translated with biotin dUTP.

Thermal cycling was performed for 30 cycles with each cycle consisting of denaturation at 94°C for 30 seconds, annealing at 60°C for 30 seconds and extension at 72°C for 1 minute 30 seconds. The duration of the extension step was increased by 5 seconds per cycle using the automatic segment extension option on a 9600 Perkin Elmer PCR machine. A number of PCR reactions were set up for each YAC digest with the respective biotinylated PCR primers. 10µl of PCR amplified products were run and analysed on a 1.2% agarose gel (section 2.1.3) and visualised by ethidium bromide staining (Figures 4.7A & 4.7B). Figure 4.7A shows a smear of YAC 33CA11 MboI
4. cDNA selection for enriched transcripts from genomic clones

Figure 4.7 Assessment of quality of linked YAC genomic DNA.

Fractionation and assessment of YACs 33CA11 MboI and A1220 AluI digested, linked and biotinylated PCR amplified products. The amplified products were detected by fractionation on 1.2% agarose, ethidium bromide stained gels (A) & (B).

(A) Shows smears of YAC 33CA11 MboI digested, linked and biotinylated PCR amplified products of a size range of 200 bp to 500 bp.

(B) Shows smears of YAC A1220 AluI digested, linked and biotinylated PCR amplified products of a size range of 200 bp to 400 bp. The results indicated that the YAC genomic DNAs were digested to completion by the MboI and AluI four base cutter enzymes, showing amplified products which did not extend beyond 500bp. The ØX174/HaeIII molecular weight marker (M.W.M) was used as a size comparison.

The quality of the linked and PCR amplified genomic material in gels (A) and (B) was assessed by hybridisation of Southern blots of these gels to human Cot-1 DNA and the TIMP-1 cDNA respectively as indicated in (C) and (D). The filters were washed at a stringency of 0.2 X SSC, 0.1% SDS at 65°C.

(C) Intense hybridisation signals to human Cot-1 DNA showed the presence of good quality linked YAC 33CA11 MboI digested genomic DNA in a size range of 200 bp to 500 bp.

(D) Hybridisation of YAC A1220 AluI digested, linked genomic DNA with the 610 bp TIMP-1 cDNA fragment gave positive signals in a size range of 200 bp to 400 bp, indicating the presence of the TIMP-1 gene within this genomic material.
digested, linked and biotinylated PCR amplified products in the size range 500bp to 200bp which is the expected size range of a tetranucleotide target for a four base pair cutter enzyme such as \textit{MboI}. This shows that the YAC DNA was well digested, probably to completion since the smear did not extend beyond 500bp. Similar results were also obtained for the YAC A1220 \textit{AluI} digested, linked and biotinylated PCR amplified products (Figure 4.7B), which show a smear in a size range of 400bp to 200bp. Similar results were also obtained for all the other YACs (data not shown).

### 4.2.2.4 Assessment of quality of linked genomic DNA using Southern hybridisation

The quality of the linked and PCR amplified genomic material was assessed by Southern blotting agarose gels containing the biotinylated PCR products. Hybridisation was performed as described in section 2.3 using human Cot-1 DNA and total yeast DNA probes. Further assessment for random representation of genomic fragments was also performed by hybridisation with known marker probes such as the TIMP-1, properdin and rhodopsin genes contained within some of the YACs used.

Figures 4.7C and 4.7D show hybridisation results obtained using Cot-1 DNA and TIMP-1 cDNA probes on YACs 33CA11 \textit{MboI} digested and A1220 \textit{AluI} digested respectively. Intense hybridisation signals to human Cot-1 DNA (Fig. 4.7C) showed the presence of good quality linked YAC 33CA11 \textit{MboI} digested DNA in the size range 500bp to 200bp as described section 4.2.2.3. Similar results were also obtained for all the other \textit{MboI} and \textit{AluI} digested and linked YACs (data not shown). Total yeast genomic DNA showed no hybridisation to any of the digested and linked genomic material in the gel blots (data not shown).

Hybridisation of YAC A1220 \textit{AluI} digested and PCR amplified genomic DNA with the TIMP-1 cDNA 610bp probe gave positive signals in the size range 400bp to 200bp (Fig. 4.7D), indicating the presence of the TIMP-1 gene within this YAC genomic material. Similar results were also obtained with the \textit{MboI} digested and PCR amplified A1220 YAC and with the C1228 YAC \textit{MboI} and \textit{AluI} digested and PCR amplified genomic DNA (data not shown). The presence of the properdin gene was evident from positive hybridisation results with the YACs 33CA11 and 4HG2 \textit{MboI} and \textit{AluI} digested and PCR amplified using the 210bp properdin cDNA probe (data not shown). Positive hybridisation signals corresponding to the rhodopsin gene (220bp cDNA fragment) were present in the rhodopsin YAC \textit{MboI} and \textit{AluI} digested and PCR amplified genomic DNA (data not shown) but not in any of the other YAC genomic material.

The presence of positive hybridisation signals with human Cot-1 DNA and negative hybridisation signals with total yeast DNA indicated clearly that the DNA from the various digested, linked and amplified YACs was human genomic DNA with no
4. cDNA selection for enriched transcripts from genomic clones

contaminating yeast DNA present. The smear of genomic DNA in the size range 
~500bp-200bp obtained from both MboI and AluI digested and PCR amplified YAC 
genomic DNAs indicated that both restriction enzymes had worked well. This size range 
of genomic DNA is ideal for hybridisation (sections 4.2.4 & 4.2.5) with λZAPII and λgt10 
≤1kb amplified cDNA inserts which fall within a similar size range (section 4.2.1.3).

4.2.2.5 Precipitation and resuspension of 5'-biotinylated PCR amplified 
YAC genomic DNA

The concentrations of the biotinylated genomic DNA for each digested YAC were 
estimated from the fractionated PCR products on the agarose gels (sections 4.2.2.3) by 
comparison with the intensity and known concentration of the ØX174/HaeIII marker 
bands. Since approximately 100ng of biotinylated genomic DNA is used per 
hybridisation, 50ng of MboI digested and PCR amplified YAC genomic DNA and 
50ng of AluI digested and PCR amplified YAC genomic DNA were pooled together. This 
was precipitated with 1/10th volume of 3M sodium acetate pH 6.0 and 2 volumes of 
absolute ethanol. The precipitate was washed in 70% ethanol, air-dried and 
re-suspended in 20μl of 0.1M sodium phosphate (pH 7.0), ready for pre-blocking and 
hybridisation as described in sections 4.2.4 and 4.2.5.

4.2.3 Preparation of Pre-blocking quenching agents (Yeast DNA, 
Ribosomal DNAs, Human Placental DNA and LINE1 DNA)

Pre-blocking agents required for quenching of repetitive elements and 
contaminating ribosomal sequences (section 4.1.8), were prepared from plasmid clones. 
These clones were initially transformed into bacteria and DNA from the recombinant 
cloned inserts from the plasmid vectors on Low Melting Point (LMP) agarose. The inserts in 
LMP agarose were purified, precipitated and the concentrations determined. They were 
then sonicated to a suitable size range, ready for pre-blocking, before hybridisation as 
described in sections 4.2.4 and 4.2.5.

Human placental DNA was obtained commercially (Gibco, BRL). Although it 
was supplied already sheared, it was sheared to a smaller size range as described in 
section 4.2.3.3. Yeast genomic DNA from strain AB1380, which is the parent host of the 
various YACs used, was obtained commercially (Gibco, BRL). This was also sonicated to 
a suitable size range before use.
4. cDNA selection for enriched transcripts from genomic clones

4.2.3.1 Isolation of insert DNA from plasmid clones

Plasmid DNA from ribosomal RNA clones (section 4.1.8) in plasmids pR7.3, pR5.8 (human ribosomal RNA 45S precursor regions), pK515 (E.coli ribosomal RNA operon) and RibH7, RibH15 (yeast ribosomal RNA inserts) and the long interspersed element (LINE1) in plasmid clone pL1A were prepared by firstly transforming the various plasmids into DH5α bacterial cells as described in section 2.4.2. The transformed clones were then grown overnight in LB media supplemented with ampicillin (50μg/ml) and DNA prepared as described in section 2.4.3.2 using the Wizard minipreparation kit (Promega). The concentration of the DNA was determined by fractionation on a agarose gel (data not shown). For long term storage purposes, glycerol stocks of the various clones were also prepared as described in section 2.4.1.1.

1μg of the plasmid DNA was digested with the appropriate restriction enzymes (section 4.1.8) to release the cloned inserts and their insert sizes determined and checked before proceeding to a larger scale preparation of the DNA. Figure 4.8A shows the separated insert fragments of the pR7.3 (7.3kb), pR5.8 (5.8kb), pK515 (7kb), RibH7 (2.6kb) and RibH15 (7kb) plasmid clones. A common pBR322 vector band of 4.4kb is seen in all digested clones. Since the insert bands obtained were of the expected sizes (section 4.1.8), a greater quantity of plasmid DNA was prepared for a larger scale digest.

Large scale cultures of the six different pre-blocking agents were set up and the plasmid DNA prepared as described in section 2.4.3.1. The concentration of the DNA obtained for each pre-blocking agent was determined by agarose gel fractionation and comparison with known concentrations of molecular weight standard markers (data not shown). A large scale digest using 10μg of plasmid DNA for each preblocking agent was carried out at 37°C with 30 units of the appropriate restriction enzyme in a total volume of 100μl. A small aliquot of the digested DNA was run on an agarose gel to check for complete digestion (data not shown). The rest of the digested DNA in each case was then fractionated on a 1% LMP agarose gel, at 40 Volts and at a temperature of 4°C. A slotted comb was used to load the larger volume of DNA as seen in Figure 4.8B, which shows an example of two of the pre-blocking agents, RibH7 and RibH15 separated in this way. A common pBR322 vector band of 4.4kb is seen in both lanes. The RibH7 and RibH15 insert DNA sizes of 2.6kb and 7kb respectively were well separated, ready for isolation and purification from the LMP agarose. The other pre-blocking agents were prepared in a similar manner.
Figure 4.8 Isolation of insert DNA from ribosomal RNA plasmid clones.

Small and large scale restriction enzyme digests of ribosomal RNA plasmid clones for release, size determination and isolation of cloned inserts used as preblocking agents against contaminating ribosomal sequences.

(A) Fractionation of restriction enzyme digested (small scale) plasmid DNA of pR7.3, pR5.8 (EcoRI digested; Lanes 1, 2 & 3), pKS15 (BamHI digested; Lane 4) and RibH7, RibH15 (HindIII digested; Lanes 5, 6 & 7) ribosomal RNA clones. The ethidium bromide stained 1% agarose gel showed the separated insert fragments of the pR7.3 (7.3kb), pR5.8 (5.8kb), pKS15 (7.3kb), RibH7 (2.6kb) and RibH15 (7kb) plasmid clones as indicated by arrows. A common pBR322 vector band of 4.4kb is present in all digested clones (Lanes 1-7). The λ/HindIII molecular weight marker (M.W.M) was used as a size comparison.

(B) Fractionation of a large scale HindIII restriction enzyme digest of the RibH7 and RibH15 plasmid DNA. The ethidium bromide 1% LMP agarose gel showed RibH7 and RibH15 insert DNA sizes of 2.6kb and 7kb respectively, which were well separated and ready for isolation and purification from the LMP agarose. The λ/HindIII molecular weight marker (M.W.M) was used as a size comparison.
4. cDNA selection for enriched transcripts from genomic clones

4.2.3.2 Recovery and purification of insert DNA from Low-melting point agarose gels

The separated insert fragments as described above were viewed under a long wave-length UV lamp and cut out from the LMP agarose gels using a clean sharp scalpel. The LMP agarose slices were transferred into small petri dishes containing TE buffer pH 8.0. Insert DNA from the LMP agarose slices was isolated and purified using two different methods as described below.

4.2.3.2.1 Glasswool method

Two eppendorfs were used for this method. One of the eppendorfs was punctured at the bottom and then 1/3 filled with glass wool. A piece of 3MM Whatman paper was cut to form a well and placed above the glass wool. The excised DNA agarose slice was then placed on top of the Whatman paper. This inner eppendorf tube was placed into another eppendorf tube which was then centrifuged at 4,000 rpm for 30 seconds. A second spin was carried out after 20-60μl of sterile water had been added to the agarose slice in the inner tube. During the second spin, the agarose slurry was caught in the glasswool and the purified DNA passed through the punctured hole of the inner tube to be collected in the outer eppendorf. The concentrations of the purified insert DNA were determined by gel electrophoresis.

4.2.3.2.2 β-Agarase method

An alternative method to recover and purify DNA from LMP agarose (Burmeister and Lehrach, 1989) involves using β-Agarase I (NEB). As this method is gentle it is useful for the isolation of high molecular weight DNA. β-Agarase I digests agarose, releasing trapped DNA and producing carbohydrate molecules which can no longer gel. It can be used to purify both large (>50 kb) and small (<50 kb) fragments of DNA from gels, and the resulting carbohydrates are removed by phenol/chloroform extraction.

The insert DNA-containing LMP agarose was equilibrated and washed twice in 2 volumes of 1 X agarase buffer (NEB). The equilibrated agarose slices were then transferred to eppendorf tubes and melted at 65°C for 10 minutes. The tubes were then cooled to 40°C and the molten agarose incubated with 1 unit of β-Agarase I for 1-2 hours at 40°C. This procedure digests up to 200μl of 1% LMP agarose. The resulting carbohydrates in the digested agarose were removed by phenol/chloroform extraction. After extraction the DNA in the supernatant was precipitated using 0.3M sodium acetate and 2 volumes of cold isopropanol after centrifugation at 13,000 rpm for 10 minutes.
4. cDNA selection for enriched transcripts from genomic clones

The resulting pellet was washed with cold 70% isopropanol, spun as before and the pellet air-dried and re-suspended in 10μl sterile water. The quality and quantity of the recovered DNA was determined by running 1μl aliquots of the pre-blocking agents on 1% agarose gels, as shown in Figure 4.9. Concentrations were estimated by comparison with known concentrations of λ/HindIII standard molecular weight markers. The concentrations of the ribosomal DNAs pR7.3, RibH7 and RibH15 were estimated to be ~300ng/μl, and that of pR5.8 and the LINE1 DNA- pL1A to be ~500ng/μl (Figure 4.9). Both the glasswool and β-agarase methods gave successful recoveries of insert DNAs from LMP agarose, with the β-agarase method giving the best yield. The quenching DNAs were then sonicated as described below to smaller sized fragments before use in the pre-blocking of genomic DNA (sections 4.2.4 and 4.2.5).

4.2.3.3 Sonication of quenching agents

Quenching agents (sections 4.1.8, 4.2.3) including human placental DNA and yeast DNA of high molecular weight sizes were sonicated to smaller fragments in the size range ~500bp-200bp using an Ultrasonic disintegrator Soniprep 150 (FISONS). The Soniprep 150 transmits sound energy to the medium through high frequency vibration at the tip of an interchangeable probe. The sonication effect is produced by the energy dissipated as the high intensity ultrasound waves create many series of micro-cavities in the dispersion medium.

Pre-blocking DNAs were sonicated to a size range of ~600bp-200bp by controlling the amplitude (microns) and time (minutes) in order to obtain the optimum conditions which produced such size fragments. This was obtained after a few attempts by changing the amplitude and time parameters and running an aliquot of the unsonicated and sonicated material after each attempt on 1.2% agarose gels, to determine whether the DNA had been sonicated to the required size fragments. Figure 4.10A shows a comparison of size fragments obtained from unsonicated and sonicated pre-blocking human placental and yeast DNAs. Human placental DNAs obtained commercially (Gibco) were supplied pre-sheared, but on fractionation on a 1.2% agarose gel were found to lie in a large MW size range of 23kb-200bp (Fig. 4.10A, lane1). This was sonicated further to the size range of ~650bp-200bp (Fig. 4.10A, lane 2). Unsonicated yeast genomic DNA of the AB1380 strain of >23kb (Fig. 4.10A, lane 3) was found to be only partially sonicated, with a size range of 3kb-200bp (Fig. 4.10A, lane 4) and was hence further sonicated to the optimal size of ~500bp-200bp (Fig. 4.10A, lane 5).

Equal quantities of the five different pre-blocking ribosomal DNAs and the LINE1 DNA (pL1A) prepared as described in section 4.2.3 were pooled together and sonicated as described above. An aliquot of this sonicated DNA was run on a 1.2% agarose gel as shown in Fig 4.10B, which shows sonicated size fragments of ~500bp-
Figure 4.9 Purified insert DNAs isolated from LMP agarose using β-agarase 1.

The ethidium bromide stained 1% agarose gel showed the quality and quantity of precipitated insert DNAs from the ribosomal and LINE 1 plasmid clones used as quenching agents. The concentrations were estimated by comparison with known concentrations of λ/HindIII standard molecular weight markers (M.W.M) of 500ng (Lane 8) and 1µg (Lane 9). The concentrations of ribosomal DNAs pR7.3 (Lanes 1&2), RibH7 (Lane 6) and RibH15 (Lane 7) were estimated to be ~300ng/µl and that of pR5.8 (Lanes 3&4) and the LINE 1 DNA-pL1A (Lane 5) to be ~500ng/µl, indicating successful recovery of pre-blocking insert DNAs from LMP agarose using β-agarase 1.
Figure 4.10. Qualitative and quantitative analysis of sonication products of quenching agents.

Qualitative analysis of sonicated DNA of the various quenching agents used was determined by comparison of size fragments in order to obtain the optimal size range of ~500-200bp as described in Figure (A). Quantitative analysis by determining the concentrations of the various sonicated DNAs is described in Figure (B).

(A) The ethidium bromide stained 1.2% agarose gel show comparison of size fragments obtained of unsonicated and sonicated pre-blocking human placental and yeast DNAs. Commercially obtained, pre-sheared but unsonicated (UnS) placental DNAs lie in a size range of 23kbp-200bp (Lane 1). This was then sonicated to a more optimal size range (OS) of ~650-200bp (Lane 2). Unsonicated yeast genomic DNA of AB1380 strain of >23kbp (Lane 3) was partially sonicated (PS) initially to a size range of 3kbp-200bp (Lane 4) and then further sonicated to a optimal size range of ~500-200bp (Lane 5). Standard molecular weight markers (M.W.M) of λ/HindIII and ØX174/HaeIII were used as size comparisons.
Figure 4.10B

The ethidium bromide stained 1.2% agarose gel show concentrations of the sonicated products of the human placental DNA and yeast DNA at 500ng/μl (Lanes 1 & 3 respectively). An aliquot of diluted yeast DNA was determined at 100ng/μl (Lane 2). Equal quantities of pooled ribosomal DNAs and LINE1 DNA (pL1A) after sonication gave a concentration of 500ng/μl (Lane 4). All DNAs of the different quenching agents used in pre-blocking were sonicated to the optimal size range of ~600-200 bp as indicated in lanes 1-4. Standard molecular weight markers (M.W.M) of λ/HindIII and ØX174/HaeIII were used as size comparisons.
4. cDNA selection for enriched transcripts from genomic clones

200bp similar to the sonicated human placental DNA (Figure 4.10B, Lane 1) and yeast DNA (Figure 4.10B, Lanes 2 & 3). The concentrations of these sonicated DNAs were determined as before and estimated to be 500ng/μl in each lane except lane 2 (diluted yeast DNA) which was 100ng/μl. The pooled and sonicated ribosomal and LINE1 DNAs were then diluted to a concentration of 200ng/μl, ready for pre-blocking of YAC genomic DNA. Sonication of high molecular weight pre-blocking agents to a smaller size range of ~600bp-200bp was essential in order to facilitate the solution prehybridisation (sections 4.2.4.1 & 4.2.5.1) of these quenching agents to the biotinylated genomic DNA which had been prepared to a size range of 500bp-200bp (sections 4.2.2.3 & 4.2.2.4).

4.2.4 First round of direct cDNA selection

The first round of direct cDNA selections was performed by initially suppressing non-specific binding to target biotinylated genomic DNA by prehybridisation with sonicated quenching agents. Pre-blocked genomic DNA was then hybridised in solution to amplified cDNA inserts. After hybridisation, biotinylated hybrids were immobilised by streptavidin-coated magnetic beads and subjected to post-hybridisation washes. Captured hybrids were eluted and amplified using cDNA vector specific primers. PCR amplified primary selected cDNAs were then purified, precipitated and prepared ready for the second round of cDNA selection.

4.2.4.1 Pre-blocking of genomic DNA with quenching agents

100ng of biotinylated YAC genomic DNA in 20μl of 0.1M sodium phosphate pH 7.0 (section 4.2.2.5) was denatured by heating at 95°C for 5 minutes. To block repeat and ribosomal sequences the denatured genomic DNA was pre-hybridised for competition with a mixture of sonicated quenching agents comprising 2μg (500ng/μl) human placental DNA (BRL), 1μg (500ng/μl) of yeast DNA (AB1380) and 1.2μg total quantity of 200ng/μl concentration of pooled sonicated ribosomal and LINE1 DNAs, prepared as described in section 4.2.3. 200ng of of LINE1 DNA (pL1A) and each of the five different ribosomal DNAs, from ribosomal RNA clones pR7.3, pR5.8 (human ribosomal RNAs), pK515 (E. coli ribosomal RNA) and RibH7, RibH15 (yeast ribosomal RNAs) were used per pre-hybridisation reaction. Pre-hybridisation was carried out by adding the quenching agents immediately to denatured genomic DNA to a final volume of 32μl with preblocking at 65°C for about 4 hours, using a 9600 Perkin Elmer PCR machine to avoid the use of mineral oil.
4.2.4.2 Solution hybridisation of preblocked genomic DNA to cDNA inserts

1 μg (1μl) of prepared cDNA inserts (section 4.2.1.4) of each of the amplified λZAPII and the λgt10 (pooled ≥ 1kb and ≤ 1kb size cDNAs) libraries were denatured independently at 95°C for 5 minutes. cDNA inserts from each library were then hybridised at 65°C in 0.1 M sodium phosphate pH 7.0, to the blocked biotinylated YAC genomic DNA (100ng) in separate hybridisation reactions for 16-20 hours using a 9600 Perkin Elmer PCR machine to avoid the use of mineral oil. In a similar manner separate hybridisations were carried out for each of the blocked target genomic DNA from the five different YACs used (section 4.1.6), using each of the two different amplified cDNA library inserts of λZAPII and λgt10 as hybridising cDNAs. Ten separate first round hybridisation reactions were set up.

4.2.4.3 Capture, wash and elution of specific primary selected cDNAs

The specific cDNAs bound to the biotinylated YAC genomic DNA were captured using streptavidin coated magnetic beads (Dynabeads M280, Dynal). Dynabeads M-280 Streptavidin were supplied as a suspension containing 6-7 x 10^8 Dynabeads/ml (10mg/ml). The Dynal Magnetic Particle Concentrator (MPC) acted as a solid phase which allowed the isolation and subsequent handling of target molecules in a highly specific manner.

Since approximately 100ng of biotinylated genomic DNA was used per hybridisation reaction (section 4.2.4.2), this represented about 200 pmoles of biotin molecules which is the minimum amount that could bind to 1mg of streptavidin coated magnetic beads. Therefore 100μl of a 10mg/ml solution of Dynabeads (1mg total) was rinsed in 6 X SSC solution (100μl) to remove any preservatives before use. A Dynal Magnetic Particle Concentrator (MPC) was used to facilitate this rinsing procedure, by placing the tube containing the rinsed Dynabeads in the Dynal MPC for a minute which carries down the paramagnetic beads. The supernatant (rinse solution) was aspirated with a pipette, then the tube was removed from the Dynal MPC and rinsing repeated twice.

After the rinsing with 6 X SSC, the beads were re-suspended in 200μl of TE buffer pH 8.0 plus 1M NaCl. The hybridisation mixture (section 4.2.4.2) was added to the beads and allowed to sit at room temperature for 15 minutes, in order to capture the biotinylated genomic DNA on the beads. The unbound cDNAs were pipetted off after precipitating the beads with bound cDNAs using a MPC for 1 minute as before. The beads were then subjected to the following washes in a volume of 100μl: 1 X SSC, 0.1% SDS, (two washes at room temperature), followed by 0.1 X SSC, 0.1% SDS,
4. cDNA selection for enriched transcripts from genomic clones

(three washes at 65°C), each wash for 15 minutes. The beads were precipitated as before and the wash solutions pipetted off into another tube. The unbound cDNAs and wash solutions were saved each time for further experiments (section 4.2.4.4).

After the washings, the magnetic beads containing captured biotinylated genomic DNA with bound cDNAs were resuspended in 100µl TE buffer pH 8.0. The specific primary selected cDNAs were eluted by heating the sample to 95°C for 5 minutes. The beads were precipitated as before and the specific primary selected cDNAs were transferred to a fresh eppendorf tube.

4.2.4.4 PCR amplification of primary selected cDNAs

The unbound DNA, DNA from the wash solutions and eluted specific cDNAs from the first round of selection were PCR amplified using 10µl from each solution as a template in a 100µl PCR reaction. PCR was carried out for 30 cycles with the same cycling conditions using cDNA vector-specific primers (ie. λZAPII or λgt10 specific primers) as described in section 4.2.1.3. The PCR products were separated by electrophoresis on a 1.2% agarose gel and visualised by ethidium bromide staining. An example of the amplified PCR products obtained is shown in Figure 4.11A and 4.11B from the λZAPII cDNAs of YAC A1220 and λgt10 cDNAs of YAC 33CA11 respectively.

The PCR amplified products of unbound DNAs after first round selection of λZAPII cDNAs for YAC A1220 were present in a short smear of about 500bp to 200bp, and most concentrated at 200bp (Figure 4.11A, Lane 1). These seem to be very short fragment cDNAs which had not bound to the target genomic DNA, and were probably contaminant clones and repeat, ribosomal cDNAs which had been pre-blocked in the genomic DNA but were present in the λZAPII cDNA library. The smear of PCR products obtained from the wash solutions 1 to 5 (Fig. 4.11A, Lanes 2 - 6) fall in a size range of ~1.0 kbp to 300bp, the smear being most concentrated at 800bp to 600bp. This smear probably represents non-specific cDNAs which had loosely bound to the target genomic DNA but had been removed after washings with increased stringency (section 4.2.4.3). The smear of PCR products obtained from eluted, bound λZAPII cDNAs (Fig. 4.11A, Lane 7) is similar in size range to that obtained from the washing (Fig. 4.11A, Lane 2 - 6) with greater concentration again at 800bp to 600bp, the increased intensity implying the more abundant sized cDNAs present. The λZAPII non-specific (Lanes 2-6) and bound (Lane 7) cDNAs of YAC A1220 were obtained in a size range of 1kb-300bp since the λZAPII cDNA library contained shorter random primed and oligo dT primed sized cDNA inserts. PCR products from primary selected specific cDNAs of YAC A1220 (Fig. 4.11A, Lane 7) showed no discrete bands and appeared more uniform in intensity in comparison to the amplified starting λZAPII cDNAs (Figure 4.4A) described in section 4.2.1.3. This could indicate that the more abundant contaminant DNAs such as
Figure 4.11  Comparison of PCR amplified products from first round and second round cDNA selection.

PCR amplified products from the unbound DNAs, wash solutions DNAs and eluted primary and secondary selected specific cDNAs after first and second round of cDNA selection respectively are shown in 1.2% agarose ethidium bromide stained gels in (A), (B) and (C). The φX174/HaeIII and the λ/HindIII standard size molecular weight markers (M.W.M) were used as size comparisons.

Figure (A) shows PCR amplified products obtained at various stages after the first round of cDNA selection of λZAPII cDNAs from YAC A1220. The PCR amplified products from unbound DNAs were present in a short smear of about 500bp to 200bp, being most intense at 200bp (Lane 1). The smear of PCR amplified products obtained from the wash solutions 1 to 5 (Lanes 2-6) fall in a size range of ~1.0kbp to 300bp, being most concentrated at 800-600bp. This represents non-specific cDNAs which were loosely bound to the target genomic DNA but were removed with high stringency washing. PCR amplified products obtained from specific primary selected λZAPII cDNAs (Lane 7) are similar in size range to those obtained from the wash DNAs with greater intensity at 800-600bp, the increased intensity indicating the more abundantly sized cDNAs present. The no DNA negative result (Lane 8) indicated absence of contaminating DNAs in the PCR reactions.

Figure (B) shows PCR amplified products at various stages after first round of cDNA selection of λgt10 cDNAs from YAC 33CA11. The PCR amplified products from unbound DNAs were present in a uniform smear of 800-300bp (Lane 1). The smear of amplified products from the wash solutions 1 to 6 (Lanes 2-7) were of a size range of ~4kbp to 300bp. Amplified products from the bound specific primary selected λgt10 cDNAs (Lane 8) were present in a size range of 1.5kbp-300bp (Lane 8), showing specific binding of these eluted cDNAs to biotinylated target YAC 33CA11 DNA (size range 500-200bp, Figure 4.7A). These bound specific λgt10 primary selected cDNAs (Lane 8) were of uniform intensity as compared to the starting λgt10 cDNAs (Lane 9) of ~4kbp-300bp which were of greater intensity from 1kbp-600bp. The no DNA negative result (Lane 10) indicated absence of contaminating DNAs in the PCR reactions.

Figure (C) shows PCR amplified products at various stages after second round of cDNA selection of λZAPII cDNAs from YAC A1220. The amplified products from unbound DNAs were present in a short smear of 500-200bp (Lane 1) which indicated short fragment cDNAs which had not bound to the target YAC genomic DNA. The smears of amplified products from the second round selected wash solutions 1 to 5 (Lanes 2-6) and the bound specific secondary selected cDNAs (Lane 7) were present in a size range of 800-200bp, the smears of amplified products being more uniform here compared to that of the first round selection products in figure (A). The secondary selected λZAPII cDNAs (Lane 7) were present in a smaller size range of 800-200bp as compared to the primary selected cDNAs from the same YAC (Lane 8) and the original λZAPII starting cDNAs (Lane 9) both of which were present in a size range of 1.5kbp-200bp. This indicated enrichment upon selection for specific sized cDNAs in a smaller size range of 800-200bp which had bound to similar sized genomic target DNAs. A greater decrease in the abundant and/or common contaminating cDNAs was observed in the uniform intensity of the secondary selected cDNAs (Lane 7) as compared to that from the primary selected (Lane 8) and starting cDNAs (Lane 9) which show greater intensities at 600bp and 300bp.
repeat and ribosomal elements had been selected against to a certain extent after first round of selection. The control with no DNA showed a negative result (Fig. 4.11A, Lane 8), showing no other contaminating DNA in the PCR reactions.

The PCR amplified products of unbound DNAs after first round selection of λgt10 cDNAs for YAC 33CA11 were present in a uniform smear of about 800bp to 300bp (Figure 4.11B, Lane 1). As described above these are short fragment cDNAs which had not bound to the target genomic YAC 33CA11 DNA, probably being contaminant clones and repeat, ribosomal cDNAs. The smear of PCR products obtained from the wash solutions 1 to 6 (with an additional 6th wash) (Fig. 4.11B, Lanes 2 - 7) were of a size range of ~4 kbp to 300bp, probably representing non-specific cDNAs as described above in Figure 4.11A. Since the λgt10 cDNAs used were of the pooled >1kb (4kb-600bp) and ≤1kb (800bp-200bp) size ranges (Figs. 4.4B & 4.4C, section 4.2.1.3) the non-specific cDNAs obtained were hence of a larger size. This was in contrast to the eluted bound λgt10 cDNAs from YAC 33CA11 (Fig. 4.11B, Lane 8) which were obtained in a size range of 1.5kbp-300bp since most of the specific cDNAs here, had specifically bound to biotinylated target genomic YAC 33CA11 DNA which had been prepared in a size range of 500bp-200bp as described in section 4.2.2.3 (Figure 4.7A). Also the eluted specific cDNAs (Fig.4.11B, Lane 8) were clearly of uniform intensity when compared to the re-amplified starting λgt10 pooled cDNAs (Fig. 4.11B, Lane 9), present as a smear of PCR products from ~4kb-300bp, being of greatest intensity from 1kb-600bp.

4.2.4.5 Purification and precipitation of primary selected cDNAs

The PCR amplified products of the λZAPII and λgt10 first round selected and eluted specific cDNAs for each of the five different YACs were purified separately using the Wizard PCR Preps System (Promega) as described in section 4.2.1.4. The different purified cDNAs were then precipitated, washed and re-suspended in sterile water. A 1μl aliquot of this cDNA was run on a 1.2% agarose gel, to quantify the amount of DNA by comparison with a known concentration of a standard molecular weight marker such as φX174/HaeIII (data not shown). The volumes of the re-suspended primary selected cDNAs were then adjusted to a concentration of 1μg/μl, ready for the second round of cDNA selection/amplification. If necessary, the PCR for the primary selected cDNAs was scaled up in order to yield a total of 1μg of PCR products required per second round hybridisation reaction (section 4.2.5.2).
4.2.5 Second round of direct cDNA selection

Each of the amplified primary selected cDNA products (section 4.2.4) were subjected to another cycle of hybridisation and capture. The secondary selected cDNA products obtained from the second round of enrichment were again PCR amplified, purified, precipitated and prepared ready for cloning into a plasmid vector.

4.2.5.1 Preblocking of genomic DNA with quenching agents

As described in section 4.2.4.1, a fresh aliquot of 100ng of biotinylated YAC genomic DNA in 20μl of 0.1M sodium phosphate pH 7.0 (section 4.2.2.5) was heat denatured and prehybridised for competition with a mixture of sonicated quenching agents (sections 4.2.3 & 4.2.3.3), at 65°C for about 4 hours.

4.2.5.2 Solution hybridisation of genomic DNA to primary selected cDNA inserts

1μg (in 1μl) of prepared primary selected cDNA inserts (sections 4.2.4.4 & 4.2.4.5) of each of the amplified λZAPII and λgt10 libraries were denatured independently and hybridised to a fresh aliquot of preblocked YAC genomic DNA (section 4.2.5.1) in separate hybridisation reactions for 16-20 hours as described in section 4.2.4.2. In total ten separate second round hybridisation reactions were set up using the five different pre-blocked YAC genomic DNA and their respective primary selected λZAPII and λgt10 cDNA inserts.

4.2.5.3 Capture wash and elution of specific secondary selected cDNAs

This step is identical to that described in section 4.2.4.3. After hybridisation, biotinylated hybrids were again immobilised using streptavidin-coated magnetic beads and a magnetic particle concentrator (MPC) and subjected to posthybridisation washes as before (section 4.2.4.3). The specific secondary selected cDNAs were eluted in TE buffer pH 8.0 and precipitated again using the MPC as previously described (section 4.2.4.3).
4. cDNA selection for enriched transcripts from genomic clones

4.2.5.4 PCR amplification of secondary selected cDNAs

PCR amplification of secondary selected cDNAs was carried out as described in section 4.2.4.4. To avoid PCR artifacts while providing the desired overall amplification, internal primer sets closest to the EcoRI cloning site were used for the amplification of each of the secondary selected λZAPII and λgt10 cDNAs. The primer sets used were:

λZAPII
- SK Primer: 5' - TCT AGA ACT ACT CCA TC - 3'
- KS Primer: 5' - CGA GGT CGA CGG TAT CG - 3'

λgt10
- Forward Primer: 5' - AGC CTG GTT AAG TCC AAG CTG - 3'
- (Set B) Reverse Primer: 5' - CTT CCA GGG TAA AAA GCA AAA G - 3'

An example of the amplified PCR products obtained from the unbound DNA, wash solution DNA and eluted specific cDNAs after the second round of selection of the λZAPII cDNAs from YAC A1220 genomic DNA is shown in Figure 4.11C. As described in section 4.2.4.4 (Figure 4.11A), the amplified products from unbound DNAs (Figure 4.11C, Lane 1) were present in a short smear of 500bp-200bp indicating short fragment cDNAs which had not bound to the target genomic DNA. The smear of amplified products obtained from the second round selected wash solutions 1 to 5 (Fig. 4.11C, Lanes 2-6) and the bound specific cDNAs (Fig. 4.11C, Lane 7) were present in a size range of 800bp-200bp, the smears of secondary selected cDNAs being more uniform here unlike those seen in Figure 4.11A, of the first round selected products. The PCR products obtained from the wash solution DNA were non-specific λZAPII cDNAs (section 4.2.4.4) and those from the eluted cDNAs (Fig. 4.11C, Lane 7) were bound specific secondary selected λZAPII cDNAs.

The secondary selected λZAPII cDNAs from YAC A1220 (Fig. 4.11C, Lane 7) were found to be present in a smaller size range of 800bp-200bp compared to the primary selected cDNAs from the same YAC (Fig. 4.11C, Lane 8) and the original λZAPII starting cDNAs (Fig. 4.11C, Lane 9), which were present in a size range of 1.5kbp-200bp. This showed enrichment upon selection for specific sized cDNAs in a smaller size range of 800bp-200bp which had bound to similar sized genomic target DNA. Also the intensity of the smear obtained on fractionating PCR products of the secondary selected cDNAs (Fig. 4.11C, Lane 7) was more uniform compared to that from the primary selected cDNAs (Fig. 4.11C, Lane 8) and the starting λZAPII cDNAs (Fig. 4.11C, Lane 9), which showed greater intensity at sizes between 300bp and 600bp. This probably indicated a decrease in the abundant and/or common contaminating cDNAs present after the second round of selection. The no DNA controls were again
negative (Figure 4.11C, Lane 10) showing no other contaminating DNAs in the PCR
reactions used.

4.2.5.5 Purification and precipitation of secondary selected cDNAs

The PCR amplified products of the λZAPII and λgt10 second round selected and
eluted specific cDNAs from the five different YACs were each purified separately using
the Wizard PCR Preps System (Promega) as described in section 4.2.1.4. The purified
cDNAs were then precipitated, washed and re-suspended in sterile water
(section 4.2.1.4). The concentrations of the secondary selected cDNAs were determined
by comparison with a known concentration of a standard MW marker as described in
section 4.2.4.5.

4.2.6 Qualitative evaluation for enrichment and depletion of
cDNA transcripts by screening starting, primary and
secondary selected cDNAs

The relative abundance of specific transcripts in the re-amplified starting,
primary and secondary selected λZAPII and λgt10 cDNAs from each of the YACs was
determined in order to evaluate the extent of enrichment of selected transcripts such as
the reporter control genes TIMP-1 and rhodopsin and depletion of non-specific
transcripts for example ribosomal, yeast and β-actin DNAs. To assess qualitative
enrichment, equal amounts of PCR-amplified cDNA inserts from the starting λZAPII
and λgt10 libraries and from the primary and secondary enriched λZAPII and λgt10
sub-libraries for each of the specific YACs were dot-blotted onto membrane filters.
Size distribution and efficiency of the selection procedure was also assessed by
examining the nature of the captured cDNAs using Southern blots, prepared by
fractionating equal amounts of starting, primary and secondary selected cDNAs.
Hybridisation of control cDNAs and other contaminant DNA probes to both dot-blots
and Southern blots helped assess the specificity of the enrichment process.

4.2.6.1 Bio-dot preparation

The Bio-Dot microfiltration apparatus (Bio-Rad, USA) can be used for
applications requiring rapid immobilisation and screening of unfractionated or purified
proteins, nucleic acids or macromolecular complexes on nitrocellulose or Hybond N+
membranes. The entire apparatus was assembled according to manufacturer's
instructions to allow for a vacuum to be applied and a in-line 3-way flow valve to be
controlled during assay procedures.
The Hybond N+ membrane was pre-wetted in 6 X SSC, in order to ensure proper drainage of solutions during the blotting procedure. The membrane was then removed from the wetting solution and excess liquid drained from the membrane, which was then laid on the sealing gasket in the apparatus such that it covered all of the holes. The gasket was then aligned above the support plate, which is situated over the vacuum reservoir. A vacuum was then applied and the flow valve adjusted so that the vacuum manifold was open to atmospheric pressure. 100μl of TE buffer was applied to the sample wells using a Octapette pipette (CoStar, UK), which was carried out in order to rehydrate the membrane following the vacuum procedure. The buffer was then gently removed from the wells by vacuum, after which the vacuum was disconnected. 50ng and 100ng of sample target DNA was denatured by adding 0.4M NaOH, 10mM EDTA (final concentration), and heating the samples at 100°C for 10 minutes. The denatured DNA was then neutralised by adding an equal volume of cold 2M ammonium acetate, pH 7.0, to the target DNA solution. The denatured DNAs were then applied to individual wells of the 96-well sample template in concentrations of 50ng and 100ng and were bound to the membrane under a gentle vacuum. The blotted membrane was then rinsed in 2 X SSC, air dried and baked at 80°C for 30 minutes before being used for hybridisation. Multiple blots were made depending on the number of the probes to be used. Hybridisations of dot-blots to the various probe DNAs were performed as described in section 4.2.6.3.

4.2.6.2 Southern blot preparation

The efficiency of the selection procedure was also monitored using Southern blots. In order to evaluate DNA concentrations and size ranges of selected cDNA inserts, 100ng of λZAPII and λgt10 cDNAs from each stage of selection (starting, primary selected, and secondary selected) were fractionated by electrophoresis on 1.2% agarose gels, and Southern blots were prepared as described in section 2.2.1. The blots were then hybridised to the various probe DNAs as described in section 4.2.6.3 in order to assess starting and final abundance of both reporter control genes and contaminating transcripts.

4.2.6.3 Hybridisation of dot-blots and Southern blots with β-actin cDNA, ribosomal DNA, yeast DNA and a known retinal gene in the region.

The Bio-dots and Southern blots were hybridised (section 2.3) with β–actin, ribosomal and yeast radiolabelled DNA probes to evaluate for depletion of non-specific transcripts. They were also probed with known cDNAs from the region such as TIMP-1 and the positive control rhodopsin, to evaluate for enrichment of selected transcripts.
4. cDNA selection for enriched transcripts from genomic clones

After hybridisation the blots were rinsed in 2 X SSC/0.1% (w/v) SDS at room temperature for 10 minutes. This step was repeated, and depending on the required stringency, the filters were washed twice with 2 X SSC/0.1% SDS at 65°C for 15 minutes. This was followed by two 15 minute washes at 65°C with 1 X SSC/0.1% SDS. Stringency was further increased by two 15 minute washes at 65°C in 0.5 X SSC/0.1% SDS. All filters were kept moist to prevent irreversible binding of the probe. The filters were exposed to X-OMAT AR (Kodak) film overnight.

The autoradiograph results from the dot-blot analysis of λgt10 selected cDNAs from YAC A1220 probed with TIMP-1 cDNA (Figure 4.12A), demonstrated a dramatic enrichment of the positive control reporter gene TIMP-1 in the secondary selected cDNAs after two rounds of enrichment. Similar results were also observed with the λZAPII selected cDNAs from the same YAC (data not shown). Dot-blot analysis of the hybridisation of the rhodopsin gene with the λZAPII and λgt10 selected cDNAs from the rhodopsin YAC indicated that this abundant retinal gene has been selected and enriched for (Figure 4.12B). The presence of the rhodopsin gene in the λZAPII and λgt10 starting cDNA libraries was evident by the faint hybridisation signals obtained with the rhodopsin gene probe. The primary selected λZAPII and λgt10 cDNAs also showed faint signals in comparison to the strong signals in the λZAPII and λgt10 secondary selected cDNAs, indicating a good enrichment for this retinal gene. However the TIMP-1 gene showed greater enrichment compared to the rhodopsin gene since it gave no hybridisation signal with the starting cDNA libraries, indicating the low abundance of this gene in the retinal libraries used.

There was no apparent increase of β-actin cDNA hybridisation (Figure 4.12C) in the λgt10 primary and secondary selected cDNAs from the YAC A1220 compared to the λgt10 starting cDNAs. Instead there was a small decrease in the λgt10 secondary selected cDNAs, indicated by a fainter hybridisation signal with the β-actin cDNA probe. This showed a selective loss of the negative-control reporter gene β-actin after the second round of enrichment. There was no increase of ribosomal sequence transcripts in the rhodopsin YAC λgt10 primary selected cDNAs in comparison to the starting λgt10 cDNAs on dot-blot analysis (Figure 4.12D). However, there was a decreased abundance of ribosomal cDNAs (less signal) in the λgt10 secondary selected cDNAs from the same YAC, indicating suppression of ribosomal sequences. The depletion of yeast contaminating transcripts in the primary and secondary selected cDNAs, on dot-blot analysis was similar to that discussed for the ribosomal sequences, showing no increase in the selected cDNA material when compared to the starting cDNAs (data not shown).

Southern Blot analysis (section 4.2.6.2) was also used to monitor the selection and enrichment of internal control genes and depletion of non-specific transcripts. In contrast to dot-blot analysis, the approximately equal amounts of PCR amplified DNA present in the Southern blots were visualised on ethidium bromide stained gels (Figures 4.13A,
4. cDNA selection for enriched transcripts from genomic clones

Figure 4.12 Qualitative evaluation for enrichment and depletion of cDNA transcripts.

The relative abundance of specific transcripts of both the TIMP-1 and rhodopsin genes determined in the starting cDNAs and primary and secondary selected cDNAs are shown in (A) and (B). Equal amounts of PCR-amplified cDNA inserts from the starting library and the enriched sublibraries for each of the YACs were dot-blotted onto membrane filters and probed with the known retinal cDNAs present in these YACs. Filters were washed to a stringency of 0.5 X SSC, 0.1% SDS at 65°C.

(A) Hybridisation of the dot-blotted YAC A1220 λgt10 selected cDNAs with the 610 bp TIMP-1 cDNA fragment showed no hybridisation signals in the λgt10 starting library and primary selected cDNAs. The strong signals present in the λgt10 secondary selected cDNAs demonstrated a dramatic enrichment for the positive control reporter gene TIMP-1.

(B) Dot-blot results of hybridisation of the rhodopsin cDNA fragment (200 bp) to the Rhodopsin YAC λZAPII and λgt10 selected cDNAs. Faint hybridisation signals were obtained with the λZAPII and λgt10 starting cDNAs and primary selected cDNAs in comparison to the stronger signals present in the secondary selected cDNAs of these libraries. This indicated a good enrichment for the rhodopsin gene.
Assessment of depletion of non-specific transcripts of β-actin and ribosomal DNAs determined in the starting library, primary and secondary selected cDNAs. Preparation and use of dot-blot filters were as described in (A) and (B).

(C) Hybridisation results of the β-actin cDNA onto YAC A1220 selected cDNAs showed no apparent increase of β-actin cDNA transcripts in the λgt10 primary and secondary selected cDNAs compared to the starting cDNAs. Instead there is a fainter signal and hence a small decrease in the λgt10 secondary selected cDNAs which showed a selective loss of the negative control reporter gene β-actin after the second round of enrichment.

(D) Dot-blot analysis showed no increase of ribosomal sequence transcripts for the rhodopsin YAC λgt10 primary selected cDNAs in comparison to the starting λgt10 cDNAs. However the decreased abundance of ribosomal cDNAs (smaller signal) in the λgt10 secondary selected cDNAs of the same YAC indicated some suppression of ribosomal sequences.
4. cDNA selection for enriched transcripts from genomic clones

4.14A & 4.15A) and then compared to the autoradiographs which showed hybridisation signals produced by the radiolabelled control and contaminant DNAs (Figures 4.13B, 4.14B & 4.15B). This helped both in gaining a qualitative assessment of the degree of enrichment and depletion of eluted cDNAs and also contributed to the evaluation of size distribution of the selected cDNA material.

Autoradiograph results (Figure 4.13B) show the presence of TIMP-1 transcripts clearly in the λgt10 secondary selected cDNAs from YACs A1220 and C1228, and its absence in the 33CA11 YAC secondary selected cDNAs. The striking enrichment for the rare TIMP-1 gene in both the YACs A1220 (Figure 4.13B i, Lane 3) and C1228 (Figure 4.13B ii, Lane 9) secondary selected material is also distinctly evident from the strong hybridisation signals with the TIMP-1 cDNA probe, in contrast to the lack of hybridisation signals in the starting λgt10 cDNAs (Figure 4.13B, Lanes 1& 7) and primary selected cDNAs (Figure 4.13B, Lanes 2&8) from the same YACs.

Ethidium bromide stained gels (Figures 4.13A) show equal amounts of cDNAs from the λgt10 starting cDNA library, primary selected cDNAs and secondary selected cDNAs from YACs A1220, C1228 and 33CA11 used in the Southern blot analysis. These blots allowed qualitative assessment of enrichment and depletion of cDNA transcripts. Although the size distribution of the λgt10 selected cDNAs from YACs A1220, C1228 and 33CA11 after one and two rounds of direct selection ranged from 200bp to 4kbp (Figure 4.13A) in length, the representation of the TIMP-1 gene in the secondary selected material from the YACs A1220 and C1228 (Figure 4.13B, Lanes 3 and 9 respectively) was found in a smaller average size range of 200bp to 600bp, indicating that the direct selection procedure preferentially selected for smaller sized cDNA transcripts.

Dramatic enrichment was also observed for the rhodopsin gene transcripts from the λZAPIII secondary selected cDNAs (Figure 4.14B, Lane 3) from the rhodopsin YAC. This was clearly evident from autoradiograph results which show strong hybridisation signals with the rhodopsin cDNA probe, in contrast to the lack of hybridisation signals in the starting λZAPIII cDNAs (Figure 4.14B, Lane 1) and primary selected λZAPIII cDNAs (Figure 4.14B, Lane 2) from the same YAC. These results were similar to that observed in Figure 4.13B for the enrichment of the TIMP-1 gene transcripts. In contrast the rhodopsin gene was found to be less enriched in the λgt10 secondary selected cDNAs (Figure 4.14B, Lane 9), since the strength of the hybridisation signal obtained here was similar to that present in the λgt10 starting cDNAs (Figure 4.14B, Lane 7). Also the abundance of the rhodopsin gene transcripts seems to be reduced in the λgt10 primary selected cDNAs (Figure 4.14B, Lane 8) as compared to the λgt10 starting cDNAs. This indicated some level of abundance normalisation (Morgan, 1992; Lovett, 1994).

Ethidium bromide stained gels (Figures 4.14A) show equal amounts of cDNAs from the λZAPIII and λgt10 starting cDNAs, primary selected cDNAs and secondary selected cDNAs from the rhodopsin YAC used in the Southern blot analysis. Although
4. cDNA selection for enriched transcripts from genomic clones

![Image of gel and Southern blot analysis](image)

**Figure 4.13 Monitoring enrichment of the internal control TIMP-1 cDNA using Southern blot hybridisation analysis.**

Qualitative assessment for enrichment and evaluation of size distribution of selected TIMP-1 cDNA transcripts as shown in (A) and (B).

(A) Ethidium bromide stained gels (i), (ii) and (iii) show equal amounts of PCR amplified cDNA inserts from the λgt10 starting, primary and secondary selected sub-libraries for the A1220 (i), C1228 (ii) and 33CA11 (iii) YACs. The size distribution of these YAC specific λgt10 selected cDNAs after one and two rounds of direct selection ranged from 200bp to 4kbp. The no DNA results of Lanes 4, 10 & 16 indicated absence of any contaminating DNAs. The ΦX174/ HaeIII and λ/HindIII standard molecular weight markers (M.W.M) were used as size comparisons.

(B) Hybridisation results of Southern blots of gels in (A) probed with the 610 bp TIMP-1 cDNA fragment. Filters were washed to a stringency of 0.2 X SSC, 0.1% SDS at 65°C. The presence of TIMP-1 gene is clearly seen in the A1220 and C1228 YACs secondary selected cDNAs (Lanes 3 & 9 respectively) but not in the 33CA11 YAC secondary selected cDNAs (Lane 15). A striking enrichment for the rare TIMP-1 gene was also obtained in both the YACs A1220 (Lane 3) and C1228 (Lane 9) secondary selected material as indicated by the strong hybridisation signals, in contrast to the lack of any signals in the λgt10 starting cDNAs (Lanes 1 & 7) and primary selected cDNAs (Lanes 2 & 8) from the same YACs. TIMP-1 cDNA transcripts in YACs A1220 and C1228 secondary selected cDNAs were present in a size range of 200bp-600bp, indicating preferential selection of smaller sized cDNA transcripts during direct cDNA selection.
4. cDNA selection for enriched transcripts from genomic clones

Figure 4.14 Monitoring selection of the external control Rhodopsin cDNA using Southern blot hybridisation analysis.

Qualitative assessment of selection and size distribution of rhodopsin cDNAs as shown in (A) and (B).

(A) Ethidium bromide stained gels showing equal amounts of PCR amplified cDNA inserts from the λZAPII and λgt10 starting, primary and secondary selected cDNAs from the rhodopsin YAC. The size distribution of the λZAPII primary and secondary selected cDNAs ranged from 200 bp to 2 kbp (Lanes 2 & 3). The λgt10 secondary selected cDNAs were present in a size range of 200bp - 1kbp (Lane 9), while the λgt10 starting and primary selected cDNAs (Lanes 7&8) were present in a larger size range of 500bp-4 kbp. The no DNA tracks (Lanes 4 & 10) indicated no contaminating DNAs. The Ox174/HaeIII and λ/HindIII molecular weight markers (M.W.M) were used as size comparisons (Lanes 5, 6, 11 & 12).

(B) Hybridisation results of Southern blots of gels in (A) with the 220 bp rhodopsin cDNA fragment. Filters were washed to a stringency of 0.2 X SSC, 0.1% SDS at 65°C. The abundant retinal gene rhodopsin was selected for in both the λZAPII and the λgt10 secondary selected cDNAs (Lanes 3 & 9 respectively) and present in a size range of 200bp-1kb indicating preferential selection for smaller sized cDNA transcripts after the second round of selection. Comparatively greater enrichment was seen for the Rhodopsin gene transcripts in the λZAPII secondary selected cDNAs (Lane 3) in contrast to that present in the λgt10 secondary selected cDNAs (Lane 9) because of a relatively greater representation of the rhodopsin gene in the λgt10 starting cDNA library (Lane 7) as compared to the λZAPII starting cDNA library (Lane 1). Normalisation for the abundant gene rhodopsin was observed, since it was less abundant in the λgt10 primary and secondary selected cDNAs (Lanes 8 & 9) as compared to the λgt10 starting cDNAs (Lane 7).
the size distribution of the λZAPII selected cDNAs from the rhodopsin YAC after one and two rounds of direct selection ranged from 200bp to 2kbp (Figure 4.14A, Lanes 2&3), hybridisation results showed the rhodopsin gene transcripts in the λZAPII secondary selected material were in a smaller average size range of 200bp to 1kbp (Figure 4.14B, Lane 3). This was also clearly seen in the λgt10 starting and primary selected cDNAs (Figure 4.14B, Lanes 7&8) of 800bp to 4kbp as compared to relatively smaller size range of 200bp to 1kbp from the λgt10 secondary selected cDNAs (Figure 4.14B, Lane 9). This clearly demonstrated that after the second round of selection smaller sized cDNA transcripts were preferentially selected for. These hybridisation results correlated with the ethidium bromide gel results (Figure 4.14A, Lanes 7, 8 & 9) showing the selection of λgt10 secondary selected cDNAs in a size range of 200bp-1kbp compared to the larger size range present for the λgt10 starting and primary selected cDNAs of 500bp-4kbp. The selection of rhodopsin gene transcripts in the λZAPII and λgt10 secondary selected cDNAs in the size range 200bp-1kbp showed specific binding of these selected transcripts to similarly sized genomic target DNAs and indicated that solution hybridisation of these two components had occurred effectively.

The β-actin cDNA, (a negative control gene), was used to determine the depletion of non-selected transcripts such as abundant housekeeping genes from the selected cDNAs. The autoradiograph results shown in Figure 4.15B indicated a great reduction in the hybridisation of β-actin cDNAs in the λgt10 primary selected cDNAs (Figure 4.15B, Lane 7) from YAC 4HG2 and a further depletion in the λgt10 secondary selected cDNAs (Figure 4.15B, Lane 8) from the same YAC, as compared to the strong signal observed in the λgt10 starting library (Figure 4.15B, Lane 6). The ethidium bromide stained gel (Figure 4.15A) again showed equal quantities of λgt10 starting, primary and secondary selected cDNAs from the 4HG2 YAC used in Southern blot hybridisations and qualitative analysis of non-selected transcripts. Similar depletion of yeast and ribosomal contaminating transcripts was observed in the primary and secondary selected cDNAs for each of the YACs used on Southern blot hybridisation analysis (data not shown).

4.2.6.4 Evaluation of enrichment of selected transcripts and depletion of non-selected transcripts.

After two rounds of enrichment, dot blot and Southern blot analysis of the selected cDNA material indicated that the TIMP-1 gene was confirmed present in the secondary selected cDNAs from the YACs A1220 and C1228 and absent in the 33CA11 YAC. This correlated with the PCR results which showed the absence of TIMP-1 in YAC 33CA11 but its presence in the YACs A1220 and C1228. TIMP-1 was dramatically more abundant in the second round selected cDNA material as compared to the original cDNA libraries and the primary selected cDNAs indicating enrichment and the
4. cDNA selection for enriched transcripts from genomic clones

Figure 4.15 Monitoring depletion of the negative control β-actin cDNAs using Southern blot hybridisation analysis.

Qualitative assessment for reduction of unwanted transcripts of the β-actin housekeeping gene as shown in (A) and (B).

(A) The ethidium bromide stained gel show equal amounts of PCR amplified cDNA inserts from the λgt10 starting, primary and secondary selected cDNAs (Lanes 1, 2 & 3) from the YAC 4HG2. The no DNA track (Lane 4) showed absence of any contaminating DNAs. The ØX174/HaeIII molecular weight marker (M.W.M) was used as a size comparison (Lane 5).

(B) Hybridisation results of Southern blotted gel (A) with a β-actin cDNA fragment. The filter was washed to a stringency of 0.2 X SSC, 0.1% SDS at 65°C. Non-selected transcripts of β-actin cDNAs were greatly depleted as shown by the faint signal in the secondary selected cDNAs (Lane 8) as compared to the strong signal observed in the λgt10 starting cDNAs (Lane 6). The second round of selection was essential, since primary selected cDNAs (Lane 7) were not as reduced in abundance for the β-actin cDNA transcripts as compared to the secondary selected cDNAs (Lane 8). Pre-blocking using quenching agents therefore produced effective depletion of contaminating β-actin cDNAs which were not selected.
4. cDNA selection for enriched transcripts from genomic clones

successful application of the direct cDNA selection technique (Figures 4.12A & 4.13A&B). These results also showed that a second round of selection was essential since the first round of selection showed no enrichment for the rare retinal gene TIMP-1. The more abundant retinal gene rhodopsin was selected for in both the \( \lambda ZAPII \) and the \( \lambda g t 10 \) secondary selected cDNAs. A second round of selection was necessary for enrichment of the rhodopsin gene transcripts, since they were present in the \( \lambda ZAPII \) secondary selected cDNAs but absent in the \( \lambda ZAPII \) primary selected cDNAs. Comparatively greater enrichment was seen for the rhodopsin gene transcripts in the \( \lambda ZAPII \) secondary selected cDNAs (Figure 4.14B, Lane 3) as compared to that present in the \( \lambda g t 10 \) secondary selected cDNAs (Figure 4.14B, Lane 9). This was because of a relatively greater representation of the rhodopsin gene in the \( \lambda g t 10 \) starting cDNA library as compared to that present in the \( \lambda ZAPII \) starting cDNA library, possibly due to the differences in the cDNA synthesis protocols used for the preparation of these two retinal libraries. The direct selection procedure also showed normalisation for the abundant gene rhodopsin, since it was less abundant in the \( \lambda g t 10 \) primary and secondary selected cDNAs (Figure 4.14B, Lanes 8 & 9) as compared to the \( \lambda g t 10 \) starting cDNAs (Figure 4.14B, Lane 7).

Non-selected transcripts of ribosomal, yeast and \( \beta \)-actin cDNAs were found to be greatly depleted in the secondary selected cDNAs isolated from the YACs. The second round of selection was therefore essential, since primary selected cDNAs were not as reduced in abundance for these transcripts, this was more obvious from the Southern blot analysis (Figure 4.15B) as compared to the dot-blot analysis (Figure 4.12C). Pre-blocking using the quenching agents as described in sections 4.2.4.1 & 4.2.5.1 seems to have been effective, demonstrating a depletion of these non-selected, contaminating transcripts.

4.3 Discussion

Comparison of PCR products obtained from unbound, wash solutions and eluted DNAs after first round (section 4.2.4) and second round (section 4.2.5) direct cDNA selection indicated a decrease in the abundant and/or common contaminating cDNAs at a gross level. This was evident from the more uniform smears of PCR amplified products obtained from the eluted secondary selected cDNAs as compared to the primary selected and starting cDNAs. The reduced background in the selected material also indicated a decrease in sequence complexity as compared to the starting libraries. This probably indicates that representation of certain common elements such as \( Alu \), LINE1 repeat sequences, ribosomal cDNAs and abundant cDNAs such as \( \beta \)-actin cDNAs found in specific size ranges such as 300bp and 600bp (indicated by greater intensity bands) have been eliminated to a certain degree after the first round and more so in the second round.
of selection. The selection of secondary selected cDNAs in a size range of 200bp-800bp showed specific binding of selected cDNAs to similarly sized genomic target DNAs and indicated that solution hybridisation of selected cDNAs to genomic target DNA had occurred effectively.

No discrete bands were seen in either the primary or secondary selected cDNAs. Such discrete bands are usually artifacts, but if the contig is short they may be real. The PCR products obtained from unbound DNAs were short fragment cDNAs which had not bound to target genomic DNAs, for example contaminant clones such as repeat elements and ribosomal cDNAs which had been pre-blocked in the genomic DNA but were present in the λZAPII and λgt10 cDNA libraries. The PCR products obtained from the DNA present in wash solutions could be non-specific cDNAs, which had loosely bound to the target genomic DNA but had been removed after washing at increased stringency.

PCR amplifications of the eluted material from the primary and secondary selections could be studied to check the length distribution and sequence complexity of the inserts, in comparison to those in the starting libraries. This may require modification of PCR annealing temperatures, extension times, Mg$^{2+}$ concentration or salt concentration in some cases (Innis and Gelfand, 1990). However some problems cannot be resolved and represent inherent limitations in the method. For example, some sequence bias in PCR may occur so that specific DNA sequences are preferentially amplified and others are poorly amplified. Short exons can also present a problem because they may not hybridise well to their homologous cDNA in the selection and thus adequate enrichment may not occur.

It is possible to attribute the apparent decrease in the level of abundance of the retinal gene rhodopsin in both the λgt10 primary and secondary selected cDNAs, to reflect some level of abundance normalisation. Thus the first round of selection was conducted using an excess of genomic DNA; this was not the case for the secondary round, (Lovett, 1994). This resulted in large increases in the relative abundance of some cDNA species during the first round of selection, but during the second round insufficient genomic target was present to capture all of these abundant cDNAs. Therefore the abundantly selected cDNAs such as the rhodopsin gene were reduced in abundance on secondary selection, and less abundant cDNAs such as the TIMP-1 gene were increased in relative abundance on secondary selection.

Although the depletion of contaminating transcripts such as ribosomal, yeast and β-actin cDNAs was ascertained qualitatively by dot-blot and Southern blot analysis, a human placental DNA probe was not used in these hybridisation studies, to assess the depletion of repetitive sequences such as Alu repeats. This was because such Alu repeat elements were found to be present within certain selected cDNA clones, so it is difficult to determine whether such repeats are present as spurious contaminants or as selected
4. cDNA selection for enriched transcripts from genomic clones

clones. The depletion of such repetitive elements was more accurately monitored by analysing for their presence or absence within individual clones of cDNAs from the enriched sub-libraries as described in section 5.4.1.1.

Although there was a high level of enrichment for both the TIMP-1 gene transcripts in the λgt10 secondary selected cDNAs (Figure 4.13B, Lanes 3 & 9) and the rhodopsin gene transcripts in the λZAPII secondary selected cDNAs (Figure 4.14B, Lane 3) the exact level of enrichment could only be determined by screening the cloned sub-libraries of the enriched cDNAs with these control cDNA probes. Hence quantitative analysis for the enrichment and depletion of selected cDNAs was determined by screening the cloned sub-libraries of the enriched cDNAs with control and contaminant DNA probes (section 5.4) and comparing it with the representation of these cDNAs in the starting cDNA libraries (section 5.2). This makes it possible to correlate the enrichment and depletion data derived from starting libraries and selected mini-library screening experiments with the gross qualitative assessment made here through dot-blot and Southern blot analysis.

The screening of the enriched cDNAs using dot-blots and Southern blots was essential for qualitative monitoring of the level of enrichment of the control cDNAs and the level of depletion of the contaminant cDNAs such as β-actin, yeast and ribosomal cDNAs, before proceeding to clone and analyse the cDNAs in the enriched sub-libraries. If the selection had not been successful and enrichment of the low abundant control cDNA such as TIMP-1 was not seen after dot-blot or Southern blot analysis, a third round of selection could have been undertaken, with the amount of quenching agents increased to improve the enrichment process.

4.4 Conclusion

The direct cDNA selection procedure was successfully established, as indicated initially at a gross level by a decrease in the abundant and/or contaminating cDNAs as observed by the reduced background, and a decrease in the sequence complexity of the amplified selected cDNAs as compared to the starting cDNAs.

Dot blot and Southern blot analysis and comparisons of the selected material and starting cDNAs corroborated this by demonstrating a dramatic level of enrichment for the rare retinal internal control gene TIMP-1, after the second round of selection in the secondary selected cDNAs. The abundant retinal gene rhodopsin was enriched in the λZAPII selected cDNAs, but was abundance normalised in the λgt10 selected cDNAs. The hybridisation analysis was also effective in qualitatively monitoring the level of depletion of the contaminant transcripts such as β-actin, yeast and ribosomal cDNAs in the enriched selected cDNAs.
CHAPTER FIVE

Analysis of selected sublibrary clones
CHAPTER FIVE

ANALYSIS OF SELECTED SUB-LIBRARY CLONES

5.1 Introduction

The analysis of region-specific cDNA sublibraries represents a simple, rapid and efficient tool for the generation of a regional transcriptional map covering a 500 kb region around the DXS426 locus found in the Xp11.23 section of the XLRP2 critical region. PCR products from the second-round selected λgt10 cDNAs for each YAC were cloned into the pDIRECT plasmid vector using the PCR-Direct Cloning system (Clontech, USA) for the preparation of YAC specific, enriched cDNA sublibraries. A random set of about 200-250 clones from each YAC sub-library were then selected for cDNA content analysis. This analysis involved hybridisation of arrayed cDNA clones to confirm the presence of control reporter genes to quantitatively evaluate the enrichment and depletion levels of cDNAs in comparison with levels in the starting cDNA libraries. A representation of 200,000 phage clones in the original retinal cDNA libraries was therefore plated and probed with rhodopsin and TIMP-1 cDNAs in order to determine the level of enrichment.

Hybridisation analysis was carried out in combination with sequencing, which was performed to determine overlap of cDNAs, their coding potential and to allow database searching for identity and similarity to previously sequenced cDNAs. Short cDNAs were fully sequenced while longer cDNAs were partially sequenced from both ends using vector primers. Sequence analysis was performed on 200-300 bp of sequence data for each clone using the Gapped BLAST program (Altschul et al., 1997). PCR primers derived from these sequences were then used in mapping experiments.

5.2 Pre-screening λZAPII and λgt10 starting cDNA libraries with internal control genes rhodopsin and TIMP-1

In order to quantitate the extent of enrichment of the positive control genes rhodopsin and TIMP-1, 200,000 plaques for each of the λZAPII and λgt10 (combined ≥ 1kb and ≤ 1kb libraries) were plated at a density of 100,000 pfu/plate (20cm X 20cm). The filters were lifted from each of the plates and prepared for hybridisation as described in section 2.4.5. Duplicate filters were probed with rhodopsin or TIMP-1 cDNAs and the
abundance of clones estimated by positive hybridisation signals. The relative abundance of specific rhodopsin and TIMP-1 cDNAs clones in the selected YAC sub-libraries were compared to evaluate the level of enrichment (section 5.4.2).

The starting adult retinal cDNA libraries were shown to contain both the two internal control genes TIMP-1 and PFC, and the external control gene rhodopsin by PCR analysis as described in section 4.2.1.2. This was confirmed by hybridisation analysis which gave quantitative results on the level of representation of these genes in the starting libraries. Only one TIMP-1 cDNA clone out of 200,000 was found to be present in the starting λgt10 cDNA library as indicated by the single positive hybridisation signal shown in Figure 5.1A. Since the TIMP-1 gene was rare in the starting library, it was used to determine the efficiency of the enrichment strategy. The rhodopsin gene was found to be abundant as expected in the original library comprising 100 out of 100,000 clones in the λZAPII cDNA library and 600 out of 100,000 clones in the λgt10 cDNA library as shown by the large number of positive hybridisation signals in Figure 5.1B. The presence of about six times more rhodopsin clones in the λgt10 cDNA library as compared to the λZAPII cDNA library could be due to the difference in the methodologies used in their construction.

As only one TIMP-1 transcript was found to be present in a total of 200,000 cDNA clones screened in the original λgt10 cDNA libraries, to obtain a 1,000-fold enrichment one would expect 1 in 200 clones to contain the TIMP-1 gene from the the YACs A1220 and C1228 second round selected cDNA material. To further confirm TIMP-1 and rhodopsin selection and to determine their level of enrichment, sub-library screening was performed for YACs A1220, C1228 and the rhodopsin YAC secondary selected cDNAs (section 5.4.2).

5.3 Construction of YAC specific cDNA sub-libraries from secondary selected cDNAs

The Clontech PCR-Direct Cloning System (Clontech, USA) provides a method for rapid, highly efficient, directional cloning of PCR products into an all-purpose cloning vector, pDIRECT (Figure 5.2A). Once a gene of interest has been cloned, transcripts can be produced using the T7 and T3 promoters which flank the cloning site.

5.3.1 Cloning of secondary selected cDNAs

Using the PCR-Direct Cloning kit (Clontech, USA), PCR products from the secondary selected cDNAs were directionally cloned into the linearised pDIRECT plasmid (Figure 5.2B). The two 5' ends of the pDIRECT vector have different PCR-Direct
Figure 5.1 Prescreening the λgt10 starting adult retinal cDNA library.

Representative autoradiograms from prescreening experiments of the plated λgt10 retinal cDNA library with the TIMP-1 and rhodopsin cDNA probes. Two filters containing a total of ~200,000 bacteriophage pfu were used to screen each cDNA. Filters were washed to a stringency of 0.2 X SSC, 0.1% SDS at 65°C.

(A) Hybridisation of the TIMP-1 cDNA to a set of duplicate filters containing ~100,000 pfu of λgt10 retinal cDNAs indicates one TIMP-1 cDNA transcript identical in the replica filter (not shown). The second set of duplicate library filters containing another ~100,000 pfu of λgt10 retinal cDNAs showed absence of any hybridisation signals (hybridisation data not shown).
Figure 5.1B

Hybridisation of the rhodopsin cDNA to the same set of duplicate filters as described in Figure 5.1A containing ~100,000 retinal cDNAs. A total of 1,200 rhodopsin gene transcripts were scored as positives from screening two sets of duplicate λgt10 cDNA library filters. These positive signals were identical in the replica filter (not shown).
Figure 5.2A  Schematic map of the pDIRECT vector.

The linearised pDIRECT vector contains non-complementary 5' single stranded ends (11 nucleotides long on one side and 12 nucleotides long on the other) for cloning PCR products. The cloning site and the multiple restriction enzyme sites flanking it are located in the middle of the lacZ gene, allowing for blue/white colour screening for recombinant plasmids. T3 and T7 promoters flank the cloning site in opposite orientations.
5. Analysis of selected sub-library clones

PCR products from the secondary selected cDNAs are directionally cloned into the linearised pDIRECT plasmid. The vector cannot recircularise by itself since it contains non-complementary 5' single-stranded ends which practically eliminates non-recombinant plasmids.

Figure 5.2B Schematic diagram of the PCR-Direct Cloning System
single stranded tails, 11 and 12 nucleotides long respectively (Figure 5.2A), which are not complementary to each other. Thus the vector cannot re-circularise by itself and non-recombinant plasmids are practically eliminated.

5.3.1.1 PCR primer design for PCR-Direct cloning and amplification of cDNA inserts

The cDNA fragments to be cloned were first amplified by PCR using primers with PCR-Direct sequences (12 and 13 nucleotides long) at their respective 5' ends. The PCR-Direct sequences incorporated into the primers are complementary to the vector PCR-Direct sequences, thus allowing for directional cloning in the pDIRECT vector (Figure 5.2B). The DNA sequences for the various amplification primers used were as follows:

**λZAPII SK-sense primer:**
5' - CTC GCT CGC CCA - TCT AGA ACT ACT GGA TC - 3'
* pDIRECT vector sequence - λZAPII SK primer sequence

**λZAPII KS-Antisense primer:**
5' - CTG GTT CGG CCC A - CGA GGT CGA CGG TAT CG - 3'
* pDIRECT vector sequence - λZAPII KS primer sequence

**λgt10 Forward-sense primer:**
5' - CTC GCT CGC CCA - AGC CTG GTT AAG TCC AAG CTG - 3'
* pDIRECT vector sequence - λgt10 Forward primer sequence (Set B)

**λgt10 Reverse-antisense primer:**
5' - CTG GTT CGG CCC A - CTT CCA GGG TAA AAA GCA AAA G-3'
* pDIRECT vector sequence - λgt10 Reverse primer sequence (Set B)

Eluted specific cDNAs from the second round of selection were PCR amplified using 10µl from each sub-library as templates in 100µl PCR reactions. Reactions were performed in a 9600 Perkin Elmer PCR machine and thermal cycling was performed for 30 cycles as described in section 4.2.1.3. The reaction products of the PCR amplified cDNAs were analysed on a 1.2% agarose gel using ethidium bromide visualisation. The amplified PCR products obtained were similar to the λZAPII secondary selected cDNAs from YAC A1220 as described in section 4.2.5.4 (Figure 4.11C, Lane 7) and the λgt10 secondary selected cDNAs from YAC A1220 as shown in Figure 4.13A (Lane 3) of section 4.2.6.3.
5.3.1.2 Purification of PCR product using CHROMA SPIN Columns

The PCR products to be cloned were purified by gel filtration chromatography using CHROMA SPIN + TE-100 columns provided in the PCR DIRECT Cloning kit (Clontech) according to the manufacturer's instructions. The concentration of the purified DNA was estimated after analysis of a small sample of the purified product on a 1.2% agarose gel as described in section 2.1.3.

5.3.1.3 Treatment of PCR products with T4 DNA polymerase

The purified PCR products were treated with T4 DNA polymerase in the presence of dTTP; the 3'→5' exonuclease activity of this enzyme removed 11 and 12 nucleotides at the respective 3' ends of the PCR products, resulting in DNA fragments with 5' extended single-stranded tails of defined PCR-Direct sequence at both ends as shown in Figure 5.2B. The presence of an excess amount of dTTP in the reaction mixture ensured that the exonuclease activity will not remove any base beyond the first thymidine that is encountered, since the polymerase activity will replace the dTTP as quickly as it is removed. The PCR-Direct sequences for the upstream and downstream primers were not complementary to each other; thus PCR products with 5' single-stranded ends will not form concatamers.

For each PCR sample, the following components were added to an eppendorf tube:

<table>
<thead>
<tr>
<th>Reagents</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purified PCR product (50-100ng)</td>
<td>1-7 μl</td>
</tr>
<tr>
<td>10 X T4 DNA Polymerase buffer</td>
<td>1 μl</td>
</tr>
<tr>
<td>25 mM dTTP</td>
<td>1 μl</td>
</tr>
<tr>
<td>T4 DNA Polymerase</td>
<td>1 μl</td>
</tr>
</tbody>
</table>

Total volume adjusted with sterile water to 10 μl

The contents of the eppendorf tube were then mixed and centrifuged briefly, and incubated at 37°C for 30 minutes. 2μl of SE solution (2.5 M NaCl, 100 mM EDTA) was then added to the tube, mixed well and centrifuged briefly in order to bring the contents to the bottom of the tube. The T4 DNA polymerase was then heat-inactivated by incubating the tube at 75°C for 15 minutes. The first time this protocol was performed, a parallel PCR-Direct cloning experiment using the control PCR-amplified DNA fragment (mouse β-actin DNA) included in the kit was used to monitor, a) the successful incorporation of PCR-Direct sequences in the cDNA-specific primers; b) the successful
5. Analysis of selected sub-library clones

generation and purification of cDNA-specific PCR products and c) the efficiency of cloning the insert (amplified DNA fragments) into the pDIRECT vector.

5.3.1.4 Hybridisation of PCR insert to pDIRECT vector

Before bacterial transformation, the single-stranded tails which are complementary between the vector and insert DNA (Figure 5.2B) were allowed to anneal. This was done by adding 1µl of 50ng/µl linear pDIRECT vector to each microcentrifuge tube containing the insert DNA. The contents of the tube were then mixed, centrifuged briefly and allowed to incubate at room temperature for at least one hour prior to transformation. Since the 5' overhanging sticky ends are GC-rich, the hybrids formed by the annealing of the insert DNA to the vector were stable enough to be used directly to transform the bacterial host cells, thus circumventing the enzymatic ligation step.

5.3.2 Transformation of cloned cDNAs and blue/white screening for recombinant plasmids

The plasmid/insert DNA hybrid mixture (2-4 µl) was transformed as described in section 2.4.2.2. Since the efficiency of transformations can vary between experiments the supercoiled pUC 19 DNA provided in the kit was used each time as a positive control. The DH5α (LacZ') bacterial host cells were plated on LB agar containing ampicillin (50µg/ml) for plasmid selection and X-gal and IPTG to screen for colonies receiving a recombinant plasmid. The cloning site in the pDIRECT vector (Figure 5.2A) is located in the middle of the plasmid's lacZα gene, hence if no insert is present, functional β-galactosidase is produced and the transformed bacterial colony was blue. However when the host cell received a recombinant plasmid (i.e., one containing a DNA insert or deletion in the lacZα gene), the resulting transformed colony was white (LacZ').

5.3.3 Transfer of recombinant clones onto master plates, glycerol stocks and preparation of sub-library filters

About 200-250 white colonies from transformations of each secondary selected cDNA sub-library for each of the five different YACs were randomly picked, arrayed and grown on ampicillin plates for ordering and storage, these being master plates. The same colonies were also picked and transferred into 96-well microtiter plates containing LB bacterial media and ampicillin. After overnight growth at 37°C an equal volume of 50% glycerol was added to each of the wells and the resultant 25% glycerol stocks were then stored at -80°C. The clones from the Master plates were re-streaked in formatted
clone arrays on five L-agar ampicillin plates (~50 clones per plate). Colonies were transferred onto Hybond N+ membranes, which were then denatured, neutralised and baked (section 2.4.5) to produce multiple copies of sub-library filters for further hybridisation analysis (section 5.4).

5.4 Sub-library filter hybridisation screening for quantitative evaluation of secondary selected libraries

The duplicate filters for each of the secondary selected sub-libraries were screened to determine the sensitivity of the direct cDNA selection procedure and the abundance levels of various cDNAs. The filters were separately hybridised with DNA probes of repeat sequences, yeast, β-actin, ribosomal sequences and positive control genes, TIMP-1, properdin and rhodopsin cDNAs for depletion and enrichment studies respectively. All DNA probes were radiolabelled with 32P using a random-priming kit (Pharmacia, section 2.3.2). Prehybridisation, hybridisation and washing conditions were conducted under standard conditions (section 2.3.5). Duplicate positive signals were counted and the results are presented in Table 5.1.

Recombinant cDNA clones which failed to hybridise to these control and contaminating probes were isolated and analysed further by sequencing (section 5.5.2) while positive clones were excluded from sequence analysis. The combined percentages of the contaminating and enriched sequences present in each of the YAC secondary selected cDNA sub-libraries are described in Table 5.1. These results are discussed further in sections 5.4.1 and 5.4.2.

<table>
<thead>
<tr>
<th>Table 5.1 Combined percentages of the contaminating and enriched sequences of the YAC secondary selected cDNA sub-libraries</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
</tr>
<tr>
<td>1%</td>
</tr>
<tr>
<td>2%</td>
</tr>
<tr>
<td>6%</td>
</tr>
<tr>
<td>90%</td>
</tr>
<tr>
<td>2%</td>
</tr>
<tr>
<td>1,000 fold</td>
</tr>
<tr>
<td>200 &amp; 70 fold</td>
</tr>
</tbody>
</table>
5.4.1 Assessing contaminating sequences

Duplicate filters of the selected sub-libraries were hybridised with radiolabelled total human genomic DNA to eliminate recombinant cDNAs that contained intermediate repeats, as well as probes for other known contaminants such as ribosomal, yeast and β-actin cDNAs for elimination from further analysis and for determining the percentage contamination after selection.

5.4.1.1 Probing with human placental DNA and LINE-1 DNA

To identify the percentage of repeat containing cDNA clones, duplicate filters of each of the four YAC (A1220, C1228, 33CA11 and 4HG2) selected sub-libraries were hybridised with labelled total human DNA; an example is shown in Figure 5.3A. In total 63 positive clones were identified, which corresponds to 6% of the 1018 cDNA clones screened. These clones were excluded from the sequence analysis. The blocking of repetitive elements during the cDNA selection procedure appeared to have been successful since the repetitive sequence elements did not appear to constitute a serious problem in the secondary selected sub-libraries. The figure of 6% of repeat sequences found is similar to the results reported by Fan et al., (1993) on application of the direct cDNA selection technique to regions of the human major histocompatibility complex (MHC). The 63 clones are probably low-copy repeats, which are never adequately quenched and can contribute up to 10 per cent of the enriched sub-library (Tagle et al., 1993). While intermediate repeat elements such as the Alu family are present in ~10% of cDNAs (Crampton et al., 1981), these are efficiently blocked by Cot-1 DNA. Lower copy number repeats however are not as efficiently blocked by Cot-1. The 6% of clones which showed positive hybridisation to human placental DNA, could be assumed not to be blocked in the selection procedure, although it is also possible that some of these cDNA fragments were selected based on their specific hybridisation to the 500kb region at Xp11.23, and that the repeat sequences in them were effectively suppressed by the blocking step.

5.4.1.2 Probing with ribosomal DNA

Ribosomal clones represent a common contaminant in most cDNA libraries. These cDNAs are probably selected because of small quantities of contaminating yeast ribosomal DNA in the gel purified YAC DNA preparations (section 4.1.3). In selections where ribosomal sequences are not pre-blocked, ribosomal cDNAs can comprise up to 70% of selected material (Morgan et al., 1992). When the cDNA...
Figure 5.3 Screening the YAC specific λgt10 secondary selected sub-libraries.
Representative autoradiograms of arrayed recombinant clones from YAC specific λgt10 secondary selected sub-libraries, hybridised separately with total human, TIMP-1 and Properdin DNAs. A total of ~200 recombinant plasmid clones were used in screening with each DNA probe. Filters were washed to a stringency of 0.2 X SSC, 0.1% SDS at 65°C.

(A) Hybridisation of duplicate filters of the 4HG2 YAC λgt10 secondary selected sublibrary with total human DNA indicates two repeat element positives identical in the replica filter. An estimated 6% of repeat sequences were detected from the total number of screened YAC specific λgt10 secondary selected sub-libraries.
5. Analysis of selected sub-library clones

The selection procedure was applied to the four different YACS after pre-blocking with the various human, yeast and *E. coli* ribosomal sequences (section 4.1.8) only 2% of the screened selected cDNA clones were found to be ribosomal sequences (hybridisation data not shown). Hence contaminating ribosomal genes were greatly reduced by 97% in selected cDNAs indicating that efficient competition had taken place.

5.4.1.3 Probing with yeast DNA

Studies on the conservation of homologous genes between yeast and man (Tugendreich et al., 1994) have shown that sub-regions of high sequence similarity were found within proteins of these two diverse organisms, which highlight domains within proteins that are of greatest structural and/or functional significance. While some of the estimated 80,000 genes in man will have counterparts with identical function in yeast, the vast majority will be related to yeast as members of a gene family that have analogous but not identical functions. These analogous yeast genes found within cDNA libraries can be selected by contaminating yeast DNA found in the YAC genomic DNA preparations.

Total yeast DNA from the YAC host yeast strain AB1380 (obtained commercially, Gibco BRL) was used as a radiolabelled probe, and 1% of the selected cDNAs were found to be positive (hybridisation data not shown). This negligible amount of yeast contamination indicated effective quenching of any co-selected yeast sequences by contaminating yeast genomic DNA.

5.4.1.4 Probing with negative control gene, β-actin cDNA

The β-actin cDNA was used as an internal negative control since it is a housekeeping gene which is represented in both the starting retinal cDNA libraries used, being a major constituent of the cytoskeleton of non-muscle cells such as retinal cells. β-actin cDNAs were found to comprise 1% of the clones in the enriched sub-libraries (hybridisation data not shown), therefore the abundant housekeeping gene β-actin has been selected against and depleted during the cDNA selection procedure.

5.4.2 Assessing for enrichment of internal positive control genes

To estimate the enrichment of the internal positive control cDNAs TIMP-1 and rhodopsin, 100 ng of these control probes were radiolabelled and hybridised to duplicate sets of filters from YACs A1220, C1228 and the rhodopsin cloned λgt10 secondary selected YAC sub-libraries. The number of positive clones obtained was then compared to the number of positive clones detected in the starting λgt10 cDNA libraries with the
same cDNA probes (section 5.2). This gave a quantitative comparison of these genes in starting and twice-enriched selected cDNA libraries. The properdin gene also found to be present in the starting retinal cDNA libraries (section 4.2.1.2) was assessed by hybridisation to YACs 33CA11 and 4HG2 λgt10 secondary selected cDNA sub-libraries.

5.4.2.1 Probing with TIMP-1 cDNA fragment

To gain a quantitative assessment of the degree of enrichment of the TIMP-1 gene, recombinant clones from the YACs A1220 and C1228 λgt10 secondary selected sub-libraries were hybridised with the TIMP-1 600 bp cDNA probe (section 4.2.1.2). These hybridisation results were then compared to the percentage of positive signals obtained using the same TIMP-1 probe on plaque lifts from the starting λgt10 cDNA library (section 5.2). A screen of 200,000 cDNAs from the starting λgt10 cDNA library resulted in the detection of only one TIMP-1 clone as indicated in Figure 5.1A. In comparison, the secondary selected sub-libraries from YACs A1220 and C1228 each contained the TIMP-1 cDNA at a frequency of 1 in 200 recombinant clones, an enrichment factor of 1000-fold as summarised in Table 5.1. An example of this hybridisation is shown in Figure 5.3B. The TIMP-1 positive clone was partially sequenced to confirm its identity to the TIMP-1 gene entry in GenBank (data not shown). The high level of enrichment of the TIMP-1 gene in the selected sub-libraries correlated well with the qualitative evaluation of enrichment for the same TIMP-1 cDNA transcripts using dot-blot and Southern blot analysis as described in sections 4.2.6.3 and 4.2.6.4.

This result demonstrated that the overall sensitivity of the cDNA selection scheme applied to the enrichment of a rare transcript, and is in good agreement with the upper limit of enrichment previously reported (Lovett et al., 1991). The selection procedure therefore seems sufficient to allow the isolation of rare retinal transcripts such as TIMP-1 whose original abundancy was as low as 1 in 2 X 10^5 clones in the starting λgt10 retinal cDNA library. The estimation of enrichment is only approximate as the long insert, oligo(dT) primed λgt10 adult retinal cDNA library was used and worked under stringent conditions after each round of selection and hence recovered only small insert cDNA clones (average length of 500bp). While screening the original λgt10 cDNA library can identify all clones, the factor of enrichment, in this case, can only roughly reflect the abundance of a certain transcript in the library and hence in the retinal tissue examined.
Figure 5.3 (continued)

(B) Hybridisation of duplicate filters of the A1220 YAC λgt10 secondary selected sub-library with the TIMP-1 cDNA indicates a single positive clone identical in the replica filter. One TIMP-1 cDNA clone detected out of 200 recombinant clones screened showed 1,000 fold enrichment for this rare retinal gene.
5.4.2.2 Probing with rhodopsin cDNA fragment

The degree of enrichment for the rhodopsin gene was quantitatively assessed by hybridisation of rhodopsin YAC λgt10 and λZAPII selected cDNA sub-libraries (section 5.3.3) with the 220 bp rhodopsin cDNA fragment (section 4.2.1.2). These hybridisation results were then compared to the number of rhodopsin positives obtained from hybridisation of the same rhodopsin probe on the starting λgt10 and λZAPII cDNA libraries (section 5.2). A screen of 200,000 cDNAs from the starting λgt10 and λZAPII retinal cDNA libraries resulted in the detection of 1,200 (Figure 5.1B) and 200 rhodopsin positive clones respectively. In comparison there were 80 (40%) and 40 (20%) rhodopsin positive cDNAs present in a screen of ~200 rhodopsin YAC λgt10 and λZAPII secondary selected cDNAs respectively (hybridisation data not shown). Hence from these comparisons, the abundant rhodopsin gene transcripts were found to be enriched by 70-fold and 200-fold in the rhodopsin λgt10 and λZAPII secondary selected cDNAs respectively as summarised in Table 5.1. These results indicated normalisation of the abundant retinal gene rhodopsin and hence a low level of enrichment of 70-fold from the rhodopsin YAC λgt10 selected sub-library and a greater enrichment of about 200-fold for rhodopsin gene transcripts in the λZAPII selected cDNAs.

These quantitative results correlated well with the qualitative evaluation for rhodopsin gene enrichment obtained after Southern blot analysis (section 4.2.6.3, Figure 4.14B). The difference in enrichment results obtained for both the rhodopsin YAC λgt10 and λZAPII selected cDNAs showed that the level of enrichment was dependent on the number of rhodopsin gene transcripts present in the starting cDNA library. The rhodopsin cDNAs present in the starting λgt10 retinal cDNA library were normalised in the selected cDNAs and the rhodopsin cDNAs present in the starting λZAPII retinal cDNA library were more enriched than normalised on direct cDNA selection. Therefore, some level of abundance normalisation probably occurred in the selection of rhodopsin transcripts. Theoretically, this should be possible using direct cDNA selection methods. If the genomic DNA is limiting in the hybridisation it will dictate the final level of each cDNA. Thus, the abundant cDNAs such as rhodopsin will quickly saturate their respective genomic targets, leaving the majority of higher abundance cDNAs in solution rather than hybridised to the genomic target. Low abundance cDNAs such as TIMP-1 will also be selected, but because there are less of these to start with, they will not be present in such a vast excess over their genomic sites. The net result when both sets are eluted from the genomic DNA should be an approximate normalisation of the abundant cDNAs downwards and the low abundance classes upwards. This conceptually simple postulation obviously ignores the problems that long genes with multiple exons, pseudogenes, gene families, hybridisation networking of cDNAs, or repeats would
introduce. However, some of these anticipated problems can be addressed where repeats are blocked with reasonable efficiency.

5.4.2.3 Probing with the properdin cDNA fragment

The properdin gene is present in all four different YACs A1220, 4HG2, 33CA11, and C1228 as shown in Figure 4.1B (section 4.1.6). The properdin gene was found to be present in the λgt10 adult retinal cDNA library (section 4.2.1.2) giving a 280 bp cDNA fragment on PCR analysis. This cDNA fragment was used as a probe in hybridisation studies for the selection of the properdin gene in the secondary selected YAC sub-libraries. All four different YAC secondary selected sub-libraries gave positive hybridisation signals with the properdin cDNA probe. An example of this is shown in Figure 5.3C which demonstrates three positive hybridisation signals (indicated by arrows) from both the master and replica filters of a plate lift from the YAC 33CA11 λgt10 selected sub-library. About 2% (four positive out of 220 recombinant clones) of the screened YAC 33CA11 λgt10 selected cDNAs were found to be properdin cDNAs, indicating that the properdin gene is moderately abundant; neither as rare as the TIMP-1 gene or as abundant as the rhodopsin gene.

The enrichment data summarised in Table 5.1 indicate that an abundant cDNA such as rhodopsin has essentially plateaued in its enrichment after one cycle of selection. In contrast, the low abundance cDNA, TIMP-1 shows a thousand-fold improvement in abundance with a secondary cycle of selection. Similar abundance normalisation and dramatic enrichment can be expected for unknown moderately abundant and low abundant cDNAs respectively from the 500 bp region of Xp11.23 studied here. The modified direct selection method is clearly capable of enriching cDNAs that are in the lowest abundance classes (e.g., TIMP-1) in the starting cDNA libraries to levels where they constitute several percent of the selected cDNA sub-libraries and can then be randomly picked. Such low abundance cDNA transcripts would otherwise be entirely missed by conventional screening protocols as described in section 1.5.4. The remaining 90% (Table 5.1) of cDNA clones were probably novel cDNAs and any other contaminating sequences which had not been competed against during the direct cDNA selection procedure. The nature of these cDNAs can only be exactly determined after sequence analysis of a percentage of random clones as described in section 5.5.

In summary the sub-library colony hybridisation screening indicated successful blocking and depletion of contaminating sequences such as β-actin, yeast, human repeat and ribosomal sequences. It confirmed a 1,000 fold enrichment and presence of the rare retinal gene TIMP-1 by identifying the specific TIMP-1 cDNA clone in the YACs A1220 and C1228 λgt10 selected sub-libraries. It also confirmed normalisation and enrichment.
Figure 5.3 (continued)

(C) Hybridisation of duplicate filters of the 33CA11 YAC λgt10 secondary selected sub-library with properdin cDNA indicate three positive clones identical in the replica filter. An estimated 2% of the screened ~220 recombinant clones of the 33CA11 YAC λgt10 secondary selected sub-library were Properdin cDNAs.
of the abundant retinal gene rhodopsin in the screened rhodopsin YAC λgt10 and λZAPII selected sub-libraries respectively.

5.5 Further analysis of sub-libraries

Selected clones which did not hybridise to any of the pre-blocked contaminating gene sequences and the internal control gene sequences (section 5.4) were randomly selected for further analysis. The cDNA clones were used in PCR reactions with pDIRECT vector-specific primer pairs which generated the individual cDNA inserts. These were fractionated on agarose gels, Southern blotted, and a number of the cDNA inserts were subsequently purified and characterised by sequence analysis (section 5.5.2). Sequence analysis was performed to determine cDNA overlap, their coding potential and to search databases of sequenced genes for identity and similarity.

5.5.1 Colony PCR of selected cDNA clones

Approximately 40-50 clones from each YAC specific sub-library were picked at random. This was carried out on the assumption that enough enrichment had occurred for successful random picking of even the very low abundance cDNAs present in the sub-library. Plasmid DNA was isolated (section 2.4.3.2) for the selected cDNA clones and the insert size range determined by colony PCR (section 2.6.2.2) using the pDIRECT λgt10 Forward-sense and Reverse-antisense primers (section 5.3.1.1). PCR reactions were fractionated on 1.5% agarose gels. Figure 5.4 shows the agarose gel photograph for one set of the analysed clones. The majority of the colony inserts were found to be in the size range of 300-500 bp. This small sized range of inserts obtained from the selected cDNA clones, is possibly due to the preferential selection for smaller sized cDNA transcripts (section 4.2.5.4) during the direct selection procedure, since this is a hybridisation procedure based on PCR amplified genomic DNA (section 4.2.2.4) and cDNA inserts (section 4.2.1.3). This is evident from the secondary selected cDNA transcripts of the TIMP-1 gene which are found in the size range of 200-500bp as described in section 4.2.6.3. When the starting cDNA libraries were subjected to multiple PCR amplification cycles, the longer cDNAs were selected against and the selection is then biased towards identifying shorter cDNAs. Therefore the average size of the selected cDNAs was relatively small although a few longer cDNA clones of 1,000bp were also selected for, which are probably near full length cDNA clones. Hence the transformation of colony-PCR procedures are not the limiting factors in obtaining longer cDNAs, but this is a result of the direct selection procedure itself.
Figure 5.4  PCR analysis of recombinant clones isolated from YAC 4HG2 secondary selected sub-library.

Determination of insert sizes of YAC 4HG2 secondary selected sub-library clones by colony PCR.

The figure shows an ethidium bromide stained gel photograph of colony PCR products fractionated on a 1.5% agarose gel with the ØX174/ HaeIII molecular weight marker (M.W.M) used as a size comparison (Lane 14). Distinct single bands in a insert size range between 1,000bp and 200bp was determined for the analysed clones (Lanes 1-11 & 13) with the majority of colony inserts (7 out of 12) present in the range 300-500bp as indicated by Clones 3-8 & Clone 11. This is probably due to preferential selection for smaller sized cDNA transcripts during the direct cDNA selection procedure. The no DNA track (Lane 12) showed no amplified bands for clone insert DNAs indicating absence of contaminating DNAs in the PCR reaction mix used.
5. Analysis of selected sub-library clones

The fact that a prevalence of small insert cDNA clones were isolated makes the re-screening of the YAC specific selected sub-libraries and starting λgt10 cDNA library a subsequent inevitable step in their further characterisation. Fractionated gels such as that shown in Figure 5.4 were transferred to Hybond-N+ (Amersham) using a bidirectional transfer method as described in section 2.2.2. Since not all the colony-PCR amplified cDNA inserts were sequenced and analysed (section 5.5.2), these colony-insert filters can be used for further studies of any interesting cDNA clones or genes which map back to the YAC contig under investigation. Such short fragment cDNA clones derived from each selection can also be used as probes to isolate longer overlapping clones from the YAC specific selected sub-libraries (section 5.3.3) and from the starting λgt10 (oligo-dT primed) cDNA library which is likely to contain longer cDNA clones of the same gene. This is because most cDNA libraries are constructed to yield full length or near full length copies of the mRNA of genes.

5.5.2 Sequence analysis using Gapped BLAST

Database search programs for sequence analysis were used to compare sequences of chosen cDNAs to a database of collected sequences which can identify similarities that suggest function, structure or evolutionary relationships. The BLAST programs as described in sections 1.5.10 & 2.8.4 has been the main algorithm used in database searches to detect related sequences. These programs were tailored for sequence similarity searches and are not generally useful for searching sequence motifs. BLASTN takes a nucleotide sequence (the query sequence) and its reverse complement and compares them against a nucleotide sequence database, and was initially used in this study. The BLASTN program was designed for speed, not maximum sensitivity, and is not intended for finding distantly related, coding sequences. Each BLAST search request consists of (1) the name of the BLAST program to use, (2) the name of the database to search, (3) the query sequence, and (4) various parameters and optional settings. The recently developed and improved Gapped BLAST program (Altschul et al., 1997; sections 1.5.10 & 2.8.4) which allows for gaps to be introduced into the alignment of sequences was the main search program used. Over a particular length of sequence, matches provide a positive score, whereas mismatches are penalised relative to an exact match. During a comparison, a total score is tallied: the higher the score the greater the similarity of the compared sequences. High scoring sequences are presumed to have biological relevance to the sequence of interest, although there are some caveats to this, such as the presence of repeat elements of no particular biological importance. Additionally, areas of sequence homogeneity (eg. AT- or GC-rich regions) can also produce a high score that carries no biological relevance (section 1.5.10).
Forty to fifty clones with an insert size range of 1.5kb to 200bp (section 5.5.1) for each YAC selected sub-library (a total of 200 random recombinant clones) were sequenced using T3 and T7 λgt10 cDNA library vector primers. PCR-amplified pDIRECT cloned inserts were precipitated with absolute ethanol and sequencing was performed on at least 150-250 bp of the PCR products using the T7 (Pharmacia) double-stranded cycle sequencing kit as described in section 2.8.2. Each of the cDNA sequences (200-300bp of sequence data) were then analysed for any sequence matches in the GenBank database (section 1.5.10) using the Gapped BLAST search program (Altschul et al., 1997).

5.5.3 Further contaminants in clones: vector sequences

Of the remaining 90% of cDNA clones which contained novel cDNAs or other contaminating sequences (Table 5.1), approximately 60% were found to be vector sequences after sequence analysis. This was also confirmed by hybridisation results of the selected sub-library filters (section 5.4) of all four YACs with a vector sequence probe (hybridisation data not shown), which indicated a high level of contamination of vector sequences in the selected sub-libraries. The vector sequences in the enriched cDNA sub-libraries were found to be inherent within the starting λgt10 retinal cDNA library. This is obvious from the sequencing data of the contaminant vector clone which shows evidence of the internal vector primer sequence (Set B, used in the PCR amplification of secondary selected cDNAs; section 4.2.5.4) of the λgt10 retinal cDNA library to be present just outside and next to the cloned vector sequence as seen in Figure 5.5; the autoradiograph showing the sequence of a vector clone. The pDIRECT vector sequence (section 5.3.1.1) is present just outside the λgt10 internal vector sequence (Figure 5.5); both sequences together making the λgt10 Forward-sense primer used in the cloning of secondary selected cDNAs (section 5.3.1).

The contaminating vector sequences were probably selected by uncompeted YAC vector DNA (pYAC4) present in the YAC genomic DNA. It is necessary to pre-block against selection of this vector sequence contamination which originated from the λgt10 retinal cDNA library, possibly during construction of this library. This can be done by competing the YAC genomic DNA with the YAC vector DNA (pYAC4), during the pre-blocking steps of the direct cDNA selection procedure (sections 4.2.4.1 & 4.2.5.1) in future experiments. It was reported in the original product review (Tagle et al., 1993) of a possible contamination of the cDNA library used with vector sequences, since about 30% of selected cDNA clones studied mapped back to cosmid vector bands. The reported contamination of repeat and ribosomal sequences (section 4.1.8) from enriched sub-libraries from YACs was much more common than YAC vector sequence contamination so it was thought unnecessary to compete out the vector sequences.
Figure 5.5 Nucleotide sequence data of a contaminant vector clone.

The internal vector primer sequence (Set B) of the λgt10 retinal cDNA library is present just outside the cloned contaminant vector sequence, indicating that such vector sequences are inherent within the starting λgt10 retinal cDNA library. The pDIRECT vector sequence is present just outside the λgt10 internal vector primer sequence; both sequences together forming the λgt10 Forward-sense primer used in the cloning of secondary selected cDNAs.
5. Analysis of selected sub-library clones

5.5.4 Clones containing novel cDNAs

30% of the selected clones were found to contain genuine cDNAs after sequence analysis. This totalled twenty five individual cDNA clones, including the internal TIMP-1 and PFC genes and other overlapping clones. Eighteen of these cDNA clones are further described in section 5.6, and show similarities with other ESTs and reported genes after GenBank database sequence analysis using the Gapped BLAST similarity search program (section 5.5.2). Table 5.2 (section 5.6) lists the similarities judged most likely to be biologically relevant. Percentage representation of these clones indicated that 48% (12 clones) including the TIMP-1 and PFC genes had a 96%-100% identity to the reported database genes/sequences. The remaining 52% (13 clones) had sequences containing Alu repeat elements or sequence motifs similar to the reported subject sequences, however they could also be part of unidentified genes/ESTs not yet found in the database. Sequencing of longer cDNA fragments of the these 13 clones may help in further assessment of their identities. Some of the analysed novel clones were chosen for mapping studies on the YAC contig used (section 5.7).

5.5.5 Clones with a poly-A tail

Ten out of the twenty five genuine cDNAs (section 5.5.4) were clones containing poly-A tails indicating that these were most likely potential genes (section 5.6). This is because most eukaryotic mRNAs have a poly-A tail added onto their 3’ ends. These ten cDNAs were isolated from the oligo-dT primed starting λgt10 retinal cDNA library, which preferably represents full length cDNA clones.

The enrichment procedure generated overlapping cDNAs, a feature which can be exploited to reconstruct longer cDNA sequences. Nevertheless, the search for full length transcripts will require re-screening the starting λgt10 cDNA library with the partial coding element already isolated. The reconstruction of longer expressed sequences from partial coding elements is an important step, firstly to distinguish the elements corresponding to the same genes, and secondly to help the search for similarities in the sequence databases.

Out of the 800 selected cDNA clones picked onto sub-library filters (section 5.4), 10% are expected to be repeat, ribosomal, yeast and β-actin cDNAs as described in Table 5.1. The remaining 720 clones (90%) will probably consist of 480 contaminating vector clones (60%) (section 5.5.3) and 240 novel cDNA clones (30%) inclusive of overlapping cDNA clones of which only 25 have currently been isolated (section 5.5.4). Although a large number of these novel cDNA clones may not map back to the contig or could be
pseudogenes or genes from multigene families some potentially interesting cDNA clones mapping to the YAC contig region should have been selected and need to be investigated by sequence analysis (section 5.5.2).

5.6 Description of selected cDNA clones

For a sequence database to be informative, two criteria must be met: (1) The query sequence must have a statistically significant match to a database sequence (a score greater than one expected by chance alone; Karlin and Altschul, 1990) and (2) there must be information available about the function of the sequence matched. It is quite common, however, that the functions of matched sequences are not obvious from the search results. Often sequence titles are uninformative and the full sequence database reports may then need to be retrieved and scanned to look for annotations that may identify the biological functions of the matched sequence. Also functionally important conserved domains such as enzyme active sites may not be noted as such in sequence database records. The approach to similarity searching taken by the Gapped BLAST programs is first to look for similar segments (HSPs) (section 1.5.10) between the query sequence and a database sequence, then to evaluate the statistical significance of any matches that were found, and finally to report only those matches that satisfy a user-selectable threshold of significance.

The clones isolated from the YAC specific selected sub-libraries as described below are denoted by the names of the respective YACs used. In each case the sequenced DNA insert fragments cloned into the pDIRECT vector were PCR amplified using the λgt10 sense and antisense primers as described in sections 5.3.1.1 and 5.5.1. The T7 and T3 ends sequences of the clones were then analysed by the Gapped BLAST program (Altschul et al., 1997).

Description of twelve sequenced clones used in mapping back studies

5.6.1 Clone 4HG2-76

The 4HG2-76 clone contains a 600bp DNA insert obtained by PCR amplification. The clone does not contain a poly-A tail on either end. The T7 and T3 ends of this clone were sequenced (section 2.8.2) and 171 and 139 bases of sequence respectively was analysed by the Gapped BLAST program (Altschul et al., 1997). The results suggested that 19bp of this sequence matched with 4 other sequences as indicated in the BLAST search results in Figure 5.6.1A.

The analysis of the 139bp query sequence from the T3 end of the 4HG2-76 clone showed alignment to about six different subject sequences from human, mouse and
Figure 5.6.1A

4HG2-76 T7

atgggtgagc agcattatta cattgtatgt cctgacaaag gcactcatgg
cctgaaaact gtctgttttt ttctccacc accagaacat aagctagtgt
cctgacatttc ttttgacac ggagggcttg aaagcatgca cttatcttcc
caaacacatg gctttttcttg g

Clone 4HG2-76 T7 end sequence of 171 bp.

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Sequence description</th>
<th>Score (bits)</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens DARC gene, 5'UTR</td>
<td>38</td>
<td>0.37</td>
</tr>
<tr>
<td>Human DNA sequence from PAC 154P24 on chrom...</td>
<td>38</td>
<td>0.37</td>
</tr>
<tr>
<td>glycoprotein D=Duffy group antigen [human, blo...</td>
<td>38</td>
<td>0.37</td>
</tr>
<tr>
<td>Homo sapiens 12q13 PAC RPCI1-316M24 (Roswe...</td>
<td>36</td>
<td>1.4</td>
</tr>
</tbody>
</table>

Sequence alignments obtained after Gapped BLAST analysis of the T7 end sequence of Clone 4HG2-76.
5. Analysis of selected sub-library clones

Figure 5.6.1B

4HG2-76 T3

aggttcgact taaggccctt ttggcagta ctacacgcc aactccctctctcctctcta ccctccctta cccccctatcc acatacttgt actcccacaa aagagcacacttactccaa aatacaacaa ttacaccacc cactcactc

Clone 4HG2-76 T3 end sequence of 139 bp.

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th></th>
<th>Score (bits)</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gbu193186</td>
<td>CTU193186</td>
<td>Cercocetus torquatus interleukin-1 alpha (IL...</td>
</tr>
<tr>
<td>gbu694921</td>
<td>MMU694922</td>
<td>Mus musculus second IL1 receptor alpha chain...</td>
</tr>
<tr>
<td>dbj1987002</td>
<td>D870002</td>
<td>Human (lambda) DNA for immunoglobulin light chain</td>
</tr>
<tr>
<td>emb1280774</td>
<td>HSHU173H7</td>
<td>Human DNA sequence from cosmid U173H7, betw...</td>
</tr>
<tr>
<td>gbu1023366</td>
<td>HSNH395A</td>
<td>Human HLA class III region containing NOTCH...</td>
</tr>
<tr>
<td>gbu153342</td>
<td>CEL0F01012</td>
<td>Caenorhabditis elegans cosmid F01G12</td>
</tr>
</tbody>
</table>

Score = 36.2 bits (18), Expect = 1.1
Identities = 18/18 (100%), Positives = 18/18 (100%)

Query: 53 tcctcttaccccctatcc 70
|----------------------|
Sbjct: 25724 tcctcttaccccctatcc 25741

repeat_region 24890..25407
/note="L1MA3 repeat: matches 529..1051 of consensus"
repeat_region 25574..26468
/note="L1PA8 repeat: matches 908..1 of consensus"

Gapped BLAST analysis of the T3 end sequence of Clone 4HG2-76 showed sequence alignments to six different subject sequences from human, mouse and C.elegans. These 18-26 bp overlapping subject sequences are about 100% identical to the 38-71 bp segment of the T3 end query sequence, and are probably a L1-repeat region as indicated in the Entrez nucleotide query data of emb.no. HSU173H7 of human DNA sequence from cosmid U173H7.
5. Analysis of selected sub-library clones

*C. elegans*, as shown in the Gapped BLAST results of Figure 5.6.1B. These 18-26bp overlapping subject sequences which are about 100% identical to the 38-71bp segment of the query sequence are probably a L1-repeat region as indicated in the NCBI Entrez nucleotide queries of emb. no. HSU173H7 (entrez query data shown in Figure 5.6.1B) and gb. no. HSMHC3W5A (query data not shown). This L1 repeat element is conserved in the interleukin-1 alpha genes of the *Cercocebus torquatus* and the *Mus musculus* species (entrez query data not shown). Hence the T3 end query sequence appears to contain a random DNA sequence identical to the L1 repeat elements present in the different subject sequences. The ORF finder query indicated three open reading frames, indicating that the T3 end of the 4HG2-76 clone may be part of a coding segment of a gene. This clone was used in the mapping studies described in section 5.7.

5.6.2 Clone 4HG2-108

The 4HG2-108 clone contains a 1kb PCR amplified DNA insert and did not contain a poly-A tail at the T3 end of the sequenced clone. The T7 end of this clone was difficult to sequence and did not give readable sequence. The T3 end of this clone gave a 145 bases of readable sequence. BLAST analysis of the 145bp query sequence from the T3 end of the 4HG2-108 clone shows that it contains a segment of DNA sequence present at the 20-50bp region of the query sequence which is approximately 100% identical to about ten different human subject sequences reported in the GenBank database, as shown in the BLAST search results of Figure 5.6.2. Though this 30bp segment of DNA could be a conserved sequence present in the various subject sequences, it is likely though that it is a repeat region as indicated by the Entrez nucleotide query of gb. no. AC004015 of the human BAC clone RG333F24 which indicated a repeat region at the 8577-8596bp segment of this subject sequence (entrez query data not shown).

The 18-35bp segment of the 4HG2-108 query sequence, is 100% identical to similar regions in the *C. elegans* cosmids C40H5 (emb.no. CEC40H5) and D1053 (emb.no. CED1053) which could be a conserved sequence present in both species. This DNA sequence is part of the coding sequence of chromosome III of *C. elegans*, as described in the Entrez nucleotide query of emb.no. CEC40H5 (query data not shown). It could however be a randomly similar 18bp DNA sequence found in both species, bearing no further significance.

In summary the 4HG2-108 clone contains a repeat region also present in other reported human clones. It may also contain a part of the coding segment for an unidentified gene (ORF finder showed one open reading frame) which is conserved in *C. elegans*. Since this clone was of interest it was included in the mapping studies of section 5.7. If it does map back to the region then further sequence will be derived from both ends of this clone and also from longer clones obtained from the starting λgt10 retinal cDNA library.
5. Analysis of selected sub-library clones

Figure 5.6.2

4HG2-108 T3
tttcccaagt atggtttctt caaacctagc ttttgctgtt ttatttgttt
agacgttgcg tgttcctttcc ttcggacac ttcagatgctttctgttgtgtttctgctcattgc
Clone 4HG2-108 T3 end sequence of 145 bp.

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Query Sequence</th>
<th>Score (bits)</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens chromosome 17 clone hCIT54K1</td>
<td>40</td>
<td>0.077</td>
</tr>
<tr>
<td>Homo sapiens PAC clone DJ0754J18 from 7p21</td>
<td>38</td>
<td>0.30</td>
</tr>
<tr>
<td>Human BAG clone RG333F24 from 7q11.2-q21</td>
<td>40</td>
<td>0.077</td>
</tr>
<tr>
<td>Homo sapiens DNA sequence from PAC 13D10 o...</td>
<td>36</td>
<td>1.2</td>
</tr>
<tr>
<td>Human PAC clone DJ290B04 from Xq23, comple...</td>
<td>36</td>
<td>1.2</td>
</tr>
<tr>
<td>Caenorhabditis elegans cosmid C4OH5, comple...</td>
<td>36</td>
<td>1.2</td>
</tr>
<tr>
<td>Homo sapiens DNA sequence from PAC 324L9 on chrom...</td>
<td>36</td>
<td>1.2</td>
</tr>
<tr>
<td>Homo sapiens Xp22 BAC GSHB-600G8 (Genome S...</td>
<td>36</td>
<td>1.2</td>
</tr>
<tr>
<td>Human DNA sequence from PAC 436M11 on chrom...</td>
<td>36</td>
<td>1.2</td>
</tr>
<tr>
<td>Homo sapiens DNA sequence from PAC 380E11...</td>
<td>36</td>
<td>1.2</td>
</tr>
<tr>
<td>Human BAG clone RG333F24 from 7q11.2-q21, complete sequence [Homo sapiens]</td>
<td>40</td>
<td>0.077</td>
</tr>
</tbody>
</table>

Score = 40.1 bits (20), Expect = 0.077
Identities = 20/20 (100%), Positives = 20/20 (100%)

Query: 31 ttttgctgtttatttgttt 50
Sbjct: 8577 ttttgctgtttatttgttt 8596
Caenorhabditis elegans cosmid D1053, complete sequence
Length = 43168

Score = 36.2 bits (18), Expect = 1.2
Identities = 18/18 (100%), Positives = 18/18 (100%)

Query: 18 cttcaactaacgttttg 35
Sbjct: 355 cttcaactaacgttttg 372

Gapped BLAST analysis of the 4HG2-108 T3 end sequence showed the 20-50 bp region of the query sequence is about 100% identical to ten different human subject sequences. This 30 bp segment of DNA was indicated as a repeat region (data not shown) at the 8577-8596 bp segment of the human BAC clone RG333F24 subject sequence. The 18-35 bp segment of the query sequence was 100% identical to a similar region in the C. elegans cosmid D1053, which could be a conserved sequence present in both species.
5.6.3 Clone 4HG2-110

The 4HG2-110 clone contains a 1.3kb DNA insert obtained by PCR. The clone does not contain a poly-A tail at the T3 end of the sequenced clone, and one could not be determined at the T7 end of the clone since sequencing proved difficult. BLAST analysis produced subject sequences from three different species showing alignments to three different segments of DNA from the 4HG2-110 query sequence as shown in the BLAST results of Figure 5.6.3. The 28bp DNA segment from the human tat/rev-like sequence, which is an HIV-1 regulatory gene related sequence is 92% identical to the query sequence. This subject sequence derived from chromosome 19 probably contains a 28bp random segment of DNA sequence which is identical to a similar region in the query sequence. The second alignment indicated a 95% identity to 22bp of the Rattus norvegicus tissue-type vomeronasal neurons putative pheromone receptor V2R4 mRNA, partial coding sequence; gb.no. AF053992. The ORF finder results showed one open reading frame present from 57-164bp which falls in the 101-122bp segment of identical query sequence to the reported rat pheromone receptor V2R4 gene. The third 18bp sequence alignment to the subject sequence of the yeast K. waltii pKW1 plasmid could either be a conserved DNA sequence present in both yeast and man but is most probably a random 18bp sequence of query sequence found to be identical to the K. waltii pKW1 plasmid sequence.

The 4HG2-110 clone was used in mapping studies (section 5.7). If it maps to the YAC genomic region of interest, further sequence data can be obtained from the T7 end of this clone and also from longer cDNAs which could be isolated from the starting λgt10 retinal cDNA library.

5.6.4 Clone 4HG2-125

The 4HG2-125 clone contains a 200bp DNA insert. In total 199bp of sequence were obtained from the T3 end of this clone. No poly-A tail was present on either end of the clone. Gapped BLAST analysis of the 199bp query sequence, produced two subject sequence alignments as shown in Figure 5.6.4 of the BLAST results. Both these subject sequences gave low scores and high E values, indicating that probably they were of no relevant biological significance. The first alignment to a 18bp segment of the Pig microsatellite repeat S0096 (gb.no. PIGS0096X) mapping to chromosome 12 probably shows a random 18bp segment of sequence found to be identical to the query sequence. The second alignment to 18bp of the Methanobacterium species (gb.no. AE000913) is also most likely to be a random sequence segment identical to the query sequence and hence is of no significance. The ORF finder results showed no open reading frame indicating here that the clone contains no coding sequence.
5. Analysis of selected sub-library clones

Figure 5.6.3

4HG2-110 T3

aggttcgact taaggcggtc agacattagg gtcgatgaac
ccttcgactc caaatctcaa ggacctc aaa ctccagtctt
caatctcga tcggaccggt tgtaacactt tgggacagag
atgacttcta cgatttttaac cgaccaacac caccgcgtgg
acatcaggt cgatgagtgg tggtc

Clone 4HG2-110 T3 end sequence of 185 bp.

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Score</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>0.10</td>
</tr>
<tr>
<td>36</td>
<td>1.6</td>
</tr>
</tbody>
</table>

Score = 36.2 bits (18). Expect = 1.6
Identities = 21/22 (95%). Positives = 21/22 (95%)

Query: 101 tgtaacacttttgacagagat 122

Sbjct: 484 tgtaacacttttgacaaagat 463

Gapped BLAST analysis produced subject sequences from three different species showing alignments to three different segments of DNA from the query sequence. The second alignment indicated a 95% identity to the 22 bp segment of the Rattus norvegicus tissue-type vomeronasal neurons putative pheromone receptor V2R4 mRNA, partial coding sequence. The 28 bp DNA segment from the human tat/rev-like sequence and the 18 bp sequence alignment to the yeast K.waltii pKWL plasmid are probably random segments of DNA sequences which are identical to similar regions in the query sequence.
Analysis of selected sub-library clones

Figure 5.6.4

4HG2-125 T3

```
aggttcgact taaggcgct taagaaaaacg aaaaatggga ccttcttttat
gagttacgctg tgtagatatct gcacgaccct cagggctgga aaactcgttc
agtgggacc aattcaggtt cgacttaagg cgccttaaga aacacgaaaaa
tggaaccttc accggctttg gtcaacctag gagatctcgc cggcggttg
```

Clone 4HG2-125 T3 end sequence of 199 bp.

```
0 50 100 150
Score E
```

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Score</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

qbl950089|PIGS0096X Pig microsatellite repeat S0096. | 36 | 1.7 |
| gblAE000913|AE000913 Methanobacterium thermoautotrophicum from ... | 36 | 1.7 |

Gapped BLAST analysis produced two subject sequence alignments of low scores and high E values, which are most likely to contain 18 bp random sequences identical to similar regions in the query sequence.

Figure 5.6.5

4HG2-162 T7

```
ccgtgtcctg cgtgacgaaa gccttcacct cctttaggcg gttcagctgt
catccccgcc aggtctctac ccgggcgcgg ttaggccgct
```

The 90 bp T7 end sequence of clone 4HG2-162.

```
0 50 100
Score E
```

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Score</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Score</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

gblU68299|MCU68299 Mouse cytomegalovirus 1 complete genomic seq... | 38 | 0.23 |
| gblU55375|CELK03E6 Caenorhabditis elegans cosmid K03E6 | 34 | 3.5 |

Gapped BLAST analysis produced two overlapping subject sequence alignments of low scores and high E values, which are probably random sequences identical to similar regions in the query sequence. Both subject sequence alignments did not show identity to any reported repeat elements.
Analysis of selected sub-library clones

In summary the 4HG2-125 clone contains DNA sequence which could not be identified as similar to any of the reported GenBank subject sequences. It is not entirely a vector or repeat sequence. The two subject sequence alignments are most likely to contain 18bp random sequences identical to regions in the query sequence. However, as the 4HG2-125 clone could be a non-coding part of a gene, it was included in the mapping back studies (section 5.7). If it does map back to the region of interest then more sequence data can be obtained from longer clones from the starting retinal library, which can then be re-analysed by Gapped BLAST.

5.6.5 Clone 4HG2-162

The 4HG2-162 clone contains a 300bp DNA insert. The T7 end of this clone when sequenced gave 90bp of sequence, with a poly-A tail indicating that it may be a gene sequence. No readable sequence was obtained from the T3 end of the clone. Gapped BLAST analysis of the 90bp query sequence produced two overlapping subject sequence alignments as shown in Figure 5.6.5 of the BLAST results. Both these subject sequences were of low scores and high E values, indicating that probably they were not of any relevant biological significance. The first alignment to a 19bp segment of the mouse cytomegalovirus 1 complete genomic sequence (gb.no. MCU68299) is probably a random 19bp segment of sequence found to be identical to the query sequence. Similarly the second alignment to a 17bp sequence of the C. elegans cosmid (gb.no. CELK03E6) is also likely to be a random 17bp sequence identical to the query sequence and of no relevant biological significance. Both subject sequence alignments show no identity to any reported repeat elements. The ORF finder results showed one open reading frame (3-111bp) indicating that the clone contained a coding sequence.

In summary the 4HG2-162 clone is likely to be a coding sequence and has both a poly-A tail and an open reading frame, although no sequence matches were found on GenBank. It was included in the mapping back studies (section 5.7). If it does map back to the region of interest then more sequence data can be obtained and further analysed from longer clones isolated from the starting retinal cDNA library.

5.6.6 Clone 4HG2-164

Clone 4HG2-164 contains a 600bp DNA fragment. The T7 end of this clone was difficult to sequence and did not give readable sequence. The T3 end of this clone gave 184bp sequence and no poly-A tail. BLAST analysis of this 184bp 4HG2-164 T3 sequence produced subject sequence alignments to three different Homo sapiens clones as shown in the BLAST results of Figure 5.6.6. The first of these subject sequence alignments was the Homo sapiens chromosome 16, cosmid clone 378E2 (LANL) which is 98% identical to the entire 184bp of the query sequence (Figure 5.6.6). The NCBI Entrez nucleotide query results (gb.no. AC004035) further showed that this subject sequence maps to
Figure 5.6.6

4HG2-164 T3

atacccccaa gcaccagcag agagctctgt ggaggctggcc aagcacacag
tggggcttca cccgttcgtt cacaccttcc ctttaccctc tcacctggcc
tcttagcctt tgttgaatat ctgggctctg ggccatttca gatctactgt
cattctttct caatggctaa gccacctgac atct

The 184 bp T3 end sequence of clone 4HG2-164.

<table>
<thead>
<tr>
<th>Query</th>
<th>Score (bits)</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB1AC004035</td>
<td>Homo sapiens chromosome 16, cosmid clone 3...</td>
<td>349</td>
</tr>
<tr>
<td>AB1AC004253</td>
<td>Homo sapiens chromosome 17, clone HCIT542B...</td>
<td>38</td>
</tr>
<tr>
<td>AB1AC004035</td>
<td>Human beta-2 integrin alphaD subunit (ITGAD...</td>
<td>16</td>
</tr>
</tbody>
</table>

Gapped BLAST analysis produced subject sequence alignments to three different human clones. The Homo sapiens chromosome 16, cosmid clone 378E2 (LANL) which has 98% identity to the entire 184 bp of the query sequence had the highest score and lowest E value. This subject sequence is not a repeat region and may be a non-coding segment of a gene present in Clone 4HG2-164. The two other reported subject sequences are likely to be random segments of DNA sequences which are identical to the query sequence.
chromosome 16p13.3 and the 13698-13881bp identical segment was reported not to be present in a repeat region (entrez query data not shown). The ORF finder results showed no open reading frames for this query sequence indicating that it may be a non-coding segment of a gene present in the human chromosome 16 cosmID clone LANL.

The two other reported subject sequences: HCIT542B on chromosome 17 and the beta-2 integrin alphaD subunit (ITGAD) gene, were both of low scores and high E-values showing a 19bp, 100% identity and 22bp, 95% identity respectively to the 184bp 4HG2-164 query sequence (query data not shown). Both these subject sequence alignments are probably random segments of DNA sequence which are identical to the query sequence and are probably of no significance.

The 4HG2-164 clone therefore probably contains an identical sequence to that found in the human cosmid clone 378E2 (AC004035). It could be a non-coding segment of a gene as no open reading frames were detected. No repeat or vector elements were detected within the clone. Since the sequence of this clone could be potentially interesting it was included in the mapping back studies of section 5.7. If it positively maps to the genomic region of interest, further sequence data could be obtained from the T7 end of this clone and also from longer cDNAs which could be isolated from the starting λgt10 retinal cDNA library.

5.6.7 Clone 4HG2-218

Clone 4HG2-218 contains a 300bp DNA fragment. No poly-A tail was present on either end of this clone. The T7 and T3 ends of the clone were sequenced and 215 and 109 bases of sequence respectively were analysed by Gapped BLAST. BLAST analysis of the 215bp query sequence from the T7 end of the 4HG2-218 clone showed it was most likely to be the gene for the human retinal arrestin (S-antigen). This is indicated in the first two subject sequences with highest scores and low E values, as shown in the BLAST search results of Figure 5.6.7A. The first alignment showed 97% identity of the human mRNA for retinal S-antigen (emb.no. X12453/HSRETS) to almost the entire query sequence, from 14-191bp. The second subject sequence identity was again to the human arrestin gene, to the exon 16 coding sequence (gb.no. U70976/HSSAGART15) which is 98% identical to the 65-191bp segment of the query sequence. Another subject sequence (gb.no. U70974/HSSAGART13) from exon 14 of the same human arrestin gene showed 100% identity to the 14-56bp segment and is located before the HSSAGART15 subject sequence more towards the 5' end of the gene. The one open reading frame from 32-172bp, indicated in the ORF finder for the query sequence confirmed that the clone contains a coding sequence, in this case exons 14 and 16 of the human arrestin (SAG) gene. The human arrestin gene has been reported to map to human chromosome 2q (Ngo et al., 1990) and was isolated from retinal tissue.
5. Analysis of selected sub-library clones

Figure 5.6.7A

4HG2-218 T7

cggatttcgctggagttttgattttcgtggagctctcctcggactta
cagcctggaggagccagctaatagtttagaggtttgcctggcataatttgaggttggggagagaagaagtattcaggatgaa
ttgcagggagagccgggagagagaacaagggcagaagagagtacgacattgagttgagaattgaggttgctcagggatcggagactgtcagagattggcaaca

The 215 bp T7 end sequence of clone 4HG2-218.

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Query</th>
<th>Score (bits)</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMB</td>
<td>X12456</td>
<td>HSSAGRTSA</td>
</tr>
<tr>
<td>EMB</td>
<td>U70976</td>
<td>HSSAGRTT1</td>
</tr>
<tr>
<td>EMB</td>
<td>M24845</td>
<td>BOVANTSR</td>
</tr>
<tr>
<td>EMB</td>
<td>M15115</td>
<td>BOVANTSR</td>
</tr>
<tr>
<td>EMB</td>
<td>X98470</td>
<td>CSMTRPL6</td>
</tr>
<tr>
<td>EMB</td>
<td>U102233</td>
<td>HS474</td>
</tr>
<tr>
<td>EMB</td>
<td>U103977</td>
<td>YCH9266</td>
</tr>
<tr>
<td>EMB</td>
<td>Y844502</td>
<td>HS26H23</td>
</tr>
<tr>
<td>EMB</td>
<td>X62981</td>
<td>SMC108</td>
</tr>
<tr>
<td>EMB</td>
<td>X62216</td>
<td>CEC108</td>
</tr>
<tr>
<td>EMB</td>
<td>U03124</td>
<td>RNU53524</td>
</tr>
</tbody>
</table>

BLAST analysis showed that the T7 end sequence of Clone 4HG2-218 contains either part of the gene or a homologue of human retinal arrestin (S-antigen) as indicated by the first two subject sequences with highest scores and lowest E values. Most of the remaining sequence alignments are conserved sequence from the retinal S-antigen (arrestin) gene in cow, pig, dog, rat and mouse.
5. Analysis of selected sub-library clones

Figure 5.6.7A continued

emb|X12453|HSSRRTSA Human mRNA for retinal S-antigen (48 KDa protein)
Length = 1582
Score = 321 bits (162), Expect = 2e-86
Identities = 174/178 (97%), Positives = 174/178 (97%)

Query: 14 gagttccccaggctccctgtagctgccacctgacgtccacctgacgtcaggagccagctaatggagatcag 73
Sbjct: 1281 gagttccccaggctccctgtagctgccacctgacgtccacctgacgtcaggagccagctaatggagatcag 1340

Query: 74 gatgcaatatttagtttttgaggagtttgctcgccataatctgaaagatgcaggagaagct 133
Sbjct: 1341 gatgcaatatttagtttttgaggagtttgctcgccataatctgaaagatgcaggagaagct 1400

Query: 134 gaggaggggaagagagacaagaatgacattgatgagtgaagatgtcggctcaggatgc 191
Sbjct: 1401 gaggaggggaagagagacaagaatgacgctgatgagtgaagatgtcggctcaggatgc 1458

gb|U70976|HSSAGART15 Human arrestin (SAG) gene exon 16 and complete cds
Length = 215
Score = 236 bits (119), Expect = 1e-60
Identities = 125/127 (98%), Positives = 125/127 (98%)

Query: 65 agttatcaggatgcaaatatttagtttttgaggagtttgctcgccataatctgaaagatgc 124
Sbjct: 62 agttatcaggatgcaaatatttagtttttgaggagtttgctcgccataatctgaaagatgc 121

Query: 125 gggagaagctggagggaaagagagcaaatgctcattgtaattgtgaatgtcggctcaggatgc 184
Sbjct: 122 gggagaagctggagggaaagagagcaaatgctcattgtaattgtgaatgtcggctcaggatgc 181

Query: 185 aggatgc 191
Sbjct: 182 aggatgc 188

The first alignment showed 97% identity of the human mRNA for retinal S-antigen to almost the entire query sequence from 14-191 bp. The second sequence identity was to the human arrestin gene exon 16 coding sequence, with a 98% identity to the 65-191 bp segment of the query sequence.
Figure 5.6.7B

4HG2-218 T3

attttataac aaacaagctt tattttctgg aagaagactc
atctcttcat acgcagcata actccaaaggy actaaactgt
ggggcttgct cgtgcactgg taactacag

The 109 bp T3 end sequence of clone 4HG2-218.

BLAST analysis of the 109 bp sequence from the T3 end showed one high score for the human mRNA for retinal S-antigen (48 KDa protein) with 98% identity from 8-109 bp of this query sequence.
The Entrez query data of emb.no. X12453/HSRETSA showed a coding segment from 15-1439 bp and a polyA site from 1582 bp, indicating that the identical segment of 1470-1573 of the T3 end query sequence corresponds to the 3' untranslated region of the human retinal S-antigen gene.
The following ten different sequence identities are conserved sequences of the retinal S-antigen (arrestin) gene present in the cow, mouse, pig, dog and rat species. These show about 85%-88% identity to the 14-130bp segment of the human homologue present in the 4HG2-218 T7 query sequence. The similarity of nucleotide sequence between human and bovine is ~ 80% (Yamaki et al., 1988) and confirms the above finding. This conservation of the arrestin gene signifies its essential role in the mammalian rod phototransduction pathway (section 1.6.3). The remaining sequence identities (of gb & emb.nos. AL022333, AC003064, U10397, Z84467, X69480, Z68214 and U53524; Figure 5.6.7A) are 18-23bp subject sequences which are 95%-100% identical to the 132-158bp and the 63-80bp regions of the query sequence and are most probably similar random repeated sequence motifs present in both the subject and query sequences.

Analysis of the 109bp query sequence from the T3 end of the 4HG2-218 clone showed one high score for the human mRNA for retinal S-antigen (48KDa protein), with 98% identity from 8-109bp of this query sequence (shown in BLAST results of Figure 5.6.7B). The T3 end of the query sequence seems to be part of the 3′untranslated region of the human retinal S-antigen gene as indicated from the results of its Entrez nucleotide query of emb.no. X12453/HSRETSA. This subject sequence shows a coding segment from 15-1439bp and a polyA site from 1582bp (as shown in Figure 5.6.7B). The third subject sequence of the Canis familiaris (dog) mRNA for arrestin showed 88% identity to a 34bp (8-41bp) segment of DNA of the query sequence, indicating a conserved segment of DNA sequence between dog and human arrestin at the 3′ untranslated region of the gene. As one might expect the 3′ untranslated region of the human arrestin gene (T3 sequence) showed less identity to arrestin sequences of other mammalian species than the coding segment of the same gene (T7 sequence).

The 4HG2-218 clone is most likely to be part of the coding region (exons 14 and 16) of the human retinal S-antigen (arrestin) gene and shows 80% nucleotide sequence conservation to the arrestin gene of various other mammalian species. It also contains the 3′ untranslated region of the same gene. This gene is expressed in retinal tissue and could possibly represent either a homologue or pseudogene present in the Xp11.23 region under study. As it could be an interesting potential candidate gene involved in the phototransduction pathway, it was tested in the mapping back studies (section 5.7).

5.6.8 Clone 33CA11-239

Clone 33CA11-239 contains a 900bp DNA insert obtained by PCR. No readable sequence was obtained from the T7 end of this clone. The T3 end though, gave 214 bases of sequence with no poly-A tail.

On BLAST analysis (Figure 5.6.8) about twenty five of the reported subject sequences have a 19-25bp segment of DNA which is 95%-100% identical to the 121-172bp
Figure 5.6.8

33CA11-239 T3

ttcctgtacg atgggggata taacggggtt cactaacctt
cccatggcat tgaaacctcc cccaaacctg atggacctag
aatctctgct tgtacctcgc cggcccccac agtgggcat
tttttctctt gttccctctc tttgaaaat gtaaaataaa
accaaaatg gacaactttt ttccagccaa ttcaccata
gagaacaac cttat

The 214 bp T3 end sequence of clone 33CA11-239.

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Score (bits)</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>42</td>
<td>0.030</td>
</tr>
<tr>
<td>42</td>
<td>0.030</td>
</tr>
<tr>
<td>42</td>
<td>0.030</td>
</tr>
<tr>
<td>42</td>
<td>0.12</td>
</tr>
<tr>
<td>38</td>
<td>0.47</td>
</tr>
<tr>
<td>38</td>
<td>0.47</td>
</tr>
<tr>
<td>38</td>
<td>0.47</td>
</tr>
<tr>
<td>38</td>
<td>0.47</td>
</tr>
<tr>
<td>38</td>
<td>0.47</td>
</tr>
<tr>
<td>38</td>
<td>0.47</td>
</tr>
<tr>
<td>38</td>
<td>0.47</td>
</tr>
<tr>
<td>38</td>
<td>0.47</td>
</tr>
<tr>
<td>38</td>
<td>0.47</td>
</tr>
<tr>
<td>38</td>
<td>0.47</td>
</tr>
<tr>
<td>38</td>
<td>0.47</td>
</tr>
<tr>
<td>38</td>
<td>0.47</td>
</tr>
<tr>
<td>38</td>
<td>0.47</td>
</tr>
<tr>
<td>38</td>
<td>0.47</td>
</tr>
<tr>
<td>38</td>
<td>0.47</td>
</tr>
<tr>
<td>38</td>
<td>0.47</td>
</tr>
<tr>
<td>38</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Gapped BLAST analysis of the T3 end query sequence showed about twenty five subject sequences with a 19-25 bp size segment of DNA which is 95%-100% identical to the 121-172 bp region of the query sequence. These subject sequences probably contain a similar sequence motif which is present in both query and subject sequences. This identical sequence motif (121-172 bp region of the query region) is also conserved in the C. elegans, Mus. musculus and S. aureus species.
Analysis of selected sub-library clones

region of the query sequence. This suggested that such subject sequences may contain a motif which is similar in both the subject and query sequences. The possibility that this could be a repeat element DNA sequence such as \textit{Alu} or \textit{LINE1} was investigated by looking at some examples of Entrez nucleotide queries (query data not shown). The \textit{S. aureus} mecR1 and mec1 genes (emb.no. X63598/SAMECR1I) identical segment of 400-418bp was found to be a coding region but no further information was given. The human DNA sequence from PAC 52202 on chromosome 6q21 (emb.no. Z95329/HS52202) was found not to contain a repeat element at its identical segment of DNA from 30819-30840bp, the same was also true for the human BAC clone GS542D18 (gb.no. AC002528) from 7q31-q32 at its identical segment of 34038-34059bp. Therefore it is not possible to deduce whether the 121-172bp region of the query sequence is a repetitive element. This identical sequence motif is also conserved in some of the other species such as \textit{C. elegans} cosmid C05D10 (gb.no. U13645/ CELC05D10) and \textit{Mus musculus} p45 NF-E2 related factor 2 (NRF2) gene (gb.no. U70475/MURNRF2). Two open reading frames were determined using the ORF finder (data not shown) from 33-199bp on the reverse strand of the query sequence, suggesting that this clone could potentially contain a coding sequence for a gene.

In summary from the sequence analysis of the T3 end query sequence of the 33CA11-239 clone it can be concluded that most of the reported subject sequences may contain a motif which is probably not a repeat element. Since this clone could contain a potentially interesting sequence motif it was used in the mapping back studies described in section 5.7. If it maps back to the Xp11.23 region used in the selection procedure, then further sequence could be derived from both ends of this clone and also from longer clones obtained from the starting \textit{\lambda}gt10 retinal cDNA library.

5.6.9 Clone 33CA11-276

The 33CA11-276 clone contains a 900bp insert. This clone does not contain a poly-A tail at the T3 end which was sequenced. 199 bp of T3 end sequence was obtained and analysed. The T7 end did not give readable sequence and was not analysed. One open reading frame on the reverse strand from 1-170bp was indicated in the ORF finder results (data not shown). The analysis of the 199bp query sequence from the T3 end of the 33CA11-276 clone shows five subject sequences of varying identities (Figure 5.6.9) as described here with high scores and low E values. It indicates that it is most likely to be part of the sequence of the human NADP dependent leukotriene B4 12-hydroxydehydrogenase gene (partial coding sequence) which was isolated from human kidney tissue and is 100% identical to the 1-193bp segment of DNA of the query sequence (as shown in Figure 5.6.9). Leukotriene B4 12-hydroxydehydrogenase catalyzes the conversion of leukotriene B4 into its biologically less active metabolite, 12-oxo-
Figure 5.6.9

33CA11-276 T3

gggccctgaggg gaagtggggcc ggtttctgta taagtagaga
tggtcctccac tattgcaaatc ctctcaaatc tttctcatg
gccgataaca gttttgtggaa actctctcacc tacatattca
ataataacat cataacacct agggagagct tttctcaagg
rttcttcaac agactctacc gtctttgagt taagacgac

The 199 bp T3 end sequence of clone 33CA11-276.

Gapped BLAST analysis of the T3 end query sequence showed five subject sequences of varying identities with high scores and low E values. The Clone 33CA11-276 is likely to contain part of the sequence of the human NADP dependent leukotriene B4 12-hydroxydehydrogenase gene (partial coding sequence), which was 100% identical to the 1-193 bp segment DNA of the query sequence. The pig leukotriene gene showed conserved sequences to the human homologue with 89% identity. Two other conserved genes present in the O. cuniculus (rabbit) and rat species are probably cognate genes with 87% and 84% identity respectively.
leukotriene B4. This is an initial and key step of metabolic inactivation of leukotriene B4 in various tissues other than leukocytes (Yokomizo et al., 1996).

The *Sus scrofa* (pig) 15-oxoprostaglandin 13-reductase mRNA, complete coding sequence, isolated from pig lung tissue was found to be 89% identical to the 4-176bp segment of DNA of the query sequence, the same was also true for the pig mRNA for NADP dependent leukotriene B4 12-hydroxydehydrogenase gene (isolated from pig kidney). The pig leukotriene gene showed conserved sequences with the human homologue, as has been reported, where the human enzyme showed 97.1% homology with the porcine enzyme (Yokomizo et al., 1996). The porcine enzyme was expressed as a glutathione S-transferase fusion protein in *E.coli.*, which exhibited similar characteristics with the native enzyme. Two other conserved genes present in rabbit and rat are probably cognate genes with 87% and 84% identity respectively to the 1-193bp DNA segment of the query sequence.

In summary the 33CA11-276 clone is probably part of the human NADP dependent leukotriene B4 12-hydroxydehydrogenase gene. It also contains conserved sequences to the pig, rabbit and rat cognate genes. The human leukotriene gene has not been mapped, so it could map to the Xp11.23 locus or alternatively this sequence could represent a pseudogene or even a member of a multigene family present in this region. It was therefore used in the mapping back studies described in section 5.7.

### 5.6.10 Clone 33CA11-283

Clone 33CA11-283 contains a 700bp DNA fragment. The T7 end of this clone was difficult to sequence and did not give readable sequence. The T3 end of this clone gave a 109bp sequence. BLAST analysis of this 109bp 33CA11-283 T3 end sequence produced subject sequence alignments to two different *Homo sapiens* clones with high scores and low E values, as shown in Figure 5.6.10. The sequences dbj.no. D29011/HUMPSX and emb.no. X95586/HSMB1GENE were from the human proteasome gene, showing about 97% identity to 109bp of the query sequence. Entrez nucleotide queries of these sequences indicated that the query sequence at the T3 end (1-109bp) is identical to the 3' untranslated region of the human proteasome gene, with no repeat element (data not shown). This is confirmed by an absence of open reading frames using the ORF finder (data not shown), indicating no coding regions. Proteasomes are the proteolytic complex responsible for major histocompatibility complex (MHC) class I-restricted antigen presentation (Akiyama et al., 1994). The gene maps to human chromosome 14q11.2 (entrez query data of emb.no. X95586/HSMB1GENE) and was isolated from the human HepG2 cell line (entrez query of dbj.no. D29011/HUMPSX).

Therefore the 33CA11-283 clone is most likely to be a part of the 3' untranslated region of the human proteasome gene, which also contains some conserved sequence motifs to the human and mouse immunoglobulin genes. Although this gene maps to
5. Analysis of selected sub-library clones

Figure 5.6.10

33CA11-283 T3

gatgtgtaaa tagtgtaatg ttaacatcca agaaagttaa
acaaatagtct accccctcgac cagttcatta atgactgcta
acacatag gtccatgtg ccagacgtt

The 109 bp T3 end sequence of clone 33CA11-283.

<table>
<thead>
<tr>
<th>Query</th>
<th>Score</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbiID290111HUMPSX Human mRNA for proteasome subunit X, complete...</td>
<td>184</td>
<td>2e-45</td>
</tr>
<tr>
<td>emb</td>
<td>X0012835</td>
<td>HSMHIGENE H.sapiens MB1 gene</td>
</tr>
<tr>
<td>emb</td>
<td>AC023595</td>
<td>SPC0794 S.pombe chromosome III cosmid c794</td>
</tr>
<tr>
<td>emb</td>
<td>AC028111</td>
<td>ATF10MG Arabidopsis thaliana DNA chromosome 4, BAC...</td>
</tr>
<tr>
<td>emb</td>
<td>Z68970</td>
<td>HL100G1 Human DNA sequence from cosmid L30G1, Huntin...</td>
</tr>
<tr>
<td>gbl</td>
<td>Z00451</td>
<td>MUSIGCD18 Mouse germline IgG-3 chain gene, D-J-C regi...</td>
</tr>
<tr>
<td>gbl</td>
<td>Z81000</td>
<td>HE6766 Human DNA sequence from Fosmid 67D6 on chromo...</td>
</tr>
<tr>
<td>emb</td>
<td>X0012835</td>
<td>MMI00100 Mouse gene IgG-3 heavy chain constant region</td>
</tr>
<tr>
<td>dbi</td>
<td>AB002262</td>
<td>AB002262 Human mRNA for KIAA0364 gene, complete cds</td>
</tr>
<tr>
<td>gbl</td>
<td>AF034191</td>
<td>AF034198 Homo sapiens immunoglobulin superfamily-li...</td>
</tr>
<tr>
<td>emb</td>
<td>Y10521</td>
<td>HSOCAM Homo sapiens mRNA for immunoglobulin-like domain...</td>
</tr>
<tr>
<td>gbl</td>
<td>M97632</td>
<td>MUSGABER Mouse gamma-aminobutyric acid transporter p...</td>
</tr>
<tr>
<td>gbl</td>
<td>AC004690</td>
<td>AC004690 Homo sapiens PAC clone DJ0630C24 from 7q31...</td>
</tr>
<tr>
<td>emb</td>
<td>Z15055</td>
<td>CF7592 Caenorhabditis elegans cosmid F59B2, complete...</td>
</tr>
<tr>
<td>emb</td>
<td>Z52766</td>
<td>SPCAG8 S.pombe chromosome I cosmid c4G8</td>
</tr>
<tr>
<td>dbi</td>
<td>DJ8343</td>
<td>DJ8343 Mouse DNA for Ig gamma-chain, secrete-type an...</td>
</tr>
</tbody>
</table>

BLAST analysis of the T3 end query sequence produced subject sequence alignments to two different Homo sapiens clones of the human proteosome gene with high scores and low E values. The remainder of the fourteen subject sequence alignments are likely to contain similar sequence motifs (7-21 bp size range) which are 100% identical to similar regions of the query sequence.
Figure 5.6.10 continued

HUMPSX Human mRNA for proteasome subunit X, complete cds
Length = 1018
Score = 184 bits (93), Expect = 2e-45
Identities = 106/109 (97%), Positives = 106/109 (97%), Gaps = 1/109 (0%)

Query: 1 gagtagtgaatagtgtaatgttaacatccaagaaagtaaaacaaatagtcacccctgcag 60
       l l l l l l l l l l l l l l l I M I l l l l l l l l l l l l M I M I I I I l l l l l l
Sbjct: 1018 gagtagtgaatagtgtaatgttaacatccaagaaagtaaaacaaatagtcacccctgcag 959

Query: 61 cagctcattagactgtaacacatagaggttcaatgtgccagagctt 109
       l l l l l l l l l l l l l l l l l l l l l l M I l I I I I I I I I I I I I I I I I I I I I
Sbjct: 958 cagctcattagactgtaacacatagaggtttcaatgtgccagagcttt 911

HSMBIGENE H.sapiens MBl gene
Length = 6502
Score = 170 bits (86), Expect = 2e-41
Identities = 106/110 (96%), Positives = 106/110 (96%), Gaps = 2/110 (1%)

Query: 1 gagtagtgaatagtgtaatgttaacctcaagaaagtaaaacaaatagtcaccctgca 59
       I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I
Sbjct: 6095 gagtagtgaatagtgtaatgttaacctcaagaaagtaaaacaaatagtcaccctgca 6036

Query: 60 gcagctcattagactgtaacacatagaggttcaatgtgccagagctt 109
       l l l l l l l l l l l l l l l l l l l l l l l l l l l l l l l l l l I I I I I I I I I I I I I I I I I I I I
Sbjct: 6035 gcagctcattagactgtaacacatagaggtttcaatgtgccccagagcttt 5987

The sequences of the human proteasome gene; dbj.no. D29011/HUMPSX and emb.no. X95586/HSMBIGENE showed approximately 97% identity to the 109 bp region of the query sequence.
human chromosome 14, the possibility that clone 33CA11-283 could represent a pseudogene or a member of a multigene family located at the Xp11.23 region could not be excluded, so it was used in the mapping back studies (section 5.7).

5.6.11 Clone A1220-164

Clone A1220-164 contains a 250bp DNA fragment. The T3 end of this clone was difficult to sequence and did not give readable sequence. The T7 end is a 123bp sequence with a poly-A tail indicating that it may be a gene sequence. BLAST analysis of the 123bp sequence from the T7 end of the A1220-164 clone indicates that it may be part of the sequence of the human tyrosyl-tRNA synthetase gene, as shown in the search results of Figure 5.6.11. This is clearly shown by the 99% identity of the 2-119bp region of the query sequence to the 368-485bp segment of the human tyrosyl-tRNA synthetase gene (gb.no. U89436/HSU89436). This is confirmed by the 89% identity of this same region of the T7 end sequence to the *Bos taurus* partial tyrosyl-tRNA synthetase gene (322-439bp). Sequence alignments with five genes from various species showed an 18-19bp segment of query sequence identical to similar sequence segments in these subject alignments. These are probably identical sequence motifs or could be repeat regions similar in both subject and query sequence.

In summary the A1220-164 clone may contain a portion of the human tyrosyl-tRNA synthetase gene, which is also conserved in the *Bos taurus* species. Although the tyrosyl-tRNA synthetase gene has been mapped to human chromosome 14q32, there may be a homologue or pseudogene in the Xp11.23 region. Clone A1220-164 was therefore used in mapping back studies (section 5.7).

5.6.12 Clone A1220-238

The A1220-238 clone contains a 600bp DNA insert obtained by PCR amplification. The T7 end of this clone was difficult to sequence and did not give readable sequence. The T3 end of this clone gave 163bp of readable sequence, but does not contain a poly-A tail. BLAST analysis of the T3 end of the A1220-238 clone gave four high scoring and low E value *Homo sapiens* subject sequences which are 97%-100% identical to the 8-163bp segment of the query sequence as indicated in the BLAST search results of Figure 5.6.12. These four sequences are all probably clones containing the same coding sequence for a portion of the human neuron-specific gamma-2 enolase (ENO2) gene. The ORF finder results suggest one open reading frame from 25-150bp (data not shown). The ENO2 gene is present in a gene-rich cluster at human chromosome 12p13 and was isolated from brain tissue as shown in Entrez nucleotide query for subject sequences of gb.no. M22349/HUMENOG (Figure 5.6.12). The A1220-238 T3 sequence is probably the 3' end of the coding region of the human ENO2 gene as indicated by the
5. Analysis of selected sub-library clones

Figure 5.6.11

A1220-164 T7

gaccagctca gcgaagagta cacactagat gtgtacagac
tctctccggt gtgcacacag cagattcca agaaggcttg
agctgaggtg gtaaacaggg tggagcaccn tttgctgagc ggc

The 123 bp T7 end sequence of Clone A1220-164.

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Score</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gb</td>
<td>U40714</td>
<td>HSU40714 Human tyrosyl-tRNA synthetase mRNA, complete... 228 1e-58</td>
</tr>
<tr>
<td>gb</td>
<td>U89436</td>
<td>HSU89436 Human tyrosyl-tRNA synthetase mRNA, complete... 228 1e-58</td>
</tr>
<tr>
<td>emb</td>
<td>X86373</td>
<td>PTTYRSA B.taurus partial tyrS gene 141 2e-32</td>
</tr>
<tr>
<td>gb</td>
<td>U65481</td>
<td>CCU55481 Cyanella capillata 18S ribosomal RNA gene, par... 18 0.26</td>
</tr>
<tr>
<td>gb</td>
<td>AE000631</td>
<td>HPAE000631 Helicobacter pylori section 111 of 134 o... 16 1.0</td>
</tr>
<tr>
<td>gb</td>
<td>AF038640</td>
<td>AF038640 Mus musculus mitogen-activated protein kin... 16 1.0</td>
</tr>
<tr>
<td>emb</td>
<td>Z93242</td>
<td>HSMBPX45 Human DNA sequence from PAC 389A20 on chrom... 16 1.0</td>
</tr>
<tr>
<td>emb</td>
<td>X17288</td>
<td>HSMBP2X45 H.sapiens MBP gene, exons 4 and 5 16 1.0</td>
</tr>
</tbody>
</table>

Gapped BLAST analysis of the T7 end query sequence indicated that it is probably part of the sequence of the human tyrosyl-tRNA synthetase gene, as shown in the top two subject sequence alignments. This is confirmed by the conservation of the sequence to a part of the partial tyrosyl-tRNA synthetase gene in the Bos taurus species. The remaining sequence alignments to five genes in various species show identical random sequence motifs (18-19 bp segments) which are similar in both query and subject sequences.
5. Analysis of selected sub-library clones

Figure 5.6.11 continued

The 2-119 bp region of the query sequence is 99% identical to the 368-485 bp segment of the human tyrosyl-tRNA synthetase gene. This is also confirmed by the 89% identity of this same region of the T7 end query sequence to the *Bos taurus* partial tyrosyl-tRNA synthetase gene (322-439 bp).
5. Analysis of selected sub-library clones

Figure 5.6.12

A1220-238 T3

|ttttttgcac tgattcagac ttaatggag gtgctcattt caagtcacca gaggctggtg caactggtga acgtgcatg
gggagtgaga ggtgctcttg gtcgagctgg aggaagaaca
gggaacctag ggttggggag agatgtatag aggaModele ccc |

The 163 bp T3 end sequence of Clone A1220-238.

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Score</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>4e-84</td>
</tr>
<tr>
<td>111</td>
<td>4e-84</td>
</tr>
<tr>
<td>102</td>
<td>6e-83</td>
</tr>
<tr>
<td>78</td>
<td>4e-13</td>
</tr>
<tr>
<td>78</td>
<td>4e-13</td>
</tr>
<tr>
<td>55</td>
<td>2e-09</td>
</tr>
<tr>
<td>60</td>
<td>1e-07</td>
</tr>
<tr>
<td>50</td>
<td>1e-07</td>
</tr>
<tr>
<td>40</td>
<td>0.089</td>
</tr>
<tr>
<td>33</td>
<td>0.35</td>
</tr>
<tr>
<td>36</td>
<td>1.4</td>
</tr>
<tr>
<td>16</td>
<td>1.4</td>
</tr>
</tbody>
</table>

BLAST analysis showed four high scoring Homo sapiens subject sequences which are clones containing the same coding sequence for part of the human neuron-specific gamma-2 enolase (ENO2) gene. The T3 end sequence is likely to be the 3' end of the coding region of the human ENO2 gene as denoted by the subject sequence of emb.no. Y00691/HSNSE which is the H. sapiens 3' mRNA for neuron-specific enolase. The next five subject sequence identities are conserved 3' end sequences of the neuron-specific enolase genes present in the mouse and rat species.
5. Analysis of selected sub-library clones

Figure 5.6.12 continued

The 8-163bp region of the query sequence is 100% identical to the 2119-2274 bp segment of the human neuron-specific gamma-2 enolase gene, indicating that the Clone A1220-238 is most likely to contain a part of this gene.

The human neuron-specific gamma-2 enolase gene maps to chromosome 12p13 and was isolated from brain tissue as shown in the above entrez nucleotide query data of M22349/HUMENOG.
5. Analysis of selected sub-library clones

identity to emb.no. Y00691/HSNSE which is the *H. sapiens* 3' mRNA for neuron-specific enolase.

The next five subject sequence identities are conserved 3' end sequences of the neuron-specific enolase genes present in mouse and rat. These show 85%-91% identity to the 11-73bp region of the human homologue present in the T3-end sequence. This conservation has been confirmed by the finding that the 5' and 3'-untranslated regions are similar (82% and 68%, respectively) to the analogous regions of the rat gamma-enolase gene, suggesting that a strong selective pressure operates on noncoding segments of gamma-enolase mRNAs (Oliva *et al.*, 1989).

Therefore the A1220-238 clone is probably the 3' end of the coding region of the human neuron-specific gamma-2 enolase gene, which has conserved regions in the mouse and rat. Since this gene is expressed in neuronal tissue and may have a homologue or related pseudogene within the Xp11.23 region studied, it was used in the mapping back studies (section 5.7).

**Description of six sequenced clones not used in mapping back studies**

**5.6.13 Clone 4HG2-49**

4HG2-49 contains a 200bp DNA insert. The clone contains a poly-A tail and 160bp of sequence was obtained at the T7 end. The T3 end sequence was not analysed because it was unreadable. BLAST analysis of the 160bp T7 end sequence shows sequence alignments to various human clones with similar scores and E-values as shown in Figure 5.6.13. Various size regions in the query sequence of 1-68bp, 22-68bp and 22-96bp are 90% identical to different segments of the subject sequences. These identical regions are repeatedly present in the subject sequences; an example of this is shown in the *Homo sapiens* DNA sequence from BAC 286B10 on chromosome 22 (emb.no. Z82244//HS286B10). These identical regions of the query sequence (in a range of 1-96bp) contain Alu repeat elements, which is indicated in the 9365-9431bp, 20821-20867bp and 45781-45855bp regions of this subject sequence (data shown in Figure 5.6.13).

In summary the 4HG2-49 clone contains Alu repeat elements in a region of 1-96bp, at the T7 end of the clone. These elements are repeatedly present throughout the reported human clones. The clone may contain part of a coding segment for an unidentified gene since the ORF finder showed one open reading frame (22-159bp). This clone was not included in the mapping back studies, but may be used in future studies.

**5.6.14 Clone 4HG2-51**

Clone 4HG2-51 contains a 450bp insert. The T3 and T7 ends of this clone gave 118 and 124 bases of sequence data respectively. A poly-A tail was found to be present at the T7 end of this clone.
5. Analysis of selected sub-library clones

Figure 5.6.13

4H49-T7

gagacagagt ctcaaccttt ttgcgccagg ctggagtgca ctggcgagat ctcggctcac tgcaaccttg 
acttcctggc ctcaggcaat cctcccggaa ttcagcttgg acttaaccag gcttgggcga gcgagttgaa 
ttcgatatac agccatcga

The 160 bp T7 end sequence of Clone 4HG2-49.

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Query</th>
<th>Score (bits)</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens DNA sequence from BAC 286B10</td>
<td>88</td>
<td>4e-16</td>
</tr>
<tr>
<td>Homo sapiens Chromosome 16 BAC clone CIT...</td>
<td>89</td>
<td>4e-16</td>
</tr>
<tr>
<td>Human chromosome 16p13.1 BAC clone CIT9875K...</td>
<td>86</td>
<td>2e-15</td>
</tr>
<tr>
<td>Homo sapiens ataxia telangiectasia (ATM) gen...</td>
<td>86</td>
<td>2e-15</td>
</tr>
<tr>
<td>Human BAC clone RG114A06 from 7q31, complete...</td>
<td>82</td>
<td>3e-14</td>
</tr>
<tr>
<td>Homo sapiens Xp22 BAC GS-607H18 (Genome Sy...</td>
<td>78</td>
<td>4e-13</td>
</tr>
<tr>
<td>Human Chromosome 11 pac pBR306B4, complete...</td>
<td>76</td>
<td>2e-12</td>
</tr>
<tr>
<td>Homo sapiens chromosome 19, cosmid R34094...</td>
<td>74</td>
<td>6e-12</td>
</tr>
<tr>
<td>Homo sapiens chromosome 19, cosmid 91K3, Hunting...</td>
<td>72</td>
<td>3e-11</td>
</tr>
<tr>
<td>Human BAC clone RG126M09 from 7q21-q22, ...</td>
<td>70</td>
<td>1e-10</td>
</tr>
</tbody>
</table>

BLAST analysis of the T7 end sequence showed sequence alignments to various human clones with similar low scores and E-values.
5. Analysis of selected sub-library clones

Figure 5.6.13 continued

_esp|HS286B10 Homo sapiens DNA sequence from BAC 286B10 on chromosome 22. TOM protein, HMOX1 gene for Heme Oxygenase 1 (HO-1, EC 1.14.99.3) and the alternatively spliced MCM5 gene for DNA Replication Licensing Factor MCM5 (P1-... Length = 87889

Score = 87.7 bits (44), Expect = 4e-16
Identities = 63/68 (92%), Positives = 63/68 (92%), Gaps = 1/68 (1%)

Query: 1 gagacagagttcactctctttttgcccaggctggagtgcactggcgcgatctcggctcac 60
Sbjct: 9365 gagacagagttcactctctttttgcccaggctggagtgcactggcgcgatctcggcttac 9423
repeat_region 9365..9423
/note="AluSp repeat: matches 302..2 of consensus"

Score = 54.0 bits (27), Expect = 6e-06
Identities = 42/47 (89%), Positives = 42/47 (89%)

Query: 22 ttgcccaggctggagtttcactggcgcgatctcggctcacgctgcaacct 68
Sbjct: 20821 ttgcccaggctggagtttcactggcgcgatctcggctcacgctgcaacct 20867
repeat_region 20821..20867
/note="AluSp repeat: matches 303..1 of consensus"

Score = 61.9 bits (31), Expect = 2e-08
Identities = 64/75 (85%), Positives = 64/75 (85%)

Query: 22 ttgcccaggctggagtttcactggcgcgatctcggctcacgctgcaacctgcaaccttgacttcctgggc 81
Sbjct: 45785 ttgcccaggctggagtttcactggcgcgatctcggctcacgctgcaacctgcaaccttgacttcctgggc 45796
repeat_region 45785..45796
/note="AluJo repeat: matches 14..302 of consensus"

Various size regions in the T7 end sequence of 1-68bp, 22-68bp and 22-96bp are 85%-92% identical to different segments of the subject sequences as shown in the example above in the Homo sapiens DNA sequence from BAC 286B10 (emb.no. Z82244/HS286B10). These identical regions of the query sequence contain Alu repeat elements, which are repeatedly present throughout the subject sequence as indicated in the 9365-9431bp, 20821-20867 and 45781-45855bp regions. Clone 4HG2-49 therefore contains an Alu repeat region in the 1-96bp at the T7 end.
Gapped BLAST analysis of the T3 end sequence shows three different subject sequence alignments as indicated by the BLAST search results of Figure 5.6.14A. These subject sequences contain sequence motifs of 17-20bp in size which are 100% identical to similar regions in the query sequence. Such identical sequence motifs could be either conserved or random sequences present in both query and subject sequences. These identical regions of the subject sequences do not contain Alu-repeat elements as indicated by their entrez nucleotide queries (data not shown). The ORF finder data indicates one open reading frame, suggesting that the T3 end sequence of the clone could be part of a coding segment of a gene.

Analysis of the T7 end sequence produced sequence alignments to various different human clones, all of similar scores and E-values as shown in Figure 5.6.14B. The T7 sequence of 1-124bp is 100% identical to similar regions in the subject sequences. This identical region contains an Alu repeat element as indicated by the entrez nucleotide queries of the subject sequences (data not shown). No open reading frame is present for the T7 end sequence, indicating that it is probably non-coding.

In summary the 4HG2-51 clone contains an Alu repeat element at its T7 end which also contains a sequence motif (either conserved or random), identical in both the query and subject sequences. Since the T3 and T7 sequences analysed described only half of the clone, the remaining sequence of the clone could contain a coding segment of an unidentified gene (one ORF at the T3 end, and a poly-A tail at the T7 end). This clone was not included in the mapping back studies, but may be further analysed in the future.

5.6.15 Clone 4HG2-140

4HG2-140 is a 200bp DNA fragment containing clone. The T7 end of the clone was not analysed as it gave un-readable sequence. The T3 end of this clone does not contain a poly-A tail and gave 189 bp of sequence. BLAST analysis of the 189bp T3 end sequence showed only one subject sequence alignment (Figure 5.6.15) of a human PAC clone DJO841B21 from chromosome 7q. The Entrez nucleotide query of this sequence (gb.no. AC004140) indicated no identity of the query sequence to any of the repeat regions present (entrez query data not shown). It is probably a similar sequence motif which is 100% identical, and present in both this subject sequence and a 16-33bp segment of the query sequence. The single subject sequence alignment reported here is of no biological significance.

In summary the 4HG2-140 clone contains DNA sequence which is found not to be similar to any of the reported GenBank database sequences. The ORF finder showed one open reading frame from 1-102bp, indicating that the T3 end sequence is possibly a part of the coding segment for an unidentified gene. This clone was not included in the mapping back studies.
5. Analysis of selected sub-library clones

Figure 5.6.14A

4HG2-51 T3

ttttagttat tatataaat aaaaaacaaa tcgtgtgtc
cgtaattgct tcgcaccttc gcgtgtgtct cgtgtgtca
ttcgtaa cacgacgacg gccgtcgtc cccgcttgggc

do The 118 bp T3 end sequence of Clone 4HG2-51.

<table>
<thead>
<tr>
<th>QUERY</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Score (bits)</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>0.063</td>
</tr>
<tr>
<td>34</td>
<td>3.9</td>
</tr>
<tr>
<td>33</td>
<td>3.9</td>
</tr>
</tbody>
</table>

Gapped BLAST analysis of the T3 end sequence showed three different subject sequence alignments containing sequence motifs of 17-20 bp which were 100% identical to similar regions in the query sequence. These identical sequence motifs are either conserved or random sequences present in both query and subject sequences.

Figure 5.6.14B

4HG2-51 T7

gagatggtgct ctggctctgt cgtcagcagc gcgacgacg
tgcacaatct cggctcactg caagctccgc ctcccaggtt
ccaagtgactc tcctgcccca gcctcctgag tagctggaacctaca

The 124 bp T7 end sequence of Clone 4HG2-51.

<table>
<thead>
<tr>
<th>QUERY</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Score (bits)</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>151</td>
<td>3e-35</td>
</tr>
<tr>
<td>147</td>
<td>4e-34</td>
</tr>
<tr>
<td>145</td>
<td>4e-33</td>
</tr>
<tr>
<td>145</td>
<td>4e-33</td>
</tr>
<tr>
<td>141</td>
<td>6e-33</td>
</tr>
<tr>
<td>145</td>
<td>6e-33</td>
</tr>
<tr>
<td>121</td>
<td>2e-26</td>
</tr>
<tr>
<td>115</td>
<td>2e-26</td>
</tr>
</tbody>
</table>

BLAST analysis of the T7 end sequence produced sequence alignments to different human clones, all of similar scores and E-values. The entire T7 end sequence of 1-124 bp was 100% identical to similar regions in the subject sequences and contained an Alu repeat element as indicated by the entrez nucleotide query data of the subject sequences (data not shown).
Figure 5.6.15

4HG2-140 T3

aggttcgact taagactcac aaaaagaagga agaggattcag actcgtcagt ctggctctcc ccaatcgctg ctggctggagt gggaaggtcg tcaacct

gaaagatct tccggtgcta ctgtttacc actccttaac tttgggtcag
tgtacctgta tggtgctcct tagacctgat acaaatc

The 189bp T3 end sequence of Clone 4HG2-140.

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Query</th>
<th>Score (bits)</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gb</td>
<td>AC004140</td>
<td>AC004140 Homo sapiens PAC clone DJ0841B21 from 7q21...</td>
</tr>
<tr>
<td>gb</td>
<td>AC004140</td>
<td>AC004140 Homo sapiens PAC clone DJ0841B21 from 7q21.1-q31.1, complete seq [Homo sapiens] Length = 74918</td>
</tr>
</tbody>
</table>

Score = 36.2 bits (18), Expect = 1.7
Identities = 18/18 (100%), Positives = 18/18 (100%)

Query: 16 ctcacaaaagaaggaagg 33
Sbjct: 19343 ctcacaaaagaaggaagg 19326

Gapped BLAST analysis of the T3 end sequence showed only one subject sequence alignment with human PAC clone DJ0841B21 from chromosome 7q21.1-q31.1. No repeat element was found within this 16-33bp segment of sequence from entrez nucleotide query data (not shown). It is probably a similar random sequence motif of 100% identity to the reported subject sequence and of no biological significance.
5.6.16 Clone 4HG2-201

4HG2-201 contains a 450bp DNA insert. The clone does not contain a Poly-A tail at the T7 end from which 156bp of sequence was analysed by the Gapped BLAST program. The T3 end sequence was not analysed because it did not give readable sequence. BLAST analysis of the 156bp T7 end sequence produced alignments to various human clones all of similar scores and E-values as shown in Figure 5.6.16. These subject sequences all contain an approximately 45-66bp segment of DNA which is about 90% identical to the 90-150bp region of the query sequence. This identical region of the query sequence is repeatedly found throughout the subject sequences, an example of which is shown in Figure 5.6.16 of the Homo sapiens 16 BAC clone CIT987-SK65D3 (gb.no. U95743/HUU95743). The repeated sequence is an Alu repeat element in the 77776-77831bp and in the 97578-97632bp regions of this subject sequence as indicated in the Entrez nucleotide query of gb.no. U95743/HUU95743.

Therefore the 4HG2-201 clone contains an Alu repeat region at 90-150bp, at the T7 end of the clone which is repeatedly present throughout the reported human clones. It does not show identity to any known gene, although it may contain a part of the coding segment for an unidentified gene, since the ORF finder showed one open reading frame (1-123bp). This clone was not included in the mapping back studies.

5.6.17 Clone 4HG2-215

The 4HG2-215 clone contains a 200bp DNA sequence amplified by PCR. The T7 end of this clone from which 68bp of sequence was generated has no poly-A tail. No readable sequence was obtained from the T3 end of the clone. BLAST analysis of the 68bp T7 end sequence produced alignments to various human clones all of similar scores and E-values as shown in Figure 5.6.17A. Six of these subject sequences contain an ~50bp DNA sequence which is 90% identical to the 6-55bp region of the query sequence. Four subject sequences contain a 30bp DNA sequence which is about 90% identical to the -3-27bp region of the query region. The region of ~6-55bp of this query sequence identical to the subject sequence of the human LD78 beta gene (dbj.no. D90145/HUMLD78B) at 712-761bp is indicated as an Alu repeat unit in the Entrez nucleotide query of this subject sequence as shown in Figure 5.6.17A. Also the ~3-27bp region of the same query sequence is repeatedly identical to similar regions throughout the subject sequence of the human ataxia telangiectasia (ATM) gene (gb.no. U82828/HSU82828) (Figure 5.6.17B). This repeated sequence was found to be an Alu repeat element in the 18612-18636bp segment of this subject sequence as indicated in the Entrez nucleotide query data (Figure 5.6.17B). This segment of the query sequence (~3-35bp) is also reported to be a Alu repeat region at the other repeated regions of the same subject sequence (data not shown).
Figure 5.6.16

The 156 bp T7 end sequence of Clone 4HG2-201.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Score</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>qbIU957 43</td>
<td>HUU957 43 Homo sapiens chromosome 16 BAC clone CIT987-...</td>
<td>115</td>
</tr>
<tr>
<td>qb</td>
<td>AF017257</td>
<td>AF017257 Homo sapiens chromosome 21 derived BAC con...</td>
</tr>
<tr>
<td>dbj</td>
<td>AP000031</td>
<td>AP000031 Homo sapiens genomic DNA, chromosome 21q1...</td>
</tr>
<tr>
<td>dbj</td>
<td>AP000047</td>
<td>AP000047 Homo sapiens genomic DNA, chromosome 21q1...</td>
</tr>
<tr>
<td>gb</td>
<td>AC004222</td>
<td>AC004222 Homo sapiens chromosome 17, clone HCIT499I...</td>
</tr>
<tr>
<td>gb</td>
<td>AC001164</td>
<td>HSAAC001164 Homo sapiens (subclone 2_e10 from PAC H9...</td>
</tr>
<tr>
<td>gb</td>
<td>AC002307</td>
<td>HUAC002307 Homo sapiens Chromosome 16 BAC clone CIT...</td>
</tr>
<tr>
<td>emb</td>
<td>Z93242</td>
<td>HS389A20 Human DNA sequence from PAC 389A20 on chrom...</td>
</tr>
<tr>
<td>emb</td>
<td>X68500</td>
<td>HS358RW H.sapiens genomic DNA sequence 358-RW</td>
</tr>
<tr>
<td>emb</td>
<td>Z95152</td>
<td>HS179N16 Homo sapiens DNA sequence from PAC 179N16 o...</td>
</tr>
</tbody>
</table>

Gapped BLAST analysis of the T7 end sequence produced sequence alignments to various human clones all of similar scores and E-values. These subject sequences all contained an ~ 45-66 bp size segment of DNA which was about 90% identical to the 90-150 bp region of the query sequence.
5. Analysis of selected sub-library clones

Figure 5.6.16 continued

sbl|U95743|HU95743 Homo sapiens chromosome 16 BAC clone CIT987-SK65D3,
[Homo sapiens]
Length = 192730

Score = 63.9 bits (32), Expect = 6e-09

Identities = 50/56 (89%), Positives = 50/56 (89%)

Query: 90  aggcgcctgtagctccccagctactcagaggctgagccaggagaatggctgaacc 145

Sbjct: 77776  aggcacctgtaatcccagctactccagaggctgaggcaggagaatcgcttgaacc 77831

Score = 77.8 bits (39), Expect = 4e-13
Identities = 51/55 (92%), Positives = 51/55 (92%)

Query: 90  aggcgcctgtagctccccagctactcagaggctgagccaggagaatggctgaacc 144

Sbjct: 97578  aggcgcctgtagctccccagctactcagaggctgagccaggagaatggctgaacc 97632

repeat_region 77625..77888 repeat_region 97798..97877
/rept_family="AluSqi" /rept_family="AluY"
repeat_region complement(79752..79830) repeat_region 97798..97877
/rept_family="(GGGA)n" /rept_family="MIR"

The identical region of 90-145 bp of the query sequence was repeatedly found throughout the subject sequences, an example of which is shown in Homo sapiens 16 BAC clone CIT987-SK65D3. This repeated sequence is an Alu repeat element in the 77776-77831 bp and 97578-97632 bp regions of this subject sequence as denoted in the entrez nucleotide query data of U95743/HUU95743.
Figure 5.6.17

4HG2-215 T7

cgggtgcgt tcacactcct gagctcaggc gatccgcagc
cgggtgcgt ccaagggctt gtctgggc

The 68 bp T7 end sequence of Clone 4HG2-215.

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Score</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbj</td>
<td>D80145</td>
<td>HUMLD78B Homo sapiens LD78 beta gene</td>
</tr>
<tr>
<td>gb</td>
<td>NM5623</td>
<td>HUMPLGDA Homo sapiens hydroxymethylbilane synthase gene</td>
</tr>
<tr>
<td>gb</td>
<td>U52683</td>
<td>HISTRF4B Human lymphocyte specific interferon regulatory factor</td>
</tr>
<tr>
<td>dbj</td>
<td>AP0000050</td>
<td>AP0000050 Homo sapiens genomic DNA, chromosome 21q11</td>
</tr>
<tr>
<td>gb</td>
<td>AP001359</td>
<td>AP001359 Homo sapiens cosmids LA0730 and ID0936 from chromosome 21q11</td>
</tr>
<tr>
<td>gb</td>
<td>UB228288</td>
<td>HUMUG288 Homo sapiens ataxia telangiectasia (ATM) gene</td>
</tr>
<tr>
<td>gb</td>
<td>AC002072</td>
<td>HUMAC002072 Human PAC clone DJ218B13 from Xq23, compatible with chromosome 4q25</td>
</tr>
<tr>
<td>gb</td>
<td>AC005100</td>
<td>AC005100 Homo sapiens chromosome 4q25, BAC clone B3011B20</td>
</tr>
<tr>
<td>gb</td>
<td>AC004509</td>
<td>AC004509 Homo sapiens chromosome 16, cosmid clone R0222H02</td>
</tr>
<tr>
<td>emb</td>
<td>Z84440</td>
<td>HS528L19 Human DNA sequence from PAC 528L19 on chromosome 5</td>
</tr>
</tbody>
</table>

BLAST analysis of the T7 end sequence produced sequence alignments to various human clones of similar scores and E-values. Six of these subject sequences contained an ~50 bp size DNA sequence which is about 90% identical to the 6-55 bp region of the query sequence. Four subject sequences contained a ~30 bp size DNA sequence which is about 90% identical to the -3-27 bp region of the query sequence.
The region of ~6-55 bp of the T7 end sequence was 94% identical to the 712-761 bp region of the human LD78 beta gene subject sequence (dbj.no. D90145/HUMLD78B) and which is indicated as an Alu repeat unit.

The ~3-27 bp region of the T7 end sequence is repeatedly identical to similar regions throughout the subject sequence of the human ataxia telangiectasia (ATM) gene (gb.no. U82828/HSU82828). This repeated sequence was found to be an Alu repeat element in the 18612-18636 bp region of this subject sequence. The 4HG2-215 clone contained an Alu repeat region of ~6-55 bp (at the T7 end) which is repeatedly present throughout the reported human clones.
In summary the 4HG2-215 done contains an Alu repeat region of ~6-55bp (at the T7 end of the clone) which is repeatedly present throughout the reported human clones showing identity to various human genes/clones. This end sequence does not contain a coding region since the ORF finder query showed no open reading frames. Although the clone contains an Alu repeat element at the T7 end, more sequence data needs to be obtained and analysed for a more definite conclusion to the identity of this clone. This clone was not included in the mapping back studies.

5.6.18 Clone  A1220-169

Clone A1220-169 contains a 300bp insert. The T3 and T7 ends of this clone were used to generate 207bp and 164bp of sequence respectively. A poly-A tail is present at the T7 end of this clone. Gapped BLAST analysis of the T3 end sequence shows alignments to various clones, all with similar scores and E-values as shown in Figure 5.6.18A. These subject sequences contain 18-22bp sequence motifs which are 100% identical to similar regions in the query sequence. These identical regions were found to contain Alu repeat elements as indicated by their respective entrez nucleotide queries (data not shown). The ORF finder query indicated three open reading frames, suggesting that the T3 end sequence of the clone could be part of the coding segment of a gene.

Analysis of the T7 end sequence also produced sequence alignments to various clones as shown in Figure 5.6.18B. Some of these subject sequences are similar to those produced by the T3 end sequence analysis in Figure 5.6.18A, because of a 72bp sequence overlap between the T3 and T7 end sequences. The subject sequences in Figure 5.6.18B contain 19-23bp sequence motifs which are 100% identical to similar regions in the query sequence. Some of these identical regions are Alu repeat elements and others contain identical random sequence motifs, present in both query and subject sequences (data not shown). The ORF finder indicated two open reading frames suggesting that the T7 end sequence could be part of the coding segment of a gene.

Therefore the complete sequence of clone A1220-169 contains various regions of Alu repeat elements and some sequence motifs which are identical to both query and subject sequences. It may also be a clone containing a gene not present in any of the reported subject sequences.

Table 5.2 summarises the sequence analysis for eighteen of the selected cDNA clones. Three of the twelve clones used in the mapping back studies showed sequence identity to reported tissue-specific genes isolated from neuronal tissue. The clones not used in the mapping back studies were small sized clones mainly containing Alu repeat regions and other random sequence motifs identical to those present in the reported subject sequences. Although some of these clones could be coding segments of unidentified genes, more sequence information is required before internal primers can
5. Analysis of selected sub-library clones

Figure 5.6.18A

A1220-169 T3
tagatatagc ggatggcagtc tggactca acgggtcag
ggggggtcct cttggtgccg cttggaagcg tataattctt
tcctgaggt tttagaaaaaa gaagacacg aagaaagcaa
gacagagt caaacagtgt aatcagaaga aagcaagaag
agaagcaaga cagagctca aacagtccaa tcagaggaga cctccca

The 207 bp T3 end sequence of Clone A1220-169.

Figure 5.6.18B

A1220-169 T7
tggaggagga gacaggagct ttcacccctg ccgggtgctc
agctggaggc aggaaagccg attccatagc tcagtgggag
gtctcctctg attacactgt ttgactttct gtcttgcttt
cctctgtct tctttctctt aacaatcagt agtagaatta tagc

The 164 bp T7 end sequence of Clone A1220-169.
### Table 5.2 Summary of sequence analysis of 18 selected cDNA clones

<table>
<thead>
<tr>
<th>Clone name</th>
<th>PCR size of clone</th>
<th>Size of end sequence of clone</th>
<th>ORF query data</th>
<th>Summary of Gapped BLAST sequence data analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clones used in mapping back studies</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Clone 4HG2-76</td>
<td>600 bp</td>
<td>T7 end - 171bp T3 end - 139bp</td>
<td>1 ORF 3 ORFs</td>
<td>T3 end sequence of clone contains repeat region, probably part of coding sequence for an unidentified gene.</td>
</tr>
<tr>
<td>2. Clone 4HG2-108</td>
<td>1 kb</td>
<td>T3 end - 145bp</td>
<td>1 ORF</td>
<td>T3 end sequence of clone contains repeat region, probably part of coding sequence for an unidentified gene.</td>
</tr>
<tr>
<td>3. Clone 4HG2-110</td>
<td>1.3 kb</td>
<td>T3 end - 185bp</td>
<td>1 ORF</td>
<td>Tentative sequence identity to a 22bp conserved segment of the rat tissue-type vomeronasal neurons putative pheromone receptor gene.</td>
</tr>
<tr>
<td>4. Clone 4HG2-125</td>
<td>200 bp</td>
<td>T3 end - 199bp</td>
<td>No ORF</td>
<td>Entire T3 end sequence of clone not similar to any of the reported subject sequences. Could be a non-coding sequence of an unidentified gene.</td>
</tr>
<tr>
<td>5. Clone 4HG2-162</td>
<td>300 bp</td>
<td>T7 end - 90bp</td>
<td>1 ORF</td>
<td>Sequence may contain a coding segment of an unidentified gene, not similar to any of the reported subject sequences.</td>
</tr>
<tr>
<td>6. Clone 4HG2-164</td>
<td>600 bp</td>
<td>T3 end - 184bp</td>
<td>No ORF</td>
<td>Entire sequence of clone 98% identical to a similar region in human chromosome 16, cosm id clone 378E2 (LANL).</td>
</tr>
<tr>
<td>7. Clone 4HG2-210</td>
<td>300 bp</td>
<td>T3 end - 105bp T7 end - 215bp</td>
<td>No ORF 1 ORF</td>
<td>Complete identity (100%) of entire sequence of clone to the human retinal arrestin (S-antigen) gene.</td>
</tr>
<tr>
<td>8. Clone 33CA11-239</td>
<td>900 bp</td>
<td>T3 end - 214bp</td>
<td>2 ORFs</td>
<td>Sequence motif (121-172bp) at T3 end identical to similar random motifs in 25 of the reported subject sequences. Identical regions are not repeat elements. Could be part of a coding segment of an unidentified gene.</td>
</tr>
<tr>
<td>9. Clone 33CA11-276</td>
<td>900 bp</td>
<td>T3 end - 199bp</td>
<td>1 ORF</td>
<td>T3 end sequence, part of the human NADPH dependant leukotriene B4 12-hydroxydehydrogenase gene. Isolated from kidney tissue. Contains conserved gene sequences to the pig, rabbit and rat cognate genes.</td>
</tr>
<tr>
<td>10. Clone 33CA11-283</td>
<td>700 bp</td>
<td>T3 end - 105bp</td>
<td>No ORF</td>
<td>Entire analysed sequence of 105bp is 97% identical to the 3' region of the human proteasome gene, which maps to chromosome 14q11.2. Conserved sequences to the human &amp; mouse immunoglobulin genes.</td>
</tr>
<tr>
<td>11. Clone A1220-164</td>
<td>250 bp</td>
<td>T7 end - 123bp</td>
<td>No ORF</td>
<td>Complete identity (100%) of T7 end sequence 99% identity to the human tyrosyl-tRNA synthetase gene. This gene maps to human chromosome 14q32. 89% identity to conserved gene sequence in Bos taurus species.</td>
</tr>
<tr>
<td><strong>Clones not used in mapping back studies</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Clone 4HG2-49</td>
<td>200 bp</td>
<td>T7 end - 160bp</td>
<td>1 ORF</td>
<td>Probably a clone containing an unidentified gene, with Alu repeat elements in a region of 1-96 bp. These Alu elements repeatedly present throughout the reported human subject sequences.</td>
</tr>
<tr>
<td>14. Clone 4HG2-51</td>
<td>450 bp</td>
<td>T3 end-118bp T7 end-124bp</td>
<td>1 ORF No ORF</td>
<td>Clone containing Alu repeat element at T7 end and 17-20 bp sequence motif at T3 end identical to similar random motifs in the subject sequences. Clone could also contain coding region of an unidentified gene.</td>
</tr>
<tr>
<td>15. Clone 4HG2-140</td>
<td>200 bp</td>
<td>T3 end - 189bp</td>
<td>1 ORF</td>
<td>Sequence in clone not similar to any of the reported GenBank data base sequences. Could be a coding region for an unidentified gene.</td>
</tr>
<tr>
<td>16. Clone 4HG2-201</td>
<td>450 bp</td>
<td>T7 end - 156bp</td>
<td>1 ORF</td>
<td>Clone containing Alu repeat repeat at 90-156 bp in T7 end sequence, which is repeatedly present throughout the reported human subject clones. Could be part of a coding segment of an unidentified gene.</td>
</tr>
<tr>
<td>17. Clone 4HG2-215</td>
<td>200 bp</td>
<td>T7 end - 68bp</td>
<td>No ORF</td>
<td>Clone containing Alu repeat region at 6-59bp (entire T7 end sequence), repeatedly present throughout reported human clones.</td>
</tr>
<tr>
<td>18. Clone A1220-169</td>
<td>300 bp</td>
<td>T3 end - 208bp T7 end - 163bp</td>
<td>3 ORFs 2 ORFs</td>
<td>Complete sequence contains many regions of Alu repeat elements and some identical sequence motifs to subject sequences. Could be part of a coding segment of an unidentified gene.</td>
</tr>
</tbody>
</table>
be designed which would give specific amplification products to non-repeat regions in the genomic DNA tested. Therefore, longer clones will need to be isolated from the selected sub-libraries and/or the starting retinal cDNA library, in order to obtain more sequence data for further analysis, before these can be used in mapping back studies.

Most of the subject sequences of the genes which showed ~100% identity to the query sequences from the clones, had high score values on BLAST analysis indicating that these genes may be of relevant biological significance. These genes also showed ~80% identity to similar genes from other species, indicating conservation. A number of these genes were isolated specifically from retinal and neuronal tissues, such as the human retinal arrestin gene (clone 4HG2-218) and the human neuron-specific gamma-2 enolase gene (clone A1220-238). Several clones contained sequence motifs which were identical to similar regions in genes reported as subject sequences. These could be randomly similar sequence motifs, or they may be conserved sequences of some biological significance. Such clones (e.g. clone 33CA11-239) were used in the mapping back studies, and further sequence data and analysis will be carried out if they map back to the region of interest. Sequences of clones containing repeat elements such as *Alu* and LINE1 were found to show subject sequence alignments to mainly human clones with similar repeat regions. This pattern was seen in most repeat-element containing query sequence BLAST results. Although these clones contain repeat regions, they could also contain coding regions for unidentified genes. Whilst the sequences from both ends of a number of clones do not contain poly-A tails, it cannot be assumed that these are not gene containing clones since the cloned cDNA sequence could be from the 5' end or the middle of a gene. Also, some sequence data could be from non-coding regions of a gene such as the 5' or 3' untranslated regions.

Overlapping clones can be determined by cross-hybridisation of individual clones or by sequence analysis; the latter being the preferred course in this study. For example clone 33CA11-239 of 900bp was similar on BLAST homology analysis to clone 33CA11-210 of 600bp (data not shown). These clones are probably overlapping clones encoding the same transcript. Sequencing indicated that the majority of clones selected were independantly derived. In each transcriptional unit, exon and cDNA sequences have been compared with each other for overlap. While sequencing data has not added much to the understanding of the sequences, their determination is a critical first step in analysis.
5. Analysis of selected sub-library clones

5.7 Mapping studies of twelve selected cDNA clones

Any unique clones from a selected sub-library can be confirmed to lie in the genomic contig analysed through mapping back studies. In order to rapidly characterise both the enriched material and the resulting clones different mapping strategies can be used. Such strategies include hybridisation of the cDNAs as probes onto dot-blots and pulsed field gel blots containing different YAC genomic DNAs and appropriate negative and positive control DNAs. Similar hybridisation studies on Southern blots using digested total human genomic DNA, YAC DNAs and DNA from a somatic hybrid cell line, which is known to harbour a chromosomal region corresponding to the YAC of interest, could also prove useful. PCR analysis of overlapping YACs used in the selection with the appropriate positive and negative controls, using internal primers from the cDNAs studied was the preferred strategy of mapping the chosen twelve clones. This is because PCR is quicker and more specific in comparison to hybridisation analysis. Also hybridisation of repetitive elements and other contaminating sequences such as vector sequences found within the cDNA probe is eliminated.

5.7.1 Choice, design and testing of internal primers of twelve selected cDNAs

Initially, twelve clones were chosen for mapping back studies since six of the remaining clones showed some similarities to Alu or repetitive elements and/or were also of a small size of 100-300 bp. When more sequence information has been obtained reliable internal primers can be designed for future mapping back studies. Internal primers were designed for the twelve cDNA clones (sections 5.6.1-5.6.12) internal to the pDIRECT cloning site and the λgt10 forward sense and reverse antisense primer sequences (sections 5.3.1.1). These primers were designed at the 3' end of the cDNA clones since the 3' end of a gene is more specific. The primers were approximately 20bp in length and were chosen with a random base distribution avoiding stretches of polypurines and polypyrimidines. Both the forward and reverse primers were checked against each other for complementarity in order to reduce the incidence of primer dimers. The internal forward and reverse primers for the twelve cDNAs chosen for mapping back studies are described below:

4HG2-76 Forward primer:

5'- CAG CAT TAT TAC ATT GTA TGT CC -3'

4HG2-76 Reverse primer:

5'- GAT AAG TGC ATG GTT TCA AGC -3'
4HG2-108 Forward primer:
5' - AGT ATT GGT TCT TCA AAC TTA CG -3'
4HG2-108 Reverse primer:
5' - AGA AGT CGA CGA GTG TCC TC -3'

4HG2-110 Forward primer:
5' - CAG ACA TTA GGG TCG ATG AAC -3'
4HG2-110 Reverse primer:
5' - CAC TCA TCG ACC TTG ATG TC -3'

4HG2-125 Forward primer:
5' - TCG ACT TAA GGC GCC TTA AG -3'
4HG2-125 Reverse primer:
5' - TGA ATT GGT CCG ACT TGA ACG -3'

4HG2-162 Forward primer:
5' - TGT CCT GCG TGA CGA AAG C -3'
4HG2-162 Reverse primer:
5' - CGG ACT GGA ACC GTC TGA G -3'

4HG2-164 Forward primer:
5' - AAG CAC CAG CAG AGA GCT C -3'
4HG2-164 Reverse primer:
5' - GTC AGG TGG CTT AGC CAT TG -3'

4HG2-218 Forward primer:
5' - CTA AGG AAA GTT ATC AGG ATG C -3'
4HG2-218 Reverse primer:
5' - ATC AAT GTC ATT CTT GTC TCT C -3'

33CA11-239 Forward primer:
5' - TTC ACT AAC CTT CCC TAG GC -3'
33CA11-239 Reverse primer:
5' - CTA TGC TGG AAT GGC TGA AG -3'

33CA11-276 Forward primer:
5' - CCG GTT CTG TTA TAT GTA GAG -3'
33CA11-276 Reverse primer:
5' - TTA ACT ACA AGA CGG TAG AGT C-3'
5. Analysis of selected sub-library clones

33CA11-283 Forward primer:
5'- TGA ATA GTG TAA TGT TAA CAT CC- 3'
33CA11-283 Reverse primer:
5'- CTC TGG CAC ATT GAA CCT CT -3'

A1220-164 Forward primer:
5' - AGC TCA GCA AAG AGT ACA CAC -3'
A1220-164 Reverse primer:
5' - ACG AGG TGG ACG AAA TGG TGG -3'

A1220-238 Forward primer:
5' - ATT TCA ATG CCA CAG AGG TGG -3'
A1220-238 Reverse primer:
5' - TAG GTT CCC TGT TCT TCC TC -3'

The internal primers were tested on human genomic DNA, the cloned cDNA as a positive control, the rhodopsin cDNA used as a negative control and a no DNA control. PCR reactions were carried out at different annealing temperatures (54°C, 56°C and 58°C) and Mg+ concentration in order to optimise the PCR conditions.

5.7.2 Mapping studies of twelve selected cDNAs to the YAC genomic contig

Overlapping YACs 4HG2, 33CA11, A1220 and C1228 (section 4.1.6) used in the cDNA selection procedure and from which the selected cDNAs were isolated, were used in the mapping back studies. The YAC DNAs were tested with their respective internal markers to check that they were amplifiable by PCR and not degraded. YACs A1220, C1228 were tested with the marker DXS337 and with the TIMP-1 gene exon 5 primers, and YACs 33CA11, 4HG2 with the DXS426 marker. All tested YACs gave positive results with their respective internal markers (data not shown). However, all twelve cDNAs were found not to map back to any of the four overlapping YACs, giving negative results with each of the twelve different internal primers used.

An example of the mapping studies using PCR analysis is seen in Figure 5.7, which shows mapping studies of cDNA clone 4HG2-218. No bands were obtained with any of the four overlapping YACs tested or the negative control rhodopsin YAC. No amplification product was obtained using the total X chromosome DNA indicating that the 4HG2-218 clone does not map to the X chromosome. Although the mapping back of the twelve cDNAs gave negative results, two other selected genes gave positive results. The internal positive control genes TIMP-1 (enriched 1,000-fold) and properdin were
Figure 5.7 Mapping studies of cDNA clone 4HG2-218 using PCR analysis.

The figure shows an ethidium bromide stained gel photograph of PCR products of tested DNAs fractionated on a 1.2% agarose gel. The four YACs A1220, C1228, 4HG2 and 33CA11 (Lanes 1-4) and the negative control of the rhodopsin YAC (Lane 5) showed no amplification products. A negative result was also obtained for total X chromosome somatic cell hybrid DNA (Lane 6). Positive results were seen in genomic DNA (Lanes 8-10) giving a single band of size ~110bp when compared to the marker sizes of the OX174/HaeIII molecular weight marker (Lane 12). A similar sized product was obtained with the positive control DNA from the cDNA clone 4HG2-218 (Lane 11), which indicates the size of the expected amplification product of ~110bp. A negative result was seen in the no DNA track (Lane 7), indicating absence of contaminating DNAs in the PCR reaction mix. The 4HG2-218 cDNA clone which is 95% identical to the S-arrestin retinal gene does not map back to the YAC genomic contig used in the selection procedure (section 4.1.6, Figure 4.2) or to the X chromosome.

C1, C2 & C3 - Genomic DNAs (controls)
found to be enriched and present in selected YAC A1220 (section 5.4.2.1) and YAC 33CA11 (section 5.4.2.3) sub-libraries respectively, indicating that these two retinally expressed genes (section 4.2.1.2) did map back to their respective YACs. There is a possibility that the two A1220 cDNAs which did not map back to the YAC of origin could have been selected from the chimeric end of the A1220 YAC (section 4.1.6). These and the ten other cDNAs used in mapping studies must be non-specific cDNAs which came through during the selection procedure. The twelve cDNAs represent only a small percentage of the selected cDNA clones, since about 240 cDNAs (section 5.5) inclusive of overlapping clones, pseudogenes and genes from multigene families are expected to be present in the total selected YAC sub-libraries. The six remaining clones (section 5.6), and any other interesting cDNAs obtained after further sequence analysis (section 5.5.2) will be used in similar mapping studies to determine any potentially novel genes.

5.8 Discussion

The cDNA selection method used in this study has the potential to normalise transcript levels, and hence enrich rare cDNAs to a greater extent than the abundant ones. This was the case for the rare TIMP-1 cDNA with a $1 \times 10^3$ enrichment in the A1220 YAC secondary selected cDNAs (section 5.4.2.1), and a 70-fold level of enrichment for the more abundant rhodopsin cDNA transcripts isolated from the rhodopsin YAC secondary selected cDNAs (section 5.4.2.2). The cDNA selection approach has previously been shown to enrich for clones present at 1 part per 200-500,000 in starting oligo(dT) cDNA libraries. The results obtained in this study confirm this estimate of sensitivity in that the TIMP-1 cDNA clone which was represented at less than 1 part in 200,000 in the λgt10 oligo(dT)-primed library was recovered. In a study by Parimoo et al., (1993), the frequency of the TIMP-1 cDNA clones from the TEMP YAC DNA selected sub-library was 5% whereas the abundance in the initial starting cDNA library was 0.01%, hence giving a 500-fold enrichment, 50% lower than that obtained in this study for the same gene.

The hybridisation conditions were designed so that the genomic DNA was present in limiting amounts (sections 4.2.4.2 & 4.2.5.2), and some level of cDNA abundance normalisation is theoretically possible. In this scenario, a large excess of cDNAs (1μg) play a game of musical chairs with a limiting amount of genomic DNA (100ng). In theory, each genomic locus hybridises to one homologous DNA molecule. Thus when the genomic 'chairs' are limiting, most copies of an abundant cDNA are left without 'chairs' and are not selected. However, low-abundance cDNAs are present at closer to a 1:1 ratio with their genomic 'chairs', and most of these cDNAs are therefore selected. The consequence of this is a net decrease in the relative abundance of the high-
5. Analysis of selected sub-library clones

abundance cDNAs such as rhodopsin and an increase in the relative abundance of the low-abundance cDNAs. If three rounds of selection and amplification are carried out, it should be possible to identify transcripts represented at a level of well below 1 part in a few million, the practical limit for the application of direct library screening techniques. Since the TIMP-1 low abundant retinal gene was present at a level of 1 in 2 X 10^5 in the starting λgt10 retinal cDNA library (section 5.2), the second round of selection and amplification was sufficient for a dramatic thousand fold enrichment in selected cDNA sub-libraries (section 5.4.2.1).

Some YACs may contain region-specific low-level repeats which are not competed out by the quenching cocktail. In such cases, a significant number of cDNAs with homologies to such repeats may be selected. One way of overcoming this problem would be to hybridise the selected library colonies to such repeats and to remove them physically. Not all cDNAs that contain repetitive elements (section 5.4.1.1) are necessarily non-specific background; the cDNA may have hybridised within a single-copy region that contains a repetitive element. Proving the identity of such a cDNA can be difficult though, and usually involves repeated suppression or sub-cloning of the selected cDNA. If the entire DNA sequence is determined, the cDNA can be localised by PCR using primers derived from non-repetitive stretches of sequence. In general, when choosing cDNAs to analyse further, it is better to choose the non-repeat containing cDNAs as a first priority and then return to the cDNAs containing repeats if there is reason to believe they may prove fruitful. It has been reported that 20% of the human genome (Britten et al., 1988) contain highly repetitive DNA, the major component of this type of sequence being Alu repeats. 6% of the screened cDNA clones in the sub-libraries of this study were human repeats (section 5.4.1.1), indicating a reduction in repeat sequences through the cDNA selection procedure. Studies on the frequency and position of Alu repeats in cDNAs as determined by database searches, suggest that 5% (Yulug et al., 1995) of fully spliced human cDNAs contain Alu sequences. The vast majority of Alus are found in the 3' untranscribed region (UTR), but 14% lie in the 5' UTR and rarely an Alu sequence is present within or partially within, the coding region of transcripts (Yulug et al., 1995). Since some cDNA libraries contain unspliced or partially spliced message then the high-copy repeats in the introns could account for the repeats within the cloned cDNA transcripts.

The ribosomal sequence contaminants peculiar to YAC selections (section 4.1.3) are found in cDNA libraries and primary uncloned cDNAs which have not been pre-blocked with ribosomal DNA (section 4.1.8). Such ribosomal sequences were found at a reduced level of 2% (Table 5.1) in the total secondary selected cDNAs in this study (section 5.4.1.2). This was probably due to the ribosomal DNA pre-blocking steps (sections 4.2.4.1 & 4.2.5.1) during the cDNA direct selection procedure. Another reason for the reduced level of ribosomal sequences is the purification of YAC genomic DNA on
5. Analysis of selected sub-library clones

CHEF gels (section 4.2.2.1). Whilst this procedure was time consuming, it probably eliminated the high backgrounds that result from the presence of the entire yeast genome. The direct selection would certainly still work even in cases where good separation of the YAC from the yeast chromosomes is not possible, and the YAC is substantially contaminated with one or more yeast chromosomes. This was the case with YAC 33CA11 in this study which co-migrated with native yeast chromosomes and was indistinguishable on the gel (section 4.2.2.1), but still showed good enrichment for the positive control reporter gene properdin (section 5.4.2.3). Most cDNA direct selections to date have used purified YACs or cosmids as genomic targets. Collections of purified YACs or binned sets of cosmids (cosmids that are known to be derived from a defined part of the genome but are not arranged into a linear contig) can also be used as targets. Most regions of the human and mouse genomes are not covered by cosmid contigs or binned sets of cosmids, but are covered by YACs as was the case for the genomic contig of 500kb at Xp11.23 used in this study, although using cosmids might have given higher levels of enrichment and lower backgrounds (Lovett, 1994).

No figures has to date been reported for contaminant vector clones present in selected sub-libraries using YAC genomic DNA during direct selection. The reason for the high percentage of 60% contaminant vector clones found in the selected sub-libraries (section 5.5.3) in this study is not clear and can only be postulated. Just as ribosomal cDNAs can comprise up to 70% of selected material when YAC genomic DNA is not pre-blocked for ribosomal sequences (Morgan et al., 1992), YAC genomic DNA is able to hybridise to homologous vector sequences found in the starting λgt10 cDNA library during selection when the YAC vector is not pre-blocked with pYAC vector DNA. This may have contributed to the high selection of vector-containing clones.

Some cDNAs can be highly over-represented in the sub-library, possibly to the detriment of less abundant clones. The frequencies of the cDNAs in the sub-library could reflect their frequencies in the starting material or, more likely, the efficiency of their hybridisation, amplification or cloning. The ratio of cDNAs recovered can be expected to depend to some extent on the exact conditions used during selection. Considerations governing the recovery of cDNAs will include the ratio of target molecules (exons in the YAC genomic pool) to molecules of selectable cDNA, which will depend on the abundance of the cDNA in the amplified insert preparation. This abundance will vary due to the effect of the selection in the previous round, the length and GC content of the exons, and the conditions of the hybridisation. Obviously the influence of different exon length will however remain, leading to less efficient recovery of genes containing very short exons. Artifacts observed with PCR such as point mutations are unavoidable, but do not affect the feasibility of this method as it still provides the primary function of generating probes for the isolation of cDNAs.
The enrichment procedure generated overlapping cDNAs (section 5.6), a feature which can be exploited to reconstruct longer cDNA sequences in transcriptional units. Nevertheless, the search for full length transcripts will necessitate the screening of total cDNA libraries with the partial coding elements already isolated. To allow more precise characterisation of the isolated cDNAs and identification of overlapping clones, any mapped cDNAs will be hybridised back to the sub-library filters for preliminary mapping. In parallel selected clones will be sequenced. This hybridisation and sequencing analysis can lead to the identification of several small contigs of overlapping cDNAs. The short fragment cDNA clones derived from each selection can be used as probes to isolate longer clones from the λgt10 oligo(dT)-primed cDNA library. Each oligo(dT)-primed cDNA insert can then be hybridised to the inserts obtained with all other short fragments from the same YAC (section 5.5.1). In cDNA selection experiments, genes will often be recovered as several fragments that may or may not be overlapping. By examining a dispersed set of genomic clones it is likely that each recovered fragment will originate from a different gene. In this way one can maximise the number of genes identified for a given quantity of genomic DNA.

PCR analysis of overlapping YACs in the genomic contig using internal primers from the selected cDNAs (section 5.7.1) was the preferred mapping method as opposed to hybridisation analysis. This is because evaluation of the selected cDNAs can prove difficult if the cDNAs consistently hybridise to similar-sized fragments in rodent and human DNAs, making it difficult to map the gene back to a specific chromosome. In such instances, a different restriction endonuclease digestion can prove helpful. A more frustrating instance could be when a cDNA detects several genomic fragments, only one of which segregates with the YAC. This leads to the question of whether a low-copy-number repeat, a gene family, or a pseudogene is being detected. In this case the DNA sequence can be determined, which should reveal whether the cDNA contains a known repeat or is a member of a known gene family. Assessing whether a pseudogene is present in the YAC that selected a cDNA from the expressed locus is somewhat more difficult and involves, at the very least, sequencing the genomic locus from within the YAC.

The cDNA selection scheme has been applied to the isolation of cDNAs that are encoded by a ~500 kb YAC contig and has identified the rare TIMP-1 gene in one out of 200 selected cDNAs (section 5.4.2.1). This demonstration of the sensitivity and selectivity of the technique naturally leads to the question of how many other low abundant retinal genes might be residing within the YAC contig studied. Some estimate of the number of genes that might be encoded by a 500-kb genomic clone can be gained by considering estimates of total gene number and gene distribution (Bickmore & Sumner, 1989). Based on such estimates, the 500-kb YAC genomic contig (section 4.1.6) might be expected to encode 10 or more genes. The number of genes in a particular YAC may be
overestimated if several short fragment sequences originate from the same message without forming an overlap. This problem can be dealt with in several ways, such as by RNA blots or sequencing of full-length cDNAs. The generation of hybrid cDNA molecules as a result of co-ligations or during PCR is again not a problem because they are either taken care of during the second round of selection or identified when sequencing full-length cDNAs.

Five of the twelve cDNA clones used in the mapping back studies contained query sequences which were ~100% identical to their respective subject sequences of genes. (Table 5.2). None of these clones mapped back to the 500kb genomic region at Xp11.23 used in the selection procedure (section 5.7). Although two of the clones A1220-164 and A1220-238 were derived from the chimeric YAC A1220 (section 4.1.6) they did not map back to the fraction of chromosome found at the chimeric end of this YAC. The chromosomal origin of the chimeric end of the A1220 YAC has not yet been determined by somatic cell hybrid panel mapping or by in situ hybridisation techniques (section 1.4.1). However both these clones may be non-specific cDNAs which could have come through during the cDNA selection procedure. The remaining three cDNA clones of 4HG2-218, 33CA11-276 and 33CA11-283 (Table 5.2), which did not map back to the region of interest were isolated from non-chimeric YACs 4HG2 and 33CA11 (section 4.1.6). They are probably also non-specific cDNAs which had come through during the selection procedure. This may especially be true for the 4HG2-218 cDNA of S-antigen/arrestin which maps to human chromosome 2q and is a retinal specific gene which may have enough abundance of transcripts in the starting retinal cDNA library to be non-specifically isolated during the selection procedure.

Further investigations on any interesting sequence from analysed selected cDNAs could be carried out by different approaches such as analysis of the corresponding genomic sequence to determine if the cDNA was selected by a true gene or a pseudogene present on the target chromosome; or by using this complex mixture of cDNAs to probe a normalised, cDNA library spotted on high density filters (Soares et al., 1994). This second method should allow direct access to full-length cDNA clones and avoid redundant sequencing of cDNAs. For further analysis of any putative novel genes expression studies such as Northern analysis and RT-PCR should be carried out for a subset of fragments. The subset of fragments, both positive and negative in Northern analysis, can then be used to screen large insert cDNA libraries, though interestingly frequency in cDNA libraries does not always accurately reflect Northern results (Tassone et al., 1995), which most likely reflect biases in various cloning steps.

In summary, the present results confirm that cDNA selection is a highly efficient method for obtaining probes from the mRNAs encoded in genomic fragments of several hundred kilobases in size. It is limited as a method for enumerating the genes contained in such inserts because several non-overlapping short fragments that derive from
different regions of the same mRNA may be obtained. Overall, with present automated DNA sequencing methods it may be more efficient in terms of data generation to examine a 100 or 200 kb genomic fragment by sequencing rather than by cDNA analysis. For much larger regions the advantages of cDNA selection in obtaining rapid access to the coding sequences seem substantial.

5.9 Conclusion

The technique of cDNA hybridisation selection was applied to four overlapping YAC clones (500kb) mapping to the Xp11.23 region, which contains part of the XLRP2 critical locus. A preliminary hybridisation analysis of the selected cDNAs has proved the value of this method for rapid and effective isolation and normalisation of highly represented genes such as rhodopsin and has achieved a 1000-fold enrichment for the rarer TIMP-1 gene. Eighteen distinct fragments of cDNAs have been identified on the basis of hybridisation and sequence analyses. Additional data was also obtained based on sequence comparisons of selected cDNAs with cDNAs and genes described in the GenBank database. Mapping studies on twelve of these selected cDNAs using PCR analysis indicated that none of these cDNAs mapped back to their YAC of origin. These cDNAs have hence been non-specifically selected for the region investigated. The high level of vector clones (60%) can be reduced by pre-blocking the YAC genomic DNA with pYAC vector DNA, in order to reduce hybridisation to homologous vector sequences found in the starting cDNA libraries. This could then increase the level of enrichment of specific retinal genes found in the region since the high background of vector clones would be reduced. The level of enrichment of rare retinal genes would then increase and the non-specific selection of retinal genes such as S-antigen (4HG2-218 cDNA) be decreased. An exhaustive screening of the selected sub-libraries of cDNAs per 100kb of YAC DNA may identify a greater number of specific retinal cDNAs/genes mapping to the region of study and may result in a comprehensive transcriptional map of this 500kb region at Xp11.23.
CHAPTER SIX

Investigation of the human TIMP-1 gene for a causative role in the pathogenesis of X-linked retinitis pigmentosa-2
CHAPTER SIX

INVESTIGATION OF THE HUMAN TIMP-1 GENE FOR A CAUSATIVE ROLE IN THE PATHOGENESIS OF X-LINKED RETINITIS PIGMENTOSA-2

6.1 Introduction

Cell-cell contacts in multicellular organisms are mediated through the extracellular matrix (ECM), which consists of a highly complex aggregate of secreted proteins and carbohydrates that not only provides structural support to cells, but also has a profound influence on many biological activities. In response to normal developmental changes and tissue turnover, the components of the ECM are continuously subjected to degradation and re-synthesis. Disturbances in these dynamic processes may lead to loss of ECM integrity, which is thought to play an important role in many pathologic conditions (Docherty et al., 1992).

The coordinated remodelling of the ECM requires the tightly controlled activity of a number of proteinases, including the family of matrix metalloproteinases (MMPs) (Matrisian, 1990). These are a family of Zn$^{2+}$-dependent endopeptidases which include interstitial and neutrophil collagenases, gelatinases A and B (72kDa and 92kDa gelatinases) and stromelysins (Matrisian, 1990). These MMPs are able to degrade all components of the ECM and the basement membranes (Matrisian 1990). They have overlapping specificities, which taken together include many important constituents of the ECM, including many types of collagens, gelatins, elastin, fibronectin and proteoglycans (Woessner, 1991).

Matrix turnover is a normal component of physiological remodelling processes and is controlled by the resident connective tissue cells. Metalloproteinases produced by these cells are tightly regulated at several levels, including gene expression and the extracellular activation and inhibition of the secreted proteinases (Murphy and Docherty, 1988). Controlling the activity of the MMPs has direct and indirect effects on the structural integrity of many macromolecules (Desrochers et al., 1991). By virtue of their ability to degrade the extracellular matrix, the metalloproteinases participate in a variety of biologically important processes. These enzymes have been found to be important for developmental growth, wound repair, metastasis, uterine involution, and pathological processes (Matrisian and Hogan, 1990).
The activity of these proteases in the extracellular environment is controlled by specific and potent inhibitory proteins known as tissue inhibitors of metalloproteinases (TIMPs). The TIMPs interact with specific MMPs to maintain connective tissue homeostasis. Imbalances between the metalloproteinases and their inhibitors and activators, leading to increased ECM degradation, have been implicated in, for example, osteoarthritic arthritis and tumour invasion (MacNaul et al., 1990).

To date, three members of the TIMP gene family have been identified: TIMP-1 (Docherty et al., 1985), TIMP-2 (Boone et al., 1990), and TIMP-3 (Apte et al., 1994; Silbiger et al., 1994). TIMP-1, is a ubiquitous 28.5 kDa secreted glycoprotein that forms tight stoichiometric non-covalent complexes with the active forms of all known MMPs and in addition binds to progelatinase B (Wilhelm et al., 1989). The secreted protein is comprised of 184 amino acids and contains six intramolecular disulphide bonds (Stevenson et al., 1990). The same cells which produce interstitial collagenase are capable of synthesizing and secreting TIMP-1 (Herron et al., 1986). TIMP-2, is a 21.5 kDa non-glycosylated protein which has 40% sequence identity with TIMP-1, and has a similar inhibitory activity against MMPs but preferentially binds to progelatinase A (Stevenson et al., 1990). TIMP-1 has been localised to chromosome Xp11.23-Xp11.4 (Willard et al., 1989), TIMP-2 to chromosome 17q25 (DeClerck et al., 1992), and TIMP-3 to chromosome 22q12.1-q13.2 region of human chromosome 22, using in situ hybridisation and human-hamster somatic cell hybrid DNAs (Apte et al., 1994).

The TIMP-1 gene spans 4.5 kb of genomic DNA and is situated within intron 5 of the SYN1 gene on chromosome Xp11.23 (Derry et al., 1992 & Hagemann et al., 1994). The TIMP-1 gene is found to lie in the RP2 critical interval between flanking markers DXS8083 and DXS6616 (discussed in Chapter 3, section 3.2.2) as shown in Figure 3.1.

TIMP-1 is reported to be identical to a growth factor like activity for erythroid cells [erythroid-potentiating activity (EPA)] which is produced by human T-cell leukemia virus type 2 (HTLV-2) infected T-cell lines (Uchijima et al., 1994). Thus, TIMP-1/EPA is thought to be a multifunctional protein regulating turnover of the extracellular matrix on the one hand and modulating growth properties of cells on the other (Hayakawa et al., 1992).

As well as being a positional candidate for RP2, recent investigations of TIMP-3 in retinal disease support the notion that TIMP-1 may be involved in XLRP. Firstly retinal TIMP-3 mRNA levels are dramatically increased in cases of simplex retinitis pigmentosa (RP), suggesting that extensive remodelling of the retinal ECM is occurring in this ocular disease (Jones et al., 1994). Secondly mutations in the TIMP-3 gene on chromosome 22q13.1 have been identified as the cause of Sorsby's fundus dystrophy (SFD) (Weber et al., 1994), an adult onset macular degeneration of the neural retina, pigment epithelium and choroid (Sorsby et al., 1949). SFD is characterised by complications arising from ECM disturbances in Bruch's membrane (Capon et al., 1989).
The association of TIMP-3 mutations with the pathogenesis of SFD has emphasized further the importance of the inhibitors in ECM homeostasis. This led to an investigation of the possible role of TIMP-1 in RP2.

6.2 Expression of human TIMP-1

Immunohistochemistry has demonstrated TIMP-1 protein expression in all nuclear layers of the retina (Johnson et al., 1995). To corroborate TIMP-1 expression in the retina, the cDNA was successfully amplified by PCR from phage lysates (as described in section 4.2.1.2) of two adult retinal cDNA libraries, λZAPII and λgt10 (section 4.1.7) using primers designed to the TIMP-1 cDNA sequence. Using primer pair sequences of exon 3f-5'-GGCTCATGCAGTCCATTTGAC-3' and exon 5r-5'-GGAAGGATCTCCCGTTGGAG-3' the expected 613bp fragment covering exons 3 to 5 was present in both libraries, confirming expression of this mRNA in the retina. This is clearly evident in Figure 4.3B which shows the PCR product obtained from the λgt10 cDNA library (section 4.2.1.2).

The amplified TIMP-1 cDNA was also used as a probe against the λgt10 cDNA library as described in the earlier section 5.2, to determine the prevalence of the TIMP-1 transcript in retinal tissue. One positive TIMP-1 plaque was identified out of ~200,000 plaques (Figure 5.1A, section 5.2), compared to 1,200 positives identified with a rhodopsin cDNA probe (Figure 5.1B, section 5.2), implying that the TIMP-1 transcript is probably rare in the retina.

6.3 Determination of the genomic structure of human TIMP-1

In order to perform an effective mutation screen of all exons the genomic structure of the TIMP-1 gene was determined. Several entries for human TIMP-1 (or EPA) in GenBank were retrieved; X03124, the original cDNA sequence; D11139, a partial genomic sequence encompassing exon 3 to exon 6; and D26513, the 5' upstream region containing CAT, TRE and Sp1 sites. Positions, sequence and size of introns 1 and 2 were determined. After inferring much of the genomic structure of the gene from the aforementioned database entries and donor/acceptor splice site consensus sequences, exon-specific primers were designed to amplify introns 1 and 2. Primer pair I1f-5'-CCAGCGCCAGAGACACC-3' and I1r-5'-TGCCAGAAAGCAGGGCTCA-3' designed from database entries X03124 and D26513 were used to amplify the region of TIMP-1 which potentially spans intron 1, from genomic DNA. Another primer set, I2f-5'-CCACAGAGGCGCTTCTGCAAT-3' and I2r-5'-GGCTGGTTGACTTTCTGCGGTCC-3'
designated from database entry X03124, was used to amplify a region of genomic DNA containing intron 2.

Amplified introns produced fragment sizes of approximately 1kb spanning intron 1 and 1.6kb spanning intron 2. The PCR products were purified, (Section 2.8.3.1) and partial sequences of introns 1 and 2 were obtained by direct sequencing (Section 2.8.3) using the primers above. Consensus splice sites (Mount, 1982) confirmed intron-exon boundaries throughout this gene, consisting of six exons and spanning ~4.5kb. Accession numbers for partial sequences of introns 1 and 2 are L47357 and L47361. The genomic structure of the human TIMP-1 gene is described in Table 6.1.

6.4 Mutation screening in patients with Retinitis Pigmentosa 2

DNA from affected male patients and unaffected males from potential RP2 families XLRP 88/75 (Figure 3.13, section 3.3.12) and XLRP 89 (Figure 3.5, section 3.3.4) were selected for TIMP-1 mutation screening. Each family had been previously defined as most likely segregating with the RP2 form of XLRP by detection of key recombination events during haplotype analysis (Chapter 3). Primer pairs designed are shown in Table 6.2, covering all 6 exons, exon-intron boundaries and the 5'upstream region containing promoter elements. The small size of the TIMP-1 gene facilitated a comprehensive approach to mutation screening by direct genomic sequencing. Primers designed for PCR amplification were also used to directly sequence the PCR products as described in section 2.8.3. Sequencing reactions were fractionated through 6% denaturing polyacrylamide gels and visualised by autoradiography (section 2.7.2).

Direct sequence analysis of affected males M2 and M3 from the two unrelated RP2 pedigrees XLRP 89 (Individual V-1, Figure 3.5) and XLRP 88/75 (Individual IV-2, Figure 3.13) respectively, revealed no disease specific sequence anomalies on comparison to unaffected male (M1) sequence. One common polymorphism was identified; a silent mutation in exon 5 at nucleotide position 434 (accession X003124). This converts TTC phenylalanine to TTT phenylalanine and may be a result of the common C-to-T conversion at CpG dinucleotide sites. This polymorphism is shown in Figure 6.1, which shows the reverse strand indicating a G-to-A conversion. However, no disease-specific mutations were identified, indicating it is unlikely that TIMP-1 is involved in the etiology of RP2.
TABLE 6.1 INTRON/EXON STRUCTURE AND SPLICE JUNCTIONS FOR THE HUMAN TIMP-1 GENE

<table>
<thead>
<tr>
<th>Exon</th>
<th>3' Splice Junction*</th>
<th>Exon Size</th>
<th>5' Splice Junction*</th>
<th>Intron Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>43 bp</td>
<td>GAGtaagcaggccccgggtgccccacagcagggagcag ggagcag ~1000 bp</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ccgggctgctgacatttgccacag AAC</td>
<td>129 bp</td>
<td>TCGtgctcttgatccccacactagccccacacacttgcc</td>
<td>~1600 bp</td>
</tr>
<tr>
<td>3</td>
<td>tgactcatatctctctctgca TCA</td>
<td>80 bp</td>
<td>AAGgtatccccggccctttttttccagcatttttaaa</td>
<td>188 bp</td>
</tr>
<tr>
<td>4</td>
<td>aagagacactttcccctcatccatcaacag ATG</td>
<td>127 bp</td>
<td>CTGtgaggccacgctccgegctgtgcccacacccaacagt</td>
<td>205 bp</td>
</tr>
<tr>
<td>5</td>
<td>gtcggccggccgccccctctctctccag GAA</td>
<td>125 bp</td>
<td>ACAgtgaggtgtgctgacgctccagccacaaagggagatcc</td>
<td>804 bp</td>
</tr>
<tr>
<td>6</td>
<td>ctcagagatgtttctctctctctccag GTG</td>
<td>267 bp</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Consensus) yyyyyyyyyyynagNNN NWGgtwgt

* Capital letters represent exonic sequence. Splice junction consensus is represented as follows: w, A or T; r, A or G; y, C or T; s, G or C (Mount SM. 1982). The gene consist of six exons and five introns spanning ~4.5kbp.
### TABLE 6.2 PRIMERS AND CONDITIONS FOR PCR AMPLIFICATION OF THE HUMAN TIMP-1 GENE

<table>
<thead>
<tr>
<th>Amplified region</th>
<th>Sequence of primer pairs</th>
<th>Size of PCR product</th>
<th>Annealing temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>5' Upstream 1</td>
<td>AACCGGTACCCATCTCAGAGA AGTGGAGGGAGATAAGGCAG</td>
<td>268 bp</td>
<td>62°C</td>
</tr>
<tr>
<td>5' Upstream 2</td>
<td>GACTCTGTGATCCTGGGTAAG AGTCTCCAGAATCTATCCAATCA</td>
<td>279 bp</td>
<td>62°C</td>
</tr>
<tr>
<td>5' Upstream 3</td>
<td>CACAATAAATGTGAAGGCTGA CTGAGCGCTAGAGGATAAATGT</td>
<td>383 bp</td>
<td>62°C</td>
</tr>
<tr>
<td>Exon 1</td>
<td>GATGAGTAATGCAATCCAGGAAG CGCAGGCCAAGCTGAGTGA</td>
<td>255 bp</td>
<td>62°C</td>
</tr>
<tr>
<td>Exon 2</td>
<td>GGTGGATGACCTGCCAGGAAG GGGCGAGTGGAATCATTCTAG</td>
<td>299 bp</td>
<td>62°C</td>
</tr>
<tr>
<td>Exon 3</td>
<td>GGCTCATGCGATCCATTTTGC TCAAAATGCTGGGAAAGAGGC</td>
<td>163 bp</td>
<td>62°C</td>
</tr>
<tr>
<td>Exon 4</td>
<td>TGTTGCGAGAAAGTGTTGGCTG GGACCTTGTTGTGGCAGCAG</td>
<td>210 bp</td>
<td>62°C</td>
</tr>
<tr>
<td>Exon 5</td>
<td>ATCACAGAAGCTGTTGCGGTC GGAAGATCTCCCCGTGGAG</td>
<td>209 bp</td>
<td>62°C</td>
</tr>
<tr>
<td>Exon 6</td>
<td>GCGCGGGCTCAGTGGTGAAAG AACCACATTGGGCACTCTCATC</td>
<td>459 bp</td>
<td>62°C</td>
</tr>
</tbody>
</table>

Primers for PCR amplification and sequencing of exons and upstream sequences containing promoter elements. Primers are listed as 5' to 3'. PCR and sequencing were performed as described in Materials and Methods.
Identification of a common polymorphism in TIMP-1 gene in an unrelated male patient with X-linked retinitis pigmentosa 2.

Direct nucleotide sequence analysis of PCR-amplified exon 5 DNA from control and unrelated affected males. Sequence of exon 5 was amplified using primer pairs as shown in Table 6.2. Sequencing reactions of patients and controls were loaded alongside each other according to the dideoxy-chain terminator used. Lanes 2 and 3 contain samples from RP2 affected males M2 and M3, respectively. Control DNA is derived from a healthy male M1 (lane 1).

A GT to AT conversion (indicated by arrow) was identified in affected male M2 on the reverse strand within the exonic 5 sequence. In the forward strand this results in the common CA to TA silent mutation at CpG dinucleotide, converting TTC phenylalanine to TTT phenylalanine and creating a polymorphism.
6.5 Discussion

The detection of naturally occurring mutations that disrupt genes are of use to biologists in many fields. In human genetics, these methods are used to determine whether a candidate gene is causally related to a phenotype and also to identify new alleles at known loci for diagnostic, population genetics and structure/function studies. A gene can be disrupted in many ways. Coding regions may be altered, promoter and other regulatory sequences may be disturbed, or processing signals such as splice sites may be modified. Possible alterations vary from deletions or re-arrangements that can remove an entire gene to individual nucleotide changes.

Single-base alterations are the most common type of mutation at most loci. A number of methods may be used for detecting these subtle changes such as single-stranded conformation (SSC) analysis (Orita et al., 1989), denaturing gradient gel electrophoresis (DGGE) (Myers et al., 1987 & Sheffield et al., 1989), heteroduplex analysis (HA) (Nagamine et al., 1989 & White et al., 1992), chemical mismatch cleavage (CMC) (Cotton et al., 1988 & Forrest et al., 1991) and direct sequencing (DS) (Bevan et al., 1992). All these methods are capable of detecting mutations with varying efficiencies (Grompe, 1993), but none defines precisely the nature of the change except direct sequencing. Since direct sequencing defines the location and nature of the change and therefore is a necessary final step of any mutation detection method, it was the preferred method for mutation screening of the human TIMP-1 gene in RP2 patients.

A majority of mutations, germline or somatic, are of little consequence to the organism since most of the genome appears to lack coding function. Even within protein-coding regions there is some tolerance to mutations due to the degeneracy of the genetic code and because the exchange of an amino acid may influence the function of a protein only slightly (Strong et al., 1991). With the development of increasingly efficient methods to scan large DNA segments for mutations, the need to predict the functional consequences of a mutation will become more pressing. Mutation screening methods such as direct genomic sequencing will detect nucleic acids alterations but do not necessarily define their biological significance. Some sequence changes, such as those causing frame-shifts or chain terminations, are obviously functionally deleterious to a protein. Other alterations, such as missense mutations leading to amino acid substitutions or base changes in introns and untranslated regions, can be either functionally relevant or represent sequence polymorphisms. Additional information including family studies, population studies or functional assays after in vitro expression may be needed to determine the significance of a sequence alteration.
Splice site mutations are a common cause of various diseases and result in abnormal RNA splicing and absent exon sequences. These mutations may be the result of single DNA base changes that disrupt 3' splice acceptor sites (Mitchell et al., 1986) and alter the processing and hence the structure of the mRNA such that an abnormal protein product is synthesised. Some examples include the Lesch-Nyhan syndrome (Gibbs et al., 1989), X-linked chronic granulomatous disease (De Boer et al., 1992) and the cystic fibrosis transmembrane regulator gene (Tsui, 1992). Another group of mutations are those that cause a reduction or absence of a particular protein product. This class of disorders may result from mutations that affect transcription for example, promoter box mutations in β thalassaemia, (Orkin, 1987) or processing of mRNA, for example cryptic splice sites in introns in β thalassaemia, (Thein and Weatherall, 1988) or that act at the translational level as in α thalassaemia (Higgs et al., 1989). Mutations found in the promoter region have also been found to contribute to the pathology of diseases such as Chronic Granulomatous Disease (Newburger et al., 1994) and Charcot-Marie-Tooth neuropathy (Ionasescu et al., 1994). Since deletion of exons (not detectable by direct sequencing in autosomal dominant diseases) and point mutations in exons and splice sites are the most common causes leading to disease phenotype, the mutation screening of the TIMP-1 gene included the coding regions of exons 1 to 6, adjacent splice sites and 5' upstream regions (Table 6.2, section 6.4), but not the non-coding intronic sequences.

A comparison of the deduced amino acid sequence of TIMP-1 (Gasson et al., 1985), TIMP-2 (Boone et al., 1990) and TIMP-3 (Apte et al., 1994 & Silbiger et al., 1994) demonstrated that 12 cysteine residues are highly conserved at the same relative locations within the mature TIMP proteins. These cysteines have been implicated in correct protein folding by forming six intra-chain disulphide bonds (Williamson et al., 1990). From these conformational considerations it seems likely that all members of the TIMP family have a very similar organisation in their tertiary structure. It also suggests that correct disulphide bond formation is essential for the functional properties of the TIMP protein. During the screening of TIMP-1 in RP2 patients no point mutations were found to affect these cysteine residues, and although no disease-causative mutations were detected, it is possible that such mutations may exist in controlling elements outside the regions studied.

The precise expression pattern of TIMP-1 in the nuclear layers of the retina was not pursued in this study because no disease-causative mutations were identified. However, other reports have shown that TIMP-1 is secreted by cultured retinal pigment epithelium cells, (Alexander et al., 1990) and immunohistochemistry has demonstrated expression in primate retinal cell bodies (Johnson et al., 1995). As the activity of this gene product must be tightly regulated after translation and at the level of gene expression, it is perhaps no surprise that little transcript could be detected in the retinal cDNA libraries. Potentially, a gene defect may lead to disruption of photoreceptor-matrix
interactions and induce apoptosis (Frisch and Francis, 1994), although the role of TIMP-1 in maintaining retina structure and function has not been investigated. The role of this tightly regulated protein in the normal functioning of the retina has yet to be determined.

### 6.6 Conclusion

For both RP2 families studied here, mutations in the TIMP-1 coding sequence, splice sites and the 5' upstream region as a cause of retinal degeneration in X-linked retinitis pigmentosa 2 can be excluded. However, mutations in intronic regions, promoter/enhancer mutations and an as yet unidentified regulatory element that lies outside these intervals in the TIMP1 gene may be implicated. Hence this gene may be causative for disease in other RP2 families which have not been investigated yet, as for example where mutations have not always been found in all genetically determined RP3 families for the RPGR gene (Zito et al., 1999; Buraczynska et al., 1997). Although the silent mutation identified in the affected male M2 within the exonic 5 sequence did not change the amino acid phenylalanine it could potentially contribute to disease since some apparently innocuous synonymous mutations in coding DNA may not be neutral as expected because they activate a cryptic site and can be pathogenic as was reported for the calpain gene (CANP3) (Richard I and Beckmann J.S., 1995) responsible for the aetiology of recessive chromosome 15-linked limb girdle muscular dystrophy. Until one assesses their impact on mRNA processing and stability or even codon usage bias, it may be difficult to ascertain the neutrality or non-pathogenicity of apparent polymorphisms.
CHAPTER SEVEN

GENERAL DISCUSSION
CHAPTER SEVEN
GENERAL DISCUSSION

7.1 Overview of the work described in this study

At the outset of this project, RP2 remained broadly localised to an ~13cM region of proximal Xp flanked by markers DXS7 (Xp11.3) and DXS255 (Xp11.22) (section 1.8.7.3.2), owing to a lack of informative recombination events and a paucity of suitably placed polymorphic markers. Between 8-19 polymorphic microsatellite markers and other microsatellites mapped to the region from whole-genome mapping efforts were used in this study to generate extensive haplotypes for members of twelve XLRP families in an effort to determine the disease locus segregating within each family and to further define the localisation of the RP2 gene.

Key recombination events enabled classification of two families as RP2 (XLRP 89 and XLRP 88/75) and enabled refinement of the distal boundary, localising the RP2 gene proximal to MAOA (Xp11.3). This confirmed previous localisation data for the RP2 locus to between markers DXS7 and DXS426 (Ott et al., 1990). The proximal boundary could not be refined further in these RP2 families due to a lack of crossovers at marker DXS255 and distal to it at DXS573. Hence in this study the XLRP2 critical region could not be further refined from this 13cM interval. However further investigation of XLRP2 families in this laboratory (Thiselton et al., 1996) helped to substantially refine the RP2 interval to approximately 4-5cM in a region flanked by markers DXS8083 and DXS6616.

Two families were classified as segregating the RP3 disease based on recombination events. A further family RP93 was clearly found not to be segregating the RP2 disease locus and was subsequently defined as RP3 independently by discovery of a mutation in the RPGR gene (section 3.3.8). Further genetic studies are necessary to establish which XLRP locus is segregating in the remaining 7 families. This may add to the number of families classified as RP6/RP15 or substantiate reports of still further lod for this disorder located on the X chromosome.

In parallel with the genetic studies, the technique of cDNA hybridisation selection was applied to four overlapping YAC clones (500kb) (section 4.1.6) mapping to the Xp11.23 region, being part of the XLRP2 critical locus, in order to develop a transcriptional map of the XLRP2 disease gene region. As part of the analysis of the Xp11.23 region, the direct cDNA selection procedure was successfully established as indicated qualitatively by dot blot and Southern blot analysis (section 4.2.6).
Comparisons of the selected material and starting cDNAs demonstrated a dramatic level of enrichment for the rare retinal internal control gene TIMP-1 after the second round of selection in the secondary selected cDNAs. However the more abundant retinal gene rhodopsin was normalised in the λgt10 selected cDNAs (section 4.2.6.4). This hybridisation analysis was also effective in qualitatively monitoring the level of depletion of the contaminant transcripts such as β-actin, yeast and ribosomal cDNAs in the enriched selected cDNAs.

Quantitative analysis of selected cDNAs by screening selected sub-libraries showed effective isolation and normalisation of more highly represented genes such as rhodopsin and a 1000-fold enrichment for the rare TIMP-1 gene (section 5.4.2). A total of 18 distinct cDNAs (section 5.6) were identified on the basis of hybridisation, sequence analyses and comparisons of selected cDNAs with cDNAs and genes described in the GenBank database. Mapping studies on twelve of these selected cDNAs (section 5.7) using PCR analysis indicated that none of these cDNAs mapped back to their YAC of origin. However screening more cDNA fragments per 100kb of YAC DNA could identify more rare transcripts and further analysis of the cloned enriched sub-libraries with systematic mapping of directly selected cDNA clones may contribute to a transcription map of this region. Furthermore, once the disease gene is cloned, the development of a detailed transcription map of the XLRP2 critical region will allow further assessment of the possible regulatory relationships between genes in this region.

The goal of the direct selection experimental design is to rapidly identify most of the coding sequences in large DNA fragments. A limitation of this approach is that it requires a cDNA library representing all possible mRNA from a species to be able to identify all coding sequences. In this study, retinal cDNA libraries were used as it was thought that the XLRP2 gene would be expressed in the retina. Another obvious limitation of direct selection is the need for a source of cDNAs that includes the specific cDNA being targeted. If the corresponding gene is expressed only transiently, this is a major problem. However, low levels of expression can be overcome by using cDNA pools, which allow for the efficient selection of cDNAs present at <1 X 10^-6. There are several advantages of this approach over the other approaches in identifying of coding sequences. The cDNA selection procedure is insensitive to number and size of introns, cryptic splice sites, or signals in intergenic DNA that resemble splice sites unlike the exon trapping procedure. This method can also detect 5' and 3' exons and genes that lack CpG-rich islands and introns or those genes that are not transcribed in hybrid cell lines or have highly diverged between species. The cDNA fragments that are obtained are expected to be randomly distributed over the mRNA and, therefore would link exons. The approach is also more robust than screening with labelled probes from large DNA fragments and is clearly less sensitive to the size of fragments used for selection. Selection is particularly convenient for YACs where a single gel and blot can be used to
identify cDNAs corresponding to a dozen or more YACs within a relatively short period of time. Most cDNA libraries consist of $10^6$ recombinant clones, an adequate representation of the transcripts expressed in one cell type. However, if the targeted transcript is expressed only rarely or at a low level in a complex tissue, then screening such a library may be entirely inadequate. These problems are exacerbated when commercial or amplified cDNA libraries are used, many of which have skewed sequence representation and a preponderance of artefactual clones. Direct selection can also be affected by factors associated with nucleic acid hybridisations. One problem that has been encountered in this context is where certain genes that have short exons fail to hybridise to their genomic locus efficiently (Lovett, 1994).

Since 1992, many groups have applied direct selection methods to various positional cloning projects and to intensive studies of defined genomic intervals. Now, efforts are focused on improving the latter stages of the analysis, in particular, on developing faster methods for deriving longer, near full-length, cDNAs, rather than a jigsaw puzzle of shorter overlapping fragments (selected cDNAs generally range from a few hundred basepairs to a few kilobases, averaging 0.5-1 kb). One area which could be improved, involves the use of mixed cDNA pools in a true multiplex selection. It would be useful to be able to discriminate the source of any given selected cDNA. This could be accomplished by adding multiplex DNA linkers to each cDNA source before they are mixed, so that the outer priming site is common to all cDNA sets while the inner one is specific to each cDNA source. By using the outer sites to prime amplification reactions during selection and inner sites to identify the selected cDNAs, one could immediately ascertain from which pool, and thus which tissue or developmental stage, the original transcript was isolated from.

Since none of the methods used for the construction of a regional transcription map allows the identification of all the genes residing in the region, the use of a combination of different methods is needed to make the map as complete as possible. Though the cDNA selection procedure as applied here cannot be expected to identify more than 50% of the genes of an organism, it does take its place as one of the developing set of efficient protocols, likely to speed up the identification of the genes of an organism with a reasonable amount of effort. Alternatively the enrichment protocol described here can be combined with the results of one of the other techniques (use of sequences chosen by the selection for conserved sequences or exons to specific cDNAs) to increase the purity of the selected material, and to reduce the background.

In addition to the photoreceptor-specific components (section 1.6.6), an increasing body of evidence is accumulating to implicate disturbances in protein processing in retinal degeneration. For example, both the putative RP3 gene (RPGR), and the choroideremia gene (a Rab escort protein) encode proteins involved in the regulation of GTPase activity and hence cellular signalling (section 1.8.7.3.1). TIMP-3, the gene
responsible for Sorsby's fundus dystrophy, encodes a tissue inhibitor of
metalloproteinases which is important in the processing of extracellular matrix proteins
(section 6.1). Although these genes are expressed in many tissues, mutations appear to
result solely in an ocular phenotype, which suggests that photoreceptors, with their high
levels of protein synthesis and trafficking, are particularly sensitive to disturbances in
protein turnover and processing. Several genes located within the RP2 critical region are
strong 'positional candidates' in this regard. The TIMP-1 gene was therefore screened
during this study for mutations in affected males of two RP2 families by PCR and direct
sequencing of all coding sequence, splice junctions and 5'UTR promoter region.
Although a neutral polymorphism was discovered in exon 5, no disease-associated
sequence alterations were found (section 6.4), suggesting that TIMP-1 is unlikely to play
a causal role in the etiology of XLRP.

Another strong positional candidate gene found in the RP2 critical region is the
ubiquitin C-terminal hydrolase (UHX1), which has a high level of retinal expression and
similarity to the Drosophila gene fat facets, the gene product which influences the fate of
cells destined to become photoreceptors in the developing compound eye (Fischer-Vize
et al., 1992). Given the importance of correct protein processing to retinal health and the
role of the ubiquitin system in protein modification and targetting (Deshaies, 1995;
Swanson et al., 1996), the genes UHX1 and PCTK1 were recently evaluated as candidate
genes for retinal diseases mapped to Xp21.1-p11.2 (Brandau et al., 1998). Fine mapping
techniques localised UHX1 to between markers DXS1266 and DXS337, where the CSN1
and RP2 loci are located. Evaluation of both CSN1 and XLRP2 patients revealed no
UHX1 or PCTK1 gene mutations or aberrations, using SSCP or deletion screening with
cDNA hybridisation hence excluding these as causal genes for XLRP2 and CSN1 eye
diseases (Brandau et al., 1998).

7.2 Apoptosis and potential therapeutic roles for RP disease

Whilst clinical heterogeneity is observed in RP this variability is superimposed
on a basic set of common signs and symptoms, such as night blindness, ERG
abnormalities, arteriolar narrowing and bone-spicule pigmentation. It is therefore
puzzling that such a high degree of genetic heterogeneity has been observed for RP, as
mutations in several different genes such as rhodopsin, the β-subunit of cGMP
phosphodiesterase, and the RPGR gene all result in certain phenotypic similarities,
which can be clinically categorised as RP. Apoptosis, a form of programmed cell death
(section 1.6.5.1) has been implicated as a final common pathway of the disease process,
which may link the different genotypes observed in RP to a common clinical phenotype.
In animal models of many types of retinal dystrophy, the death of photoreceptors seems
to occur by this common mechanism (section 1.6.5.2). Apoptosis has also been found to be a major mechanism of cell death in light-induced photoreceptor degeneration (Hafezi et al., 1997). The molecular basis for programmed cell death is rapidly becoming understood and several ways to inhibit it have been investigated.

Manipulation of factors which influence the apoptotic pathway could be a potential avenue to follow for ameliorating disease. Withdrawal of growth factors has been shown to induce apoptosis in neurons (Martin, 1993). Furthermore, intravitreal injection of various growth factors into the RCS rat and in a light damaged rat model delayed photoreceptor cell death for several months (LaVail et al., 1992). One explanation for this rescue could be that injection of growth factors prevents induction of the apoptotic pathway, prolonging the survival of photoreceptors. However, before such an approach can be justified in humans it would be necessary to establish the functional capacity of rescued photoreceptors, the longevity of effect and efficacy of different growth factors. Recently anti-apoptosis therapy has also been studied in two Drosophila mutant strains rdgC and ninaE^{Rh27}/+, that exhibit age-related retinal degeneration, by blocking the activity of a family of proteases that mediate programmed cell death, the caspases (Davidson and Stellar, 1998). Such caspases were blocked by making use of a broad-spectrum caspase-inhibitory baculoviral survival protein p35 (Hay et al., 1994). Mutant flies expressing p35 showed significant retention of visual function and may provide a pointer for the effort to halt the progression of rhodopsin-mediated forms of RP.

Another avenue for therapy would be the genetic manipulation of apoptosis-associated genes. In the past few years several investigators have attempted to delay retinal degeneration by modifying the expression of genes known to be involved in apoptosis. However, overexpression of bcl-2 and bcl-X, potent inhibitors of apoptosis in many systems, did not result in a significant reduction of photoreceptor apoptosis in the rd mouse (Joseph and Li, 1996). Similarly, absence of the tumor-suppressor gene product p53 did not alter apoptosis of photoreceptors in a mouse model of light-induced retinal degeneration which lacks p53 (Marti et al., 1998). Another gene product involved in the regulation of apoptosis is the proto-oncoprotein c-fos, a component of the transcription factor activator protein 1 (AP-1). Recently, it was shown that mice lacking c-fos are protected from light-induced photoreceptor cell death and c-fos was also continuously upregulated concomitant with apoptotic photoreceptor death (Hafezi et al., 1997a). By studying double mutant rd/rd, c-fos^{−/−} mice it was shown that in contrast to its role in light-induced photoreceptor degeneration, c-fos is not essential for apoptosis in the rd mouse and hence not involved in the apoptotic pathway of inherited retinal degeneration (Hafezi et al., 1998). Further studies are needed to elucidate the mechanisms of c-fos and AP-1 action in the retina and their involvement in the regulation of apoptotic photoreceptor cell death. Other studies on the mechanism of
inhibition of the apoptosis inducing genes could also be investigated including the use of antisense oligonucleotides to block translation of apoptotic mRNAs (Martin et al., 1994). An important issue which needs to be addressed with such strategies, is that it is critical to limit the scope of manipulation to the target tissue of interest.

Manipulation of the apoptotic mechanism could have significant implications for the design of therapeutic strategies for human retinal dystrophies, for which no treatment is currently available. Schemes for gene therapy (discussed in the next section) for individual mutations in RP disease could be an overwhelming task, whereas strategies aimed at manipulating apoptosis might be much more practical, as it would apply to multiple different mutations affecting different genes. The vertebrate visual system has been pushed, by evolutionary pressures, to approach its thermodynamic limitations. This exquisite tuning, however, comes at a cost. Photoreceptors are forced to carry large metabolic burdens, to renew continuously their light-sensitive outer segments and to exist under conditions of very high oxygen partial pressure (reviewed in Steinberg, 1987). Modest changes in the biochemical environment or cellular structure of photoreceptors, caused by genetic lesions or environmental insults, are sufficient to induce cell death through apoptosis. For the inherited retinal degenerations, understanding the impact of each molecular defect on the biology of photoreceptors is important, in addition to studying how these effects converge upon the apoptotic pathway. Collectively, these studies should lead to the development of rational therapies that may slow or even reverse the progression of this devastating disease process.

7.3 Potential gene therapy for the treatment of RP

The ultimate goal of providing suitable therapy which will either cure or slow the disease progression is the aim of investigating and elucidating underlying disease pathologies in human disorders. At present, there are no long-term therapies for retinal degenerative diseases. An encouraging exception is Sorby fundus dystrophy (SFD). In a report by Jacobsen et al., (1995), SFD patients with mutations in the TIMP3 gene responded favourably to high doses of vitamin A, such that photoreceptor function improved, at least briefly when sufficient levels of vitamin A were restored. Whilst this finding represents a starting point for metabolic intervention in this form of macular degeneration, further work must be carried out to determine the lowest effective dose of vitamin A, as the long term use of large doses may have toxic side effects. Although this specific therapy is not necessarily applicable to other forms of retinal degeneration, as additional causative genes are characterised it should be possible to design specialised therapies based on the knowledge of the biochemical defect.
The identification of a disease-causative gene opens the possibility to consider human gene therapy and to move beyond genetic counselling and non-specific management of many disorders. Inherited diseases of the retina are excellent candidates for gene therapy; the direct transfer of therapeutic genes to the relevant tissue. Direct access to the back of the retina, and the sub-retinal space, is possible while maintaining the spatial integrity of the retina (Ali et al., 1996). Thus therapeutic agents can be placed in contact with the target tissue while other tissues are spared. Also, excellent animal models exist (section 1.6.6), including the rd5 mouse, the Irish setter dog with mutations in PDEB (the canine homologue of the β subunit of PDE) (Suber et al., 1993) and several transgenic animals expressing human mutations. Furthermore, because the gene defect can be identified early in the course of disease, while most of the retina is still intact, therapy to prevent degeneration is feasible. The principal difficulty with gene therapy for these diseases is that cells in the mature retina do not divide, therefore use of viral vectors that integrate into replicating DNA is precluded. Conversely, non-integrated vectors might be more stable in retinal cells than in rapidly dividing cells. The use of adenovirus or adeno-associated virus which can infect non-dividing cells should get around this problem. Finally, the efficacy of gene therapy will depend, in part, on whether the disease mutation has a negative effect on the retina, which would simply require addition of the correct DNA sequence, or whether the cause is a positive-acting mutation, which would require replacement of the defective DNA. The latter possibility being a more challenging prospect. Since the RP2 gene is not retinally specific in its expression and the mutation of the gene potentially causes a defect in the cytoskeleton (Schwahn et al., 1998) and is not of a biochemical nature, then gene therapeutic strategies may not compensate even partially for the defect in this gene.

Several attempts at gene therapy for retinal diseases in animal models have been reported. In the rd mouse photoreceptor cell degeneration was found to be significantly retarded after adenovirus-mediated delivery of a wild-type β-PDE cDNA to the photoreceptors (Bennett et al., 1996). Although these initial steps towards somatic gene therapy are encouraging, much will also depend on the stability of the transduced genes, and the targeting of the transgene to the relevant cells. To overcome such limitations as short-term expression of the delivered gene and loss of transduced cells in adenovirus use in vivo, an Ad-based vector, the encapsidated adenovirus mini-chromosome (EAM) from which all of the viral genes have been deleted was developed and used successfully to transduce cells in vitro (Kumar-Singh and Chamberlain, 1996). In a recent report (Kumar-Singh and Farber, 1998) this novel EAM vector system was used to transduce and rescue cells from the neurosensory retina in vivo. EAM-mediated delivery of the β-PDE cDNA to rd mice allowed prolonged transgene expression and rescue of rod photoreceptor cells. There are a number of other physicochemical and viral vectors that are promising vehicles for gene delivery. In a recent report, ribozymes, catalytic RNA
molecules that cleave a complementary mRNA sequence, were used in the rescue of photoreceptor cells in a transgenic rat model of ADRP (Lewin et al., 1998). A combination of strategies that enhance and prolong the expression of genes delivered to the retina as well as selection of the optimal gene delivery vehicle could ultimately lead to effective gene therapy for some types of retinitis pigmentosa.

7.4 Recent discovery of the RP2 gene and its implications

The positional candidate approach favours mutation screening of genes that are selectively expressed in the tissue most profoundly affected by the disease and also those most abundantly expressed in such tissues. There are, however, a number of genes that are ubiquitously expressed but where mutation results in damage to one or few tissues for example, the retinoblastoma gene. A more reasonable candidate selection criterion might be the expression of a gene that is expressed in or near the affected cells. If, however, the gene is transiently expressed during development and/or is never expressed at a high level, it could be absent from all but the most complete cDNA library catalogues and undetectable by northern blotting. While there is no doubt about the power of the positional candidate approach, there are some genes whose discovery will depend on more purely positional methods, as has been the case for both the RP3 and RP2 genes.

Schwahn et al., (1998) used PCR to create a hybridisation probe that consisted of a random assortment of sequences from a yeast artificial chromosome that mapped within the genetically determined RP2 disease interval. This probe was then used to screen Southern blots of genomic DNA from 26 unrelated probands affected with X-linked RP2 and discovered one with a substantive re-arrangement. The first RP2 mutation discovered was caused by the insertion of a type of retrotransposon known as a long interspersed nuclear element (LINE 1 or L1, Kazazian and Moran, 1998) into intron 1 of the gene. This mutation was of sufficient size to be detected by the hybridisation strategy used by the investigators but would not have been detected by typical PCR-based mutation screening of the coding sequence. This could account for why more mutations have not been detected in the RPGR gene (Buraczynska et al., 1997) or in the recently discovered ABCR gene involved in Stargardt disease (Allikmets et al., 1997), considering the possibility that genomic re-arrangement mutations may occur more frequently than is commonly supposed.

The RP2 gene is homologous to cofactor C, a protein which mediates the folding of β-tubulin. The photoreceptor outer segments are connected to the inner segments by a thin cytoplasmic bridge containing a non-motile cilium. This structure contains nine pairs of microtubules, and it is plausible that a mild error in microtubule assembly or
function could be ultimately fatal to the photoreceptor cells while allowing such ubiquitous microtubule-associated functions as cell division to proceed normally. This could be the first case where a defect in the cytoskeleton has been shown to cause RP, although the functional relationship to cofactor C may be so distant (it has a similarity of 44% over less than half of the amino acid sequence) that RP2 may have nothing to do with microtubule assembly. The precise folding pathways of α- and γ-tubulins (both expressed in the retina), and associated cofactors have not been determined. Thus the pathogenesis of RP2 could involve the accumulation of incorrectly-folded photoreceptor or neuron-specific tubulin isoforms, followed by progressive cell death. If so, other tubulin-folding defects may give rise to similar phenotypes, and these findings may then aid in identification of the genes responsible for other forms of retinal degeneration.

The RP2 gene can be further characterised by a number of approaches. In vitro expression systems may be of use in order to determine the function of the normal protein and to study potential interactions with other proteins. Since no known mouse model exists for XLRP2 a suitable application in the future may be to use gene targeting for gene replacement or knock out studies since gene function is not always inferred from primary sequence database identities. A mouse model for this form of XLRP could be created by the construction of transgenic mice carrying the respective disease-causative mutations identified in individual families. This technology allows the evaluation of the effect on the organism of absent gene function or presence of dysfunctional product and a more detailed study of the histology and pathology of the disease, which is difficult in living patients. In addition, such a model would provide a tool for testing potential drug or environmental therapies, and could be used in experiments on gene therapy based on antisense RNA. It would also be useful to screen for mutations of the RP2 gene in a panel of patients with various retinal dystrophies, such as CSNB and AIED (section 1.8) which may have allelic heterogeneity to the RP2 gene, as was found for a RPGR mutation in a CSNBX family (Herrmann et al., 1996). Different mutations in this gene may account for different clinical phenotypes, as has been found for the rhodopsin and peripherin/RDS genes. Mutation screening of other XLRP2 families to determine additional new mutations in the gene could also enable rapid and accurate identification of carriers which in turn would permit more accurate genetic counselling and prognosis. The application of conventional biochemistry and cell biology approaches, combined with the accrual of additional RP2 mutations, mean that despite its extremely low expression in the retina, the true function of RP2 is unlikely to remain in the dark. This in turn should lead to a better understanding of normal eye function, provide clues to assist in the discovery of other RP genes and contribute to the search for a cure for RP.
It is interesting that a number of preconceived ideas about what type of genes would be involved in retinal dystrophies have now been abandoned. Initially it was thought that retina-specific diseases would be caused by mutation of retina-specific genes. The association of Sorsby fundus dystrophy with the ubiquitously expressed gene TIMP3 (Weber et al., 1994) has disproved this. RPGR mutations in X-linked RP3 (section 1.8.7.3.1) and RP2 gene mutations in X-linked RP2 (section 1.8.7.3.2) suggests that severe retinopathy need not relate to mutations of genes expressed at high levels in the retina. Histopathological studies originally led to the hypothesis that the Stargardt macular dystrophy gene would be expressed in the RPE. Subsequently, the relevant mutant gene was found to be expressed in rod photoreceptors (Allikmets et al., 1997). This is despite the fact that Stargardt macular dystrophy is, clinically, mainly a disease of cone photoreceptors. Future research efforts therefore need to be broad based, making as few assumptions as possible.

7.5 Future work

The construction of extensive transcript maps are necessary for exploring the functional significance of chromosomal organisation and for isolating disease genes. The direct isolation of expressed sequences from ordered genomic libraries by exon amplification or direct cDNA selection is an appropriate strategy for integrating physical and transcriptional data and has been used in positional cloning approaches and for megabase-scale isolation of cDNAs (The Huntingdon's Disease Collaborative Research Group. 1993).

A catalogue of genes encoded by the 500kbp region of Xp11.23 would provide a framework for assigning roles in the etiology of XLRP2 and other retinal diseases in the region to individual gene/s. The selected cDNA fragments were initially analysed by limited DNA sequence analysis (sections 5.5.2 & 5.6), in order to design suitable oligonucleotide primer for mapping studies (section 5.7) to the YACs used in the selection procedure. Screening of the arrayed cDNA libraries (section 5.4) containing enriched cDNA clones can prove to be a rapid method to identify an initial set of novel sequences and to survey gene content. This approach however may identify only a subset of all genes present in the libraries. When an arrayed library had been essentially completely analysed in a recent study (Tassone et al., 1995), re-screening of the entire selected library (at >1000 pfu per 100 kb of YAC DNA) with all novel sequences revealed significant numbers of additional novel sequences. Often these were present at < 1% of the library and appeared to represent new genes. Thus, for comprehensive transcriptional mapping, exhaustive analysis of the selected cDNA libraries, involving
screening of several thousand clones, is required. Hence screening of the selected arrayed cDNA libraries containing ~200 novel cDNA clones of which some are overlapping will serve well in a survey analysis of gene content and may possibly identify classes of high and low frequency genes. Comprehensive transcriptional mapping is only possible after initial product analysis since comprehensive analysis of a selected cDNA sub-library practically involves two steps—identification of a set of highly represented cDNA fragments by analysis of arrays of a few hundred clones, followed by re-screening of 5,000-10,000 clones with the high frequency fragments to assure identification of clones present at levels of ~0.1%. The combined analysis and mapping studies will be used to estimate the completeness of this mapping effort and to identify procedures that would facilitate large-scale transcript mapping.

Future work on the analysis of clones can measure redundancy of transcription unit representation. A measure of how completely the transcription units in the region were identified is how often the same transcript was recognised by different cDNA fragments. Most of the redundancy can be eliminated during the preliminary screening by ignoring cDNA fragments that cross-hybridise with mapped cDNA fragments. Large transcripts would, however, still be represented by several of the small (300-500 bp) cDNA fragments (section 5.5.1). It can be reasoned that the isolation of larger cDNA clones, more closely approximating the full-length transcripts, would enable one to identify short probes that recognise the same transcript. The description of the transcriptional map of this region would be made much more complete by the addition of a predicted protein product obtained for any mapped selected cDNAs, and the 5' and 3' boundaries of such transcription units. Most methods, including cDNA direct selection, for producing transcript maps do not yield this type of information. Sequence analysis of full-length cDNA clones would provide information necessary to predict the protein product encoded by each transcription unit. Such sequence information could also be used to design oligonucleotides for mapping of the 5' and 3' boundaries of each transcription unit. However, full length cDNA clones would not be easy to obtain for long or low abundance transcripts. In the case of long transcripts, the full length sequence could be obtained by putting together the sequences of overlapping cDNA clones. The isolation of overlapping clones would be aided by the fact that multiple short cDNA fragments from each transcription unit are isolated. The protein product of low abundance transcripts would be most easily predicted by sequencing of genomic DNA.

The YAC and possibly a cosmid-based physical map of the XLRP2 critical region should allow the transcript fragments to be accurately mapped, which in turn will allow the density of transcription through the region to be evaluated. With this information, one should then be able to address issues such as how completely the transcription units have been identified, what the factors are that limit the effectiveness of the procedure, how the description of the transcriptional map of this region can be completed, and what
are the uses of this map for studies of other retinal disorders in this region of chromosome Xp. In addition to their application in the search for disease genes, the transcript maps will provide an important insight into the organisation of the human genome, the distribution of genes, their order and its possible functional relevance. Since transcripts usually are evolutionarily conserved, they will also serve as anchor points in comparative mapping efforts for example, in selecting mouse and human cDNAs across the XLRP2 critical genomic region. The availability of cDNAs from mouse embryos at various stages of development makes this a particularly attractive approach for mapping genes that are developmentally regulated, for example those involved in vertebrate photoreceptor development. Although direct selection of genomic DNA fragments from one species using target genomic DNA from another species is technically difficult, it has already been achieved (Sedlacek et al., 1993a), and will doubtless prove valuable in producing comparative gene maps.

Further characterisation of interesting cDNA clones from the selected libraries analysed by sequence analysis will involve mapping back interesting clones to the original YAC to confirm region specific chromosomal localisation. Somatic cell hybrid mapping panels can also be utilised to regionally localise new genes and to exclude non-specific isolates. Another logical step in the high throughput and detailed mapping of XLRP2 critical region specific cDNAs, will be the use of X-chromosome specific radiation-reduced hybrids. Genomic blots of restriction enzyme digested YAC DNA will be used to detect single copy gene fragments and such restriction mapping also used to determine the presence of overlapping gene fragments. The original cDNA libraries can also be screened with small PCR probes in order to isolate full length cDNA sequences and to link partial overlapping clones. With a more complete sequence, translation of cDNA clone sequences for open reading frames (ORF) using BLASTN database searches, GRAIL and protein database searching to determine possible function/s for the gene product can be undertaken. Zooblots can also be used to detect conservation of gene sequences in other animal species.

Expression studies using Northern blots, RT-PCR and tissue in-situ hybridisation experiments will be used to verify that selection products represent genuine genes and to study their expression pattern in retinal tissue. Northern blotting analyses will be conducted on selected clones to assess the distribution of expression in various adult human tissues and to determine transcript sizes. The failure of cDNA fragments to detect transcripts could be a reflection of their low levels of transcription, their tissue specificity, or the low sensitivity of probes especially those in a size range of 200-250 bp. RT-PCR should provide increased sensitivity in examining sequences with low levels of expression. Negative RT-PCR results with adult retinal mRNA as template may indicate that the corresponding mRNAs are expressed at very low levels or are regulated
developmentally. Clones that may give positive results with RT-PCR but which do not detect bands on Northern blotting may represent transcripts of low-copy number.

As reported by Schwahn et al., (1998) only 18% of XLRP2 patients screened revealed disease-associated mutations in the reported RP2 gene (section 1.8.7.3.2). Perhaps mutations may lie in unidentified parts of the gene or alternative transcripts, or common mutations were not identified by SSCP analysis and/or there is heterogeneity within RP2, with the major locus still to be identified. Hence any potentially interesting genes isolated will then be screened for mutations in RP2 patients and such mutation screening also applied to DNAs of patients with retinal disorders such as AIED and CSNBX. Unravelling the genetic basis of XLRP2 will require a detailed characterisation of the entire critical region, including comprehensive analysis of any candidate genes selected. Furthermore, once any retinal disease gene/s is cloned, the development of a detailed transcription map of the XLRP2 critical region will allow further assessment of the possible regulatory relationships between genes in this region.
CHAPTER SEVEN A

SUMMARISED GENERAL DISCUSSION

7.1A Overview of the work described in this study

Extensive haplotypes were generated using polymorphic microsatellite markers for members of twelve XLRP families in an effort to determine the disease locus segregating within each family and to further define the localisation of the RP2 gene. Key recombination events enabled classification of two families as RP2 (XLRP 89 and XLRP 88/75) and enabled refinement of the distal boundary, localising the RP2 gene proximal to MAOA (Xp11.3). This confirmed previous localisation data for the RP2 locus between markers DXS7 and DXS426 (Ott et al., 1990). The proximal boundary could not be refined further in these RP2 families due to a lack of crossovers at marker DXS255 and distal to it at DXS573. Hence in this study the XLRP2 critical region could not be further refined from this 13cM interval. After this analysis four of the XLRP families (XLRP87, XLRP90, XLRP93 and XLRP1120) were reported to have new and different mutations for the RPGR gene (Zito et al., 1999; section 3.3.A). Linkage analysis of the remaining eight XLRP families using markers spanning Xp21.1 to Xp11.23 containing the RP2 and RP3 loci indicated no statistically significant lod scores (section 3.3.A1) due to the small size and lack of informative meiosis of many of these families. There was also found to be no evidence for a founder effect following haplotype comparisons in these eight families (section 3.3.A2) due to a lack of haplotype sharing.

As part of the transcriptional analysis of the Xp11.23 region, the direct cDNA selection procedure was successfully established as indicated qualitatively by dot blot and Southern blot analysis (section 4.2.6) and quantitatively by screening selected sub-libraries which showed effective isolation and normalisation of more highly represented genes such as rhodopsin and a 1000-fold enrichment for the rare TIMP-1 gene (section 5.4.2) in the secondary selected cDNAs. A total of 18 distinct cDNAs (section 5.6) were identified on the basis of hybridisation, sequence analyses and comparisons of selected cDNAs with cDNAs and genes described in the GenBank database. Mapping studies on twelve of these selected cDNAs (section 5.7) using PCR analysis indicated that none of these cDNAs mapped back to their YAC of origin. However screening more cDNA fragments per 100kb of YAC DNA could identify more rare transcripts and further analysis of the cloned enriched sub-libraries with systematic mapping of directly selected cDNA clones may contribute to a transcriptional map of this region.
7.2A Future work

Further screening and analysis of a few hundred clones of the selected arrayed cDNA libraries for novel cDNAs and genes would contribute to a comprehensive transcriptional map of the 500kbp region of Xp11.23 analysed. Further characterisation of any novel and interesting cDNA clones will involve mapping studies to the original YACs, somatic cell hybrid mapping panels and X-chromosome specific radiation reduced hybrids to confirm region specific chromosomal localisation. The original cDNA libraries can also be screened with small PCR probes in order to isolate full length cDNA sequences and to link partial overlapping clones. With a more complete sequence, translation of cDNA clone sequences for open reading frames (ORF) using BLASTN database searches, GRAIL and protein database searching to determine possible function/s for the gene product can be undertaken. Expression studies using Northern blots, RT-PCR and tissue in-situ hybridisation experiments will be used to verify that selection products represent genuine genes and to study their expression pattern in retinal tissue.

The vast amount of data generated by the Human Genome Project accessed via various databases will allow the query of future sequence data in the context of different analyses, including gene finding, protein motif identification, regulatory motif analysis, identification of repeated sequences, similarity analyses, nucleotide compositional analyses and cross-species comparisons. The use of bioinformatics in the form of databases and analytical tools will become essential for collection, analysis, annotation and storage of mapping, sequencing and expression data. Also computational methods will allow extracting, viewing, annotating and analysing genomic information efficiently. The availability of entire genome sequences has enabled a new approach to biology called functional genomics, the interpretation of the function of DNA sequence on a genomic scale, which results from the interaction of genomes with their environment. Current methods for studying DNA function can be used for comparison and analysis of sequence patterns directly to infer function, large-scale analysis of the messenger RNA and protein products of genes, and various approaches to gene disruption.

Genes could also be identified on the basis of similarity to sequences present in EST databases. Using EST data will identify and assist in modelling more genes than will be found by methods based solely on protein/cDNA homologies. The coverage of ESTs is higher than the coverage based on well-characterised proteins since the public EST databases of human and mouse ESTs contain in total more than 1 million sequences, representing >80% of the genes expressed in these organisms. Finding good matches to ESTs is a strong suggestion that the region of interest is
expressed and because the data in EST databases represent cDNA made from a large number of tissues, can give some indication as to the expression profiles of the located genes. The technology and applications of microarrays of immobilised cDNAs and oligonucleotides, in which labelled probes are used to determine complementary binding has been successfully applied to the simultaneous expression of many thousands of genes and to large-scale gene discovery, as well as mutational screening (diagnostics) and mapping of genomic DNA clones (Ramsay G, 1998). Comparatively massive cDNA microarrays for large-scale gene discovery in infant brain tissue was recently used by Drmanac et al., (1996), where 20,000 genes were identified with a further 20,000 expressed at low levels. Such DNA chip technology can be used in future expression studies, genetic mapping as well as mutational analyses of any novel retinal genes found during this study.
REFERENCES
REFERENCES


References


Bartley, J., Gies, C., Jacobson, D. (1989). Cone dystrophy (X-linked) (COD1) maps between DXS7 (L1.28) and DXS206 (XJ1.1) and is linked to DXS84 (754). *Cytogenet. Cell Genet.* 51: 959.


References

**Nucleic Acids Research.** 21:2963-2965.


References


361
References


*Science.* **238**: 806-812


References


References


References


Jones, K.A., Black, D.M., Brown, M.A., Griffiths, B.L., Nicolai, H.M., Chambers, J.A.,
Bonjardim, M., Xu, C-F., Boyd, M., McFarlane, R., Korn, B., Poustka, A., North, M.A.,
400 kb cosmid walk in the BRAC 1 region: identification and localisation of 10 genes


References


Mears, A.J., Gieser, L., Yan, D., Chen, C., Fahrner, S., Hiriyanna, S., Fujita, R.,
the RP2 gene in a North American cohort of families with X-Linked Retinitis


References


References


References


References


*Trends in Genetics.*, 8(11): 392-398


*Cell* 65: 905-914.


PUBLICATIONS
X-linked retinitis pigmentosa (XLRP) is genetically heterogeneous with two clinically indistinguishable forms (RP2 and RP3) mapping in close proximity on proximal Xp. The region containing the RP2 gene (Xp11.3-11.22) corresponds to the critical regions for two other retinal diseases (Congenital Stationary Night Blindness and Aland Island Eye Disease). We have refined the interval containing the RP2 gene and thus excluded several candidate genes expressed in the retina. Alongside XLRP family analysis, we have been constructing a physical map centered around the marker DXS426. Heterogeneity analysis suggests the RP2 gene most probably lies within the 2cM distal to DXS426. On this basis we have embarked on creating a transcriptional map in the proximity of DXS426 utilising well characterised YACs from the contig. cDNA selection was applied to 4 YACs covering approximately 500kb using 2 retinal cDNA libraries. The retinal libraries used were found to contain both PFC and TIMP1 which map to the YACs. The YACs chosen from the region of interest therefore contain these internal positive controls. The abundantly expressed gene rhodopsin was also used to isolate a YAC for use as a positive control for the technique. cDNA/YAC DNA hybrids were captured using streptavidin coated magnetic beads. Two rounds of selection were performed and preliminary data suggests at least 700 fold enrichment for positive control genes such as TIMP1. Sublibraries have been created for enriched cDNAs derived from each YAC. These libraries are presently being screened for contaminating sequences such as ribosomal genes. Novel, non repetitive clones will be mapped back to the contig and further analysed. This work should provide candidate genes for the several ocular diseases mapping to proximal Xp.
Mapping the RP2 Locus for X-linked Retinitis Pigmentosa on Proximal Xp: A Genetically Defined 5-cM Critical Region and Exclusion of Candidate Genes by Physical Mapping

Dawn L. Thiselton,1 R. Mark Hampson,1,2 Manimekelei Nayudu,1 Lionel Van Maldergem,3 Mitchel L. Wolf,4 Bratin K. Saha,5 Shomi S. Bhattacharya,1 and Alison J. Hardcastle1,6

1Department of Molecular Genetics, Institute of Ophthalmology, London, UK; 2Rayne Institute, St. Thomas’ Hospital, London, UK; 3Institut de Pathologie et de Genetique A.S.B.L., B-6280, Gerpinnes (Loverval), Belgium; 4Jewish Hospital, St. Louis, Missouri 63108; 5Winship Cancer Centre, Emory University School of Medicine, Atlanta, Georgia 30322

Genetic linkage studies have implicated at least two loci for X-linked retinitis pigmentosa (XLRP) on proximal Xp. We now report a defined genetic localization for the RP2 locus to a 5-cM interval in Xp11.3–11.23. Haplotype analysis of polymorphic markers in recombinant individuals from two XLRP families has enabled us to identify DXS8083 and DXS6616 as the new distal and proximal flanking markers for RP2. Using STS-content and YAC end-clone mapping, an ~1.2 Mb YAC contig has been established encompassing the proximal RP2 boundary and extending from TIMP1 to DXS1240 in Xp11.23. Several ESTs have been positioned and ordered on this contig, one of which is novel to the region, identified by sequence database match to a physically mapped YAC insert terminal STS. Integration of the genetic and physical data has placed four retinally expressed genes proximal to DXS6616, and thereby excluded them from a causal role in RP2. This work now provides a much needed focus for positional cloning approaches to isolation of the defective gene.

Retinitis pigmentosa (RP) is a group of hereditary progressive retinal degenerations characterized by night blindness, visual field impairment, and degenerative pigmentary changes in the retina. RP exists as autosomal dominant, autosomal recessive, and X-linked forms and displays considerable genetic heterogeneity with at least 15 distinct loci so far assigned to human chromosomes (for review, see Dryja et al. 1995). X-linked retinitis pigmentosa (XLRP) is the most severe clinical form, accounting for 7–30% of all cases, depending on the population studied, with an incidence of ~1:20,000 (Jay 1982; Heckenlively 1983). Male XLRP patients generally develop concentric visual field loss before the 20th year of life leading to severe visual handicap by the age of 40 (Bird 1975). Female carriers show variable symptoms of the disease on ophthalmological testing, with visual impairment usually beginning in middle age, although absence of ocular abnormalities does not exclude the carrier state (Arden et al. 1983).

In the absence of functional clues as to the pathophysiology of XLRP, positional cloning strategies have been adopted to isolate the defective genes. Following the first genetic linkage of an RP gene (designated RP2) to Xp11.3 in a panel of British families (Bhattacharya et al. 1984), subsequent genetic analyses have indicated the existence of at least three other XLRP loci (RP3, RP6, and RP15) located more distally on Xp (Musarella et al. 1990; Ott et al. 1990; Teague et al. 1994; McGuire et al. 1995) and the fact that the disease in some families maps to none of these locations suggests the possibility of even more XLRP loci (Aldred et al. 1994). As evidence for RP6 is to date only statistical (Ott et al. 1990), and RP15 has been demonstrated in only one family (which is
reported as a cone-rod degeneration; McGuire et al. 1995), the majority of XLRP families fall into the categories of RP2 or RP3. The ability to distinguish between RP2 and RP3 is dependent on the detection of crossovers dissecting the target region by genetic linkage/haplotype analysis, attributable to the lack of reliable clinical differences between the two disease entities (Wright et al. 1991). The clustering of XLRP genes on proximal Xp makes such genetic distinction of XLRP families difficult; however, from cumulative genetic data it appears that RP3 predominates in British and American families (Musarella et al. 1990; Ott et al. 1990; Teague et al. 1994).

Precise localization of the RP3 gene to Xp21.1 by genetic linkage analysis has been augmented by the molecular genetic analysis of RP3 patients with submicroscopic DNA deletions, confining RP3 to a 530-kb stretch of DNA flanked by genetic markers OTC and DXS1110 (Roux et al. 1994). The focusing of efforts afforded by the identification of M.O., a male RP patient harboring a 75-kb deletion within the RP3 critical region, has recently led to the isolation of a gene (RPG) in which mutations account for XLRP in a proportion of RP3 patients (Meindl et al. 1996).

The RP2 gene has remained broadly localized to an -13-cM interval in Xp11.22-11.3 flanked by DXS7 and DXS255, owing to a lack of recombination events in the critical region (Friedrich et al. 1992; Wright et al. 1991) and no detectable disease-associated deletions. Various multipoint and heterogeneity analyses further suggest different locations for the RP2 gene within this interval (Xp11.23; Teague et al. 1994; Xp11.22; Bergen et al. 1995). Narrowing of the RP2 region will improve reliability of carrier detection and may facilitate characterization of the gene and its mutations and help resolve the issue of clinical and genetic heterogeneity.

We present here two XLRP families in which the gene responsible is consistent with an RP2 location, and where key recombinants have been identified that define both the proximal and distal boundaries of the RP2 critical interval within each family. Integration of the genetic and physical maps of the region has led to the exclusion of several retinally expressed candidate genes located in Xp11.23, and the degree of refinement now renders the construction of a physical contig spanning the RP2 critical region, and isolation of candidate transcripts therein, a more manageable task.

RESULTS

Haplotype Analysis

Key recombination events within families NRP and F72 are shown in Figure 1. No deletions were detected using the 19 microsatellites listed.

Family NRP

In family NRP (Fig. 1a) individual V-1, an affected male, is a recombinant with respect to his carrier mother IV-1 between markers MAOB and DXS1055. The phase of the maternal alleles was established firmly from analysis of two obligate carrier sibs (IV-2 and IV-3) who share the mother’s affected haplotype. This crossover defines the distal boundary of the XLRP interval, with the disease-associated haplotype located proximal to MAOB, providing firm evidence for RP2 segregating in family NRP. A similar crossover is seen in individual III-4, a carrier female who has inherited her mother’s affected chromosome proximal to MAOB and passed this recombinant chromosome onto her affected son IV-7. The phase of the alleles in II-2 could be clearly deduced from the haplotypes of her affected sons III-3 and III-6, assuming crossover minimization.

The proximal boundary of the RP2 interval is defined by an inferred crossover in individual IV-4 between markers DXS1055 and DXS6616, which has been transmitted to her affected son V-3. It is assumed that this recombination event occurred in individual III-2, as individual IV-5 has the same affected haplotype as her maternal uncle III-1 (as seen in his daughters IV-1, IV-2, and IV-3). The disease-associated haplotype in individuals IV-4 and V-3 is therefore located distal to DXS6616. A crossover event in individual III-5 provides additional support for the location of the XLRP gene in this family. This affected male is a recombinant with respect to his carrier mother II-2 between markers DXS6616 and DXS6941, with the disease-associated haplotype located distal to DXS6941.

In summary, key recombination events in family NRP are consistent with RP2 segregation, and indicate an RP2 critical region flanked by MAOB (Xp11.3) and DXS6616 (Xp11.23).

Family F72

In family F72 (Fig. 1b) individual II-5 is a recombinant with respect to his carrier mother I-1 between markers DXS8083 and DXS1003. The phase of the maternal alleles was deduced from
the haplotypes of her phase-known unaffected and obligate carrier daughters. The disease-associated haplotype in affected male II-5 is located proximal to DXS8083, confirming the segregation of RP2 in this family, and providing a new distal flanking marker for the RP2 critical region.

A recombination event in individual III-1, an unaffected male, defines the proximal boundary of the RP2 interval in family F72. A crossover has occurred between markers DXS6616 and DXS6941, positioning the disease-containing interval distal to DXS6941. The phase of the material alleles was firmly established from analysis of the grandparental genotypes.

The crossover data clearly defines F72 as an RP2 family, with the disease gene critical interval flanked distally by DXS8083 (Xp11.3–11.23) and proximally by DXS6941 (Xp11.23).

**Figure 1** Pedigrees of the XLRP families used in this study, showing haplotypes constructed with the markers listed. Solid bars indicate those alleles that are linked to the XLRP mutation (i.e., disease-associated haplotype). In the case of recombinant individuals (marked by an X) the solid bar is used to depict only those alleles that can be unambiguously linked to the parental “affected” haplotype. Hatched lines represent markers that were uninformative. (A) Family NRP, with recombinant individuals IV-7, V-1, and V-3 localizing the RP2 gene to the region between markers fvlAOB and DXS6616. (B) Family F72, with recombinant individuals II-5 and III-1 positioning the RP2 gene between markers DXS8083 and DXS6941.
using markers from Xp11.23 established a contig spanning ~1.2-Mb (based on the additive size of minimal tiling path YACs) from *TIMP1* to DXS1240 (Fig. 2). YACs were initially isolated from the library by PCR screening using markers DXS426, ZNF81, MG61, DXS722, and GATA1. STSs were derived from the left and/or right ends of the human DNA inserts in several YACs (detailed in Table 1) and proved instrumental in confirming YAC overlaps and integrity. Of the 10 total YAC insert terminals isolated, two proved to derive from regions other than Xp11.23 on somatic cell hybrid analysis; 34AC5LA maps to distal Xp and C01160LA maps to chromosome 6, indicating that these YACs are chimeric. FASTA database identity searches detected matches for two YAC insert terminals with gene sequences: 4HG2LA lies within exon 2 of the *SYN1* gene, anchoring this end of the YAC firmly in Xp11.23, whereas 34AC5RA showed 99.5% identity over 210 bp to a human infant brain cDNA [Integrated Molecular Analysis of Genomes and their Expression (IMAGE) Consortium; expressed sequence tag (EST) accession no. H09726] or 99% identity over 209 bp to a human placental cDNA (IMAGE Consortium; EST accession no. N41839) for forward and reverse strands, respectively. This result suggests that these cDNAs are identical, and maps a new EST (expressed in brain and placenta) to the physical map of Xp11.23 (Fig. 2).

The contig comprises 18 YACs and encompasses 30 markers including 11 microsatellites, 10 gene/ESTs and eight new STSs derived from YAC insert terminals, to give a STS density of ~1 every 40 kb. Key markers (e.g., F0701LA, DXS6849, and DXS6950) and two YACs, ICRFy900C1022 (SYN1) and ICRFy900C01160 (2bC6), have been included to allow integration with other partial YAC contigs already existing for Xp11.23 (Coleman et al. 1994; Hagemann et al. 1994; Knight et al. 1994; Fisher et al. 1995; Kwan et al. 1995), which reflect the interest in this gene-rich region of Xp, to which many genetic diseases have been mapped (Nelson et al. 1995). The depth of the contig permitted physical ordering of most STSs/ESTs assigned to the contig, as shown in Figure 2. The order so derived is as follows: Xpter-*SYN1*(CA)ₖ-*TIMP1-*C01022RA-4HG2RA-PFC- (CA)ₖ-33CA11LA-(DXS426, F0701LA)-4HG2LA-(DXS1367, ELK1)-33CA11RA-(ZNF81, DXS6849, DXS1004E)-DXS6616-DXS6950-34AC5RA-3ODH10RA-(MG61, DXS6941)-MG81-(DXS722, MG21)-DXS1011E-MG44-C01160RA-GATA1-DXS1126-DXS1240-Xcen. Interestingly, our lo-

---

**Figure 2** YAC contig encompassing the RP2 proximal flanking marker, DXS6616, and adjacent markers in Xp11.23 (not to scale). The horizontal line with bidirectional arrows represents the chromosome, with STSs above in the order determined by YAC content. Where two STSs proved inseparable, they are assigned to a single position. The approximate position of the OATL1 pseudogene cluster is indicated by the broken horizontal line (Nelson et al. 1995). The bold horizontal lines represent YACs with sizes given where known. All YACs are derived from the ICI YAC library (Anand et al. 1990) except ICRFy900C1022 and ICRFy900C01160 (Larin et al. 1991). A filled rectangle at the end of a YAC indicates a terminal sequence STS that maps to overlapping YACs in the contig. A hatched box at the end of a YAC depicts a terminal sequence STS that does not derive from this region of Xp11.23. The asterisk denotes the map position of a new EST in Xp11.23, corresponding to 34AC5RA (see text). Vertical broken lines show the positions of STSs in YAC clones, with a solid circle indicating a verified positive. Based on the most parsimonious STS order, YAC 34AC5 appears to harbor an internal deletion for MG81 (indicated by parentheses). The shaded bar denotes the region of the contig contained within the newly defined RP2 critical interval. The new proximal flanking marker for RP2, DXS6616, is underscored, and the retinally expressed cDNAs thereby excluded from a role in the disease are in boldface.
Table 1. Details of Gene-based and YAC Insert STSs Derived in This Study

<table>
<thead>
<tr>
<th>STS name</th>
<th>Source</th>
<th>PCR primers (5’-3’)</th>
<th>(T_a) (°C)</th>
<th>Product size (bp)</th>
<th>Accession no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GATA1</td>
<td>Gene 3’UTR</td>
<td>ACAGACATGGCGCTCCAGAG AGCTTTGAAGGGTTCAGG</td>
<td>62</td>
<td>112</td>
<td>DXS9855E</td>
</tr>
<tr>
<td>MG21</td>
<td>Gene 3’UTR</td>
<td>AATTATGCCTGGACGACGCTCA CTCCGCTCAGGACGCTTGAG</td>
<td>62</td>
<td>323</td>
<td>DXS9867E</td>
</tr>
<tr>
<td>MG44</td>
<td>Gene 3’UTR</td>
<td>TGAGGCTGTTTCTGCTGTCATTA TAGATTCGTCCTCCAGCTGACTG</td>
<td>62</td>
<td>494</td>
<td>DXS9868E</td>
</tr>
<tr>
<td>MG61</td>
<td>Gene 3’UTR</td>
<td>ATCTTGACACCTCATACCTCTT GGTACCCTCCTCTGTTGCTGGTCCAAA</td>
<td>62</td>
<td>252</td>
<td>DXS9869E</td>
</tr>
<tr>
<td>MG81</td>
<td>Gene 3’UTR</td>
<td>ATCTCGATTCCTTGGCTAGGGA AAGCAAGACATCCACA CATACAGG</td>
<td>62</td>
<td>123</td>
<td>DXS9870E</td>
</tr>
<tr>
<td>C1022RA</td>
<td>YAC end</td>
<td>ATATAATCCAGGAGCATGAGTCAGA CTCTGTGCTTGCTGCTG</td>
<td>64</td>
<td>125</td>
<td>DXS9875E</td>
</tr>
<tr>
<td>4HG2LA</td>
<td>YAC end</td>
<td>CTTGAAAGGAACCAAGCAAGC TGTGTCCTGATTGATTAGGC</td>
<td>57</td>
<td>162</td>
<td>DXS9872E</td>
</tr>
<tr>
<td>4HG2RA</td>
<td>YAC end</td>
<td>TGGCCACTGAAGATGAGATGCT CCTTCAGGGGAAAATACCTTC</td>
<td>63</td>
<td>125</td>
<td>DXS9873E</td>
</tr>
<tr>
<td>33CA11LA</td>
<td>YAC end</td>
<td>ACCTGAGATTTCACCTT AACTTTAGGTCTCTGTG</td>
<td>45</td>
<td>111</td>
<td>DXS9876E</td>
</tr>
<tr>
<td>33CA11RA</td>
<td>YAC end</td>
<td>GAATTCAAGGAGACAATGCG AGCTGTCCTGCAATAGCGCT</td>
<td>56</td>
<td>134</td>
<td>DXS9877E</td>
</tr>
<tr>
<td>34AC5RA</td>
<td>YAC end</td>
<td>TACACTGTGATTAGCTGACGAGGGGATTCGATTAGTGC</td>
<td>64</td>
<td>177</td>
<td>DXS9853E</td>
</tr>
<tr>
<td>30DH10RA</td>
<td>YAC end</td>
<td>AGGTATATGTCAGGAGCCACAC ACCGAGCTGCGGGGCATGAC</td>
<td>64</td>
<td>129</td>
<td>DXS9854E</td>
</tr>
<tr>
<td>C01160RA</td>
<td>YAC end</td>
<td>TGGCTAGAGGCTCATACCTGG GTGCCATTGGATGCTTCGTG</td>
<td>58</td>
<td>172</td>
<td>DXS9874E</td>
</tr>
</tbody>
</table>

calization of DXS1004E (ZNF41) differs from that in published reports (Knight et al. 1994; Carrel et al. 1996). In contrast to the report of Knight et al. (1994), repeated testing in our laboratory has confirmed YAC ICRFy900C1022 to be positive for DXS1004E. The additional presence of DXS1004E in 1C1 YACs 3EB3, 24GHS and 31AG10 which overlap C1022, positions ZNF41 proximal to ELK1, and distal to DXS6616. The physical ordering and orientation of the retinal cDNAs MG21, MG44, MG61, and MG81 on our contig is in agreement with that depicted on the recently published physical map of Boycott et al. (1996). Although orders could not be established between all pairs of markers, our physical mapping data demonstrates clearly that DXS6616, the new proximal flanking marker for RP2, lies distal to retinally expressed genes MG21, MG44, MG61, and MG81.

**DISCUSSION**

XLRP is a progressive degenerative disease of the retina that is clinically and genetically heterogeneous. In this study we performed haplotype analysis in two XLRP families with 20 polymorphic microsatellite markers spanning the critical region for the most prevalent forms RP3 and RP2 (Xp21.1–Xp11.22). Analysis of newly positioned markers in recombinant individuals from these families has enabled us to define new proximal and distal boundaries for the RP2 gene critical...
interval and reduce significantly its size. The new flanking markers in family NRP are MAOB (Xp11.3) and DXS6616 (Xp11.23), whereas those in family F72 are DXS8083 (Xp11.3–11.23) and DXS6941 (Xp11.23). Taken together, assuming the defective gene is the same in both families, this data suggests a refined RP2 interval flanked by DXS8083 and DXS6616 on proximal Xp, spanning 4–5 cm. Positional cloning efforts can now be directed at this greatly reduced interval, facilitating the search for an RP2 candidate gene. In the event that these two families are segregating different loci, microheterogeneity within this interval may only become evident once the disease gene for one of these families has been cloned.

Genetic heterogeneity is evident in many inherited retinal degenerations, reflecting the eye’s limited repertoire of responses to a variety of genetic lesions. Another emerging pattern in ophthalmic genetics is that of “gene sharing” (allelic) in which different mutations within the same gene can cause clinically distinct ocular diseases. Two forms of autosomal congenital stationary night blindness (CSNB) have been shown to be allelic to RP: Both CSNB and RP can result from mutations in (1) the rhodopsin gene (Dryja et al. 1993; Rao et al. 1994); and (2) the gene encoding the β-subunit of the rod cGMP phosphodiesterase (Gal et al. 1994). The refinement of the RP2 critical interval described here may have significant implications for the localization of other inherited X-linked retinal disorders that have overlapping map locations in Xp11.3–Xp11.22 and which may be allelic to RP2 that is, CSNBX (Aldred et al. 1992; Bech-Hansen et al. 1992; Berger et al. 1995) and X-linked progressive cone dystrophy (Hong et al. 1994; Meire et al. 1994).

The identification of DXS6616 as the new proximal flanking marker for RP2 has also enabled us to exclude several genes as potential candidates for this disorder. Four retinally expressed genes have been mapped to the OATLI region in Xp11.23 by direct selection using an OATLI YAC to screen a retinal cDNA library (Geraghty et al. 1993). We have positioned and ordered these genes with respect to DXS6616 on a 1.2-Mb YAC contig spanning the RP2 proximal boundary and show that DXS6616 lies distal to this cluster of retinally expressed genes, thereby excluding them from involvement in RP2. With an STS density of ~1 every 40 kb, including seven novel STSs and a newly mapped EST, this contig signifi-

METHODOLOGICAL

Subjects and Samples

Two families are presented: family NRP from the USA, comprising four generations with DNA for nine affected males and eight obligate carrier females, and family F72 from Belgium, comprising three generations with DNA for
four affected males and two obligate carrier females. A diagnosis of XLRP was based on detailed family history and comprehensive ophthalmological tests including fundus examination, visual field assessment, fluorescein angiography, and electroretinogram (ERG) measurements. DNA extraction from peripheral whole blood was performed using the Nucleon II kit (Scotlab).

Detection of Microsatellite Polymorphisms

The forward primer for each microsatellite was end-labeled with [γ-32P]ATP by incubating the primer at 37°C for 45 min with T4 polynucleotide kinase (New England Biolabs). The dinucleotide repeats were then amplified from 100 ng of genomic DNA as described previously (Thiselton et al. 1995). Alleles were detected by electrophoresing the PCR products on 6% denaturing polyacrylamide gels (Promega), followed by exposure to X-ray film. Specific primer details and PCR conditions for each microsatellite can be obtained from GDB (1995).

Haplotype Analysis

Nineteen microsatellite markers spanning ~25 cM from Xp21.1 to Xp11.22 (Fain et al. 1995; Thiselton et al. 1995) were used to generate haplotypes for all sampled individuals. From Xp21.1 to Xp11.22 the order of markers is known: CYBB–DXS1110–DXS556–DXS574–DXS977–DXS228–DXS7–MAOB–DXS8080 (adm0122c1)–DXS8083 (adm0242x5)–DXS1003–DXS1055–SYN1–DXS426–DXS1367–DXS6616–DXS6941–DXS1126–DXS573 (Nelson et al. 1995; Dib et al. 1996). Haplotypes were constructed assuming the minimum number of recombination events.

Construction of a YAC Contig Spanning the RP2 Proximal Boundary in Xpll.23

Sources of Initial STSs and ESTs

Details of STSs corresponding to genetic markers SYN1, PFC, DXS426, DXS1367, DXS6616, DXS6941, DXS722, DXS1126, and DXS1240 and expressed sequences TMP-1, DXS1004E (ZNFI), and DXS1011E may be obtained from GDB (1995). Primer pairs for microsatellite DXS6941 were kindly provided by A. Meindl (Kinderpoliklinik der Universität München, Germany). Published sequence information for other genes assigned to the region was used to design EST markers from the 3' untranslated regions (GATA1, MGC21, MGC44, MGC61, and MGC81) and these are described in Table 1. STSs for genes ELK1 and ZNF81 were as reported previously (Coleman et al. 1994; Knight et al. 1994).

Identification and Initial Characterization of YACs

YACs were identified for a STS or EST marker by PCR-based screening of hierarchical pools of clones from the ICI 4X YAC library (Anand et al. 1990). Intact DNA from each positive clone was prepared in agarose plugs and analysed by PCR to verify STS/EST content, and by pulsed-field gel electrophoresis (PFGE) (CHEF DR II; Bio-Rad) to assess the size and purity of the YACs present. PFGE conditions were run time 17–22 hr, 5 V/cm, with a pulse time ramped from 60-90 sec in 0.5 x TBE at 14°C. YAC sizes were estimated by comparison with yeast chromosome size standards (Bio-Rad) on 1% agarose gels by visual inspection after staining the gel with ethidium bromide. If no distinct YAC was visible, Hybond N+ (Amersham) blots were prepared from acid-nicked gels by Southern transfer (Sambrook et al. 1989) and hybridized at 65°C overnight with [α-32P]CTP-labeled total human DNA. The contig was supplemented with YACs for markers SYN1 and 2bC6 (DXS226) from the ICRF YAC library (Larin et al. 1991) through the ICRF Reference Library Database.

Creation of YAC Insert-end STSs

YAC insert ends were isolated by two methods: ALU- vector PCR using human-specific ALU primers ALE1 (5'–CCTCCAAAGCCTCTGGAAGTACAG–3') or ALE3 (5'–CCATICTGCAACTCGCAGCTCGG–3') and primers specific for the left and right arms of the pYAC4 vector (LA 5'–CAGCGTTCTGAGACGACTCTGCGCGC–3' ; RA 5'–ATATAGGCGCCAGCAACCGACCTCTGGC–3'). Vector et al. (Riley et al. 1990), using restriction enzymes PvuII, Drl, and EcoRV with 6-bp recognition sequences, which were found generally to yield large insert terminal PCR products. PCR products were purified and sequenced directly using an internal Vector et al. unit primer or pYAC4 primer (LA 5'–GTGGTTTAAAGCGCAAG–3' ; RA 5'–GTCCGAAAGCCGGCTTCAA–3'). End-fragment sequences were subjected to FASTA data-base identity searches before PCR primer design. The new STSs developed from YAC insert ends were tested against a pair of somatic cell hybrids containing portions of the X chromosome as their only human component: Kag 2.3 (Xp21.1–Xqter) and Sin 176 (deleted for Xp22.1–Xp11.22; Laffrenie et al. 1991) to provide an initial indication that they derived from proximal Xp.

YAC Contig Construction by STS/content Mapping

The YAC contig was constructed by determining the STS content of each YAC and therefore establishing overlaps between clones. The orientation of the contig was deduced from the known order of genetic markers mapping to the contig, that is, Xpter–PFC–DXS426–DXS722–DXS1126–Xcen (Nelson et al. 1995).

ACKNOWLEDGMENTS

We gratefully acknowledge the Human Genome Mapping Project Resource Centre for invaluable YAC and primer provision, and thank the Wellcome Trust (A.J.H.; grant no. 042019/Z/94/Z) and Wellcome Trust (A.J.H.; grant no. 042019/Z/94/Z) for their support. Sample collection and diagnosis for family NRP was supported by the National Society to Prevent Blindness awarded in memory of Silas Adelsheim (B.K.S.).

The publication costs of this article were defrayed in part by payment of page charges. This article must there-
fore be hereby marked “advertisement” in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES


GENETIC AND PHYSICAL MAPPING OF THE RP2 LOCUS


GENOME RESEARCH


Received July 2, 1996; accepted in revised form September 6, 1996.
Genomic Organization of the Human TIMP-1 Gene

Investigation of a Causative Role in the Pathogenesis of X-Linked Retinitis Pigmentosa 2

Alison J. Hardcastle,* Dawn L. Thiselton,* Manimekalai Nayudu,* R. Mark Hampson,*† and Shomi S. Bhattacharya*

Purpose. To evaluate the role of TIMP-1 in inherited retinal degeneration.

Methods. The genomic structure of the TIMP-1 gene was established and male patients with x-linked retinitis pigmentosa 2 from five families were screened for sequence alterations by direct sequencing in all exons, exon–intron boundaries, and the 5' upstream region of the gene.

Results. TIMP-1 appears to be expressed in the retina at low levels and consists of six exons spanning a genomic region of approximately 4.5 kb on Xp11.23. No disease-specific sequence alterations were identified. A site substitution in exon 5 was observed in samples from control subjects and patients, but it did not alter the amino acid sequence of the protein product.

Conclusions. The results of this study exclude mutations in the TIMP-1 coding sequence, splice sites, and the upstream region of the gene.

X-linked retinitis pigmentosa is a progressive degeneration of the retina. Early stages of the disease are characterized by peripheral fundus changes, causing night blindness and constriction of the visual fields. As the disease progresses into the macula, visual acuity deteriorates, leading to blindness in affected males by the third decade. Genetic heterogeneity in x-linked retinitis pigmentosa has been well established by multiple genetic linkage studies, with two predominant loci mapping to Xp21.1 (RP3) and Xp11.3-11.22 (RP2). RP2 lies within a 15-cM interval between loci DXS57 and DXS255, and to date the most closely linked marker is DXS426. We recently redefined the genetic interval containing the RP2 gene to a 5-cM region on Xp11.23-p11.3 containing the TIMP-1 gene. TIMP-1 lies approximately 100 kb distal to DXS426 and was therefore considered a good candidate for RP2.

METHODS. Expression of TIMP-1. Primers designed to the cDNA sequence (accession X03124; F 5'-TGCGTTCTGGCATCTGTTGTTG-3' and R 5'-GCCAGGATTCAGGCTATCTGGGAAGAAG-3') were used to amplify a 613-base pair (bp) fragment covering exons 3 to 5 from phage lysate of an adult retina cDNA library in λZapII (Stratagene, STRATAGENE Ltd., Cambridge, UK) and an adult retina cDNA library in λgt10 (gift from J. Nathans). Polymerase chain reaction (PCR) cycling conditions were 94°C for 1 minute, 68°C for 30 seconds, and 72°C for 30 seconds for 30 cycles. The amplified TIMP-1 cDNA was also used as a probe (Pharmacia Biotech Oligolabeling Kit, St. Albans, UK) against 200,000 plaques plated from the λgt10 library. Stringent hybridization conditions were imposed (65°C), and washes were taken to a stringency of 0.1 × SSC at 55°C.
Genomic Structure of TIMP-1. Several entries for human TIMP-1 (or EPA) in GenBank were retrieved: X03124, the original cDNA sequence; D11199, a partial genomic sequence from exons 3 to 6; and D26513, the 5' upstream region containing CAT, TRE, and Sp1 sites. Positions, sequence, and size of introns 1 and 2 were undetermined. Primer pair I1F-5'-CCAGAGGCCCAGAGAGACACG-3' and I1R-5'-TGCCAGAAGCCAGGGGCTCA-3' designed from database entries X03124 and D26513 were used to amplify the region of TIMP-1, which potentially spans intron 1, from genomic DNA. Another primer set, I2F-5'-CCACAGACG-GGCCCTTCTGCAAT-3' and I2R-5'-GGTCTGGTTGAC-TTCTGGTGTCG-3' designed from database entry X03124, was used to amplify a region of genomic DNA containing intron 2. Reactions were carried out in a volume of 50 µl, containing 150 ng of genomic DNA, 0.5 µM of each primer, and 0.5 U Taq polymerase (BioTaq, Bioline UK Ltd., London, UK). PCR cycling conditions were 30 cycles at 94°C for 1 minute, 62°C for 2 minutes, and 72°C for 1 minute, followed by 1 cycle at 72°C for 5 minutes. The PCR products were purified using MicroSpin S-400 HR columns (Pharmacia Biotech, St. Albans, UK), and partial sequences were obtained by direct sequencing using the primers described above (see below for direct sequencing protocol).

Mutation Screening in Patients With Retinitis Pigmentosa 2. All investigations followed the tenets of the Declaration of Helsinki, and informed consent and full institutional review board approval were obtained. Affected male patients and unaffected males from five RP2 families were selected for mutation screening. Each family had been previously defined as segregating the RP2 form of x-linked retinitis pigmentosa by detection of key recombination events using haplotype analysis (two of these families are presented in reference 5). The families studied came from the United Kingdom, the United States, Italy, and Belgium. Primer pairs were designed covering all six exons, exon-intron boundaries, and the 5' upstream region containing promoter elements. Polymerase chain reaction was performed in a volume of 50 µl, containing 150 ng of DNA, 0.5 µM of each primer, and 0.5 U of Taq polymerase (BioTaq, Bioline UK Ltd., London, UK). The reaction consisted of 30 cycles at 94°C for 1 minute, 62°C for 2 minutes, and 72°C for 1 minute, with a final cycle at 72°C for 5 minutes. The PCR products were purified using MicroSpin S-400 HR columns (Pharmacia Biotech) before sequencing.

Primers designed for PCR amplification were also used for direct sequencing. Five MicroMolar µM of primer (including I1F/I1R, I2F/I2R) was end labeled using 1 µl of [β32P]dATP in a 10-µl reaction following the manufacturer's instructions (Pharmacia Biotech). Reagents from the Pharmacia T7 DNA polymerase sequencing kit were used for direct sequencing. Ten microliters of column-purified PCR product was mixed with 2 µl of labeled primer and 1 µl of 10 mM DTT. After denaturation at 95°C for 5 minutes, the samples were placed on ice and 1 µl of 10 mM DTT, 2 µl of label mix A, and 2 µl of T7 DNA polymerase (1:4 dilution) were added. Four microliters were then mixed with each 2.5-µl aliquot of prewarmed termination mix and incubated at 37°C for 5 minutes. Reactions were terminated with 4 µl of stop solution, heat denatured, and electrophoresed on 6% denaturing acrylamide gels.

RESULTS. Immunohistochemistry has demonstrated TIMP-1 protein expression in all nuclear layers of the retina. To corroborate TIMP-1 expression in the retina, we amplified the cDNA from phage lysates of two retina cDNA libraries. The expected 618-bp product was present in both libraries, confirming expression of this protein in the retina (data not shown). To determine the prevalence of the TIMP-1 transcript in retinal tissue, a retina cDNA library was plated out and probed with TIMP-1 cDNA. One positive was identified from approximately 200,000 plaques, compared to 600 positives identified with a rhodopsin cDNA probe, implying that the transcript may be rare.

To perform an effective mutation screen of all exons, the genomic structure of the TIMP-1 gene was determined. Primer pairs designed to known TIMP-1 cDNA sequence facilitated the amplification of introns 1 and 2 (fragment sizes of approximately 1 kb spanning intron 1 and 1.6 kb spanning intron 2). Introns 1 and 2 were partially sequenced and the genomic structure of the gene was inferred from database entries and donor–acceptor site consensus. Consensus splice sites confirmed intron–exon boundaries throughout this small gene, consisting of six exons and spanning approximately 4.5 kb. Accession numbers for partial sequences of introns 1 and 2 are L47357 and L47361. The genomic structure of TIMP-1 is described in Table 1.

The small size of the TIMP-1 gene permitted a comprehensive approach to mutation screening by means of direct sequencing using the primer pairs described in Table 2. Analysis of patients from five unrelated RP2 pedigrees revealed no disease-specific sequence anomalies. Only one polymorphism was caused by a silent site substitution identified in exon 5 at nucleotide position 434 (accession X03124). This converts TTC phenylalanine to TTT phenylalanine and presumably is a result of the common C-to-T conversion at CG dinucleotides. The level of polymorphism was not determined because highly informative markers were available close to TIMP-1. In conclusion, no disease-specific mutations were identified, indicating it is unlikely that TIMP-1 is involved in the etiology of RP2.
**TABLE 1. Intron/Exon Structure and Splice Junctions for TIMP-1**

<table>
<thead>
<tr>
<th>Exon</th>
<th>3' Splice Junction*</th>
<th>Exon Size (bp)</th>
<th>5' Splice Junction*</th>
<th>Intron Size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>—</td>
<td>43</td>
<td>GAGgaacggaggggtcgccagcagggaccggagc</td>
<td>~1000</td>
</tr>
<tr>
<td>2</td>
<td>cgggcctcttcgtatatattgcccacagAAC</td>
<td>129</td>
<td>TCGggagcttigtatccactccacaactctgcc</td>
<td>~1600</td>
</tr>
<tr>
<td>3</td>
<td>tgcactatatctctctctgtgccCTCA</td>
<td>80</td>
<td>AAGgtcctttgcgtccaatccgtttcccagcaggctcaaa</td>
<td>188</td>
</tr>
<tr>
<td>4</td>
<td>aagacacettcctccctcatactcaacagtGTG</td>
<td>127</td>
<td>CTGggagggggtcggcctgcacccacaacaagcttg</td>
<td>205</td>
</tr>
<tr>
<td>5</td>
<td>gtcggtccccccctctctctcttgAAA</td>
<td>125</td>
<td>ACACgtgagtgtgtcctggaccctgcctccagcaggtcctgg</td>
<td>804</td>
</tr>
<tr>
<td>6</td>
<td>ctcagagatgtttccctcctcccttccGTG</td>
<td>267</td>
<td>—</td>
<td>NWGgtrwgt</td>
</tr>
</tbody>
</table>

(Consensus) yyyyyyyyyyny^NNN

* Capital letters represent exonic sequence. Splice junction consensus is represented as follows: w, A or T; r, A or G; y, C or T; s, 0 or C.

**DISCUSSION.** The precise expression pattern of TIMP-1 in the nuclear layers of the retina was not pursued in this report because no disease-causing mutations were identified. However, other reports have observed secreted TIMP-1 by cultured retinal pigment epithelium cells and immunohistochemistry demonstrating expression in primate retina cell bodies. Because the activity of this gene product must be tightly regulated after translation and at the level of gene expression, it is perhaps no surprise that little transcript could be detected in the retina cDNA libraries. Potentially, a gene defect may lead to disruption of photoreceptor–matrix interactions and induce apoptosis, although the role of TIMP-1 in maintaining retina structure and function has not been investigated. Because members of the TIMP family are also being increasingly implicated in degenerative diseases, we investigated whether TIMP-1 had a role in the pathogenesis of RP2. Although no mutations were detected, controlling elements outside the regions studied may eventually be implicated.

Because the critical interval containing the RP2 gene is large (5 cM), many more candidates may be mapped to this gene-rich region. We are creating a physical map in the form of a Yeast Artificial Chromosome (YAC) contig. Combined with genetic mapping in families with x-linked retinitis pigmentosa, our physical map has already enabled us to exclude several candidate genes expressed in the retina. The YACs are being used in cDNA selection experiments to isolate novel transcripts from a retina cDNA library to provide further candidates for the disease.

**TABLE 2. Primers and Conditions for Polymerase Chain Reaction Amplification of the TIMP-1 Gene**

<table>
<thead>
<tr>
<th>Region Amplified</th>
<th>Primer Pairs (sense/antisense, 5' to 3')</th>
<th>Fragment Size (bp)</th>
<th>Annealing Temperature (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5' upstream 1</td>
<td>AACCCTACCATCTCCAGACA GTGGAGGGGATAAGGCGAG</td>
<td>268</td>
<td>62</td>
</tr>
<tr>
<td>5' upstream 2</td>
<td>GACTCTGTGATCTGGTGAAG AGCTCCAGAATCTATCCAATCA</td>
<td>279</td>
<td>62</td>
</tr>
<tr>
<td>5' upstream 3</td>
<td>CAGAAATATGTTGAAAGGCTGAG CTAGGAGGAGGATAAAATGT</td>
<td>383</td>
<td>62</td>
</tr>
<tr>
<td>Exon 1</td>
<td>GATGAGTAAATGCAATCCAGGAA CCAGCGCAAGCTTCTGAGA</td>
<td>255</td>
<td>62</td>
</tr>
<tr>
<td>Exon 2</td>
<td>GTGGAGATACCTGCCAACGT GGGGGACCTGAGCTTCTAG</td>
<td>299</td>
<td>62</td>
</tr>
<tr>
<td>Exon 3</td>
<td>GGCATTGGAACGGGATGTCCTG TGAATATGGCTTGGAAAGGGC</td>
<td>163</td>
<td>62</td>
</tr>
<tr>
<td>Exon 4</td>
<td>TGGGAGGAGGTGCTGTGGCTTGGTGC</td>
<td>209</td>
<td>62</td>
</tr>
<tr>
<td>Exon 5</td>
<td>ATCCACCCGCATGCTTCTGGTCGTC GAAGCATCTGGGAGTGGAG</td>
<td>459</td>
<td>62</td>
</tr>
<tr>
<td>Exon 6</td>
<td>GCAGGCTCAGTGGGCAAGA AAGGACATTGGGATCCCTCATC</td>
<td>210</td>
<td>62</td>
</tr>
</tbody>
</table>
Key Words
candidate genes, genomic structure, mutation screening, tissue inhibitor of metalloproteinases-1, X-linked retinitis pigmentosa

Acknowledgments
The authors thank Professor J. Nathans for providing them with a retina cDNA library. They also thank the Wellcome Trust and the Ulverscroft Foundation for their support.

References