Physical and Transcriptional Mapping Studies

Within the Retinitis Pigmentosa

Critical Region on Chromosome 7p.

Reshma J. Patel.

Institute of Ophthalmology
University College London

Thesis submitted to the University of London for
the degree of Doctor of Philosophy.

1998
Declaration

I declare that this thesis submitted for the degree of Doctor of Philosophy is composed by myself, and the work herein is my own, or that the author involved is clearly stated.

Reshma J. Patel B.Sc. Hons.
Abstract

Revolutionary developments in molecular genetic procedures in the past decade have accelerated the discovery of the genetic cause for numerous single gene disorders.

This project has focused on employing positional cloning approaches to identify genes that are implicated in disease. These were applied to one form of autosomal dominant retinitis pigmentosa (RP9), an inherited disease of the retina, in which progressive degeneration of the photoreceptors leads to night blindness and ultimately the complete loss of vision. RP9 had previously been genetically linked to chromosome 7, and genetic refinement located the disease within an interval of approximately 1.6 cM between D7S795 and D7S484. With the absence of candidate genes in the critical region, the focus of this study primarily involved physical mapping procedures. These were used to provide a cloned resource onto which genes and STSs could be mapped, and for use in gene identification methods to identify transcribed sequences that lie within the disease interval. These could then be treated as potential candidate genes for further analysis. Physical mapping in the RP9 disease region on 7p14-15 was initiated by screening YAC libraries with genetically linked markers and integrating YACs which were available from the chromosome 7 YAC resource. *Alu*-PCR and STS content strategies were used to identify overlaps of clones. The isolation of YAC terminal sequences permitted the progression of chromosome walking. This work contributed to a detailed contig that spans the RP9 critical region and extends distally. The contig was saturated with additional STSs which allowed the placement and ordering of six transcripts distal to the critical region. YAC clones were sized by PFGE to provide a physical size estimation of approximately 5,900 kb for the contig and 3,200 kb for the disease interval.

Since no genes in the disease region had been identified at the commencement of this study, two strategies for detecting and isolating genes were then employed. The positional candidate gene approach is rapidly becoming the favoured approach to identify genes within critical intervals. To serve the RP9 project and to benefit other chromosome 7 positional cloning endeavours, 30 known chromosome 7 ESTs and
several genes were mapped to sub-regional locations by PCR through a chromosome 7 human-mouse somatic cell hybrid panel. Five of the ESTs revealed substantial nucleotide and amino acid homology with known gene sequences in database records. One was identified as a homologue of a rat gene. Those ESTs that mapped to the 7q14-15 sub-region were tested by PCR assay for their presence in the RP9 YAC contig. Two ESTs mapped onto the contig, one within (EST02120) and one distal to the disease region. Further analysis was carried out by members of the group to determine whether EST02120 was truly part of a gene.

A second strategy involved testing a novel cDNA hybrid selection method. The sandwich selection hybridisation method (Yan and Swaroop 1994) is based on solution hybridisation of cDNAs to target genomic clones. As the sandwich selection method was new and had not been widely discussed in the literature, the main aim was to test this system using a control strategy. A cosmid containing the blue opsin gene was used to isolate its cDNA counterpart as a test for the method's efficiency and sensitivity. This demonstrated that the technique may be sensitive to the size of the genomic source used in the procedure. The advantages, limitations and certain refinements are discussed. This technique was also applied to cosmid clones from the RP9 region in a further attempt to identify and isolate transcribed sequences within the disease region. Selected cDNAs were tested to ascertain whether any mapped back to their respective target genomic clone, and to assess their nature.
Acknowledgements

I would like to thank my supervisor, Professor Shomi Bhattacharya for the tremendous support, encouragement and advice he has provided throughout the years. Also for offering superb opportunities to students and for being extremely generous in both his knowledge and time. I would also like to express my deepest gratitude to Dr. Chris Inglehearn for all his time, guidance, constant support and for being an excellent proof-reader and teacher. I wish to acknowledge the Medical Research Council for awarding the grant which supported this project.

I am indebted to two very special, hard working and generous friends, Alison and Dawn. Alison for her expert help with my “nightmare” experiments and chapters, for being constantly fun loving and care-free, and for those brilliant holidays! Dawn for her patience and time in reading this thesis, for bringing sunshine into the lab and for being such a wonderful caring friend. They have been a constant help and are extremely patient, I can’t thank you both enough. Thanks to my fellow work colleagues Jeffrey (thanks for my middle name) and Rachel for all their help and guidance; and to Dr. Anand Swaroop, Denise Yan and Ricardo Fujita at the University of Michigan.

For the wonderful organisation of the lab and for putting up with continual pestering from us all, thanks to: Kamal for his expert computer skills and for being the nicest and funniest man on Earth; Neil for being the “daddy” of us all, for his well appreciated “bullying” and for being a good listener and proof-reader; also Jill, Naheed and Zara for their brilliant organisation and massage’s! Thanks to my special friends: Binoy, for all his encouragement, generosity, humour and kindness. You deserve the world!; Debi, for being so caring, “picking me up” when I’m down and for all the laughs; Su, for entertaining me on the way home from beano’s! and for all the fun; Fabiana, for your sparkle; Alex, thanks for being a great friend, you are invited to my house for dinner any time; and to Texas Tim, for sharing your wisdom, care and friendship. Also thanks to Mai, Ro, Steph, Cheryl, Kevin, Karen, Jim, Annette, David, Marcella, Prof. David Hunt, Sin, Eranga, Sana, Mani, Emma, Zoe, Jane, Vinny, Myrto, Donna, the biochemists, and anyone who I’ve forgotten to mention here.

A big hug and thanks to Ajay, Sangeeta and Baa for being so wonderful and caring throughout. One day I’ll pay for dinner! Also to my non-scientist friends who have not forgotten me even though I have been on Mars recently!: A special thanks to Rak for being so supportive, helpful and generous; Suki for your encouragement and for always checking if I’m alive; also thanks to Suggi and Rani; and lastly thanks to Kiran, Hina, Paras and the crew for being there for the last eleven years!
Dedication

To my Mother and Father,
who have given me endless support and encouragement, and have worked very
hard to provide me with the opportunities they did not have.
Table of Contents

Declaration 2
Abstract 3
Acknowledgements 5
Dedication 6
List of Figures 13
List of Tables 16

Chapter 1 Introduction 18

1.1 The Human Genome Project. 18
1.2 Paths to disease gene identification. 18
1.3 Positional cloning. 19
  1.3.1 Genetic linkage mapping. 20
    1.3.1.1 Human genetic maps. 24
  1.3.2 Physical Mapping. 25
    1.3.2.1 In situ hybridisation. 27
    1.3.2.2 Somatic cell hybrids. 28
  1.3.3 Molecular methods of physical mapping. 29
    1.3.3.1 Sequence tagged sites (STSs), a common language for genome mapping. 30
    1.3.3.2 Pulse field gel electrophoresis technology. 30
    1.3.3.3 Yeast artificial chromosomes. 31
    1.3.3.4 Physical maps of the human genome. 34
  1.3.4 Methods of gene identification. 35
    1.3.4.1 Cross-species sequence homology. 36
    1.3.4.2 CpG island detection. 36
    1.3.4.3 Direct screening. 37
    1.3.4.4 cDNA hybrid selection. 38
    1.3.4.5 Exon trapping. 42
  1.3.5 Positional candidates and expressed sequence tags. 44
  1.3.6 Sequencing the human genome. 46
1.4 Structure of the retina. 48
  1.4.1 The neural retina. 48
  1.4.2 The photoreceptor cells. 50
1.4.3 Retinal pigment epithelium. 53
1.4.4 The rod phototransduction pathway. 53
  1.4.4.1 Excitation. 53
    1.4.4.1.1 Photoactivation. 53
    1.4.4.1.2 Transducin activation. 53
    1.4.4.1.3 Phosphodiesterase and cGMP-gated cation channels. 55
  1.4.4.2 Recovery and restoration of the dark state. 55

1.5 Retinitis Pigmentosa. 57
  1.5.1 Clinical description. 57
  1.5.2 Mode of RP inheritance. 59
  1.5.3 Autosomal dominant RP (adRP). 59
  1.5.4 Molecular genetics of adRP. 60
    1.5.4.1 Rhodopsin. 60
    1.5.4.2 rds/Peripherin 60
    1.5.4.3 Other unidentified adRP genes . 61
  1.5.5 Other inherited non-syndromic RP loci. 62
    1.5.5.1 ArRP. 62
    1.5.5.2 X-linked RP. 64
    1.5.5.3 Digenic RP. 65

1.6 AdRP on chromosome 7p13-15 (RP9). 66
  1.6.1 Genetic analysis of RP9. 66
  1.6.2 Clinical phenotype of RP9. 67

1.7 Aims of this thesis. 69

Chapter 2. Materials and Methods 70

2.1 DNA Isolation. 70
  2.1.1 Plasmid and cosmid DNA preparation. 70
  2.1.2 Purification of DNA from preparations or PCR. 71
    2.1.2.1 Phenol chloroform extraction. 71
    2.1.2.2 Ethanol precipitation. 71
    2.1.2.3 Sephacryl -S200 and S400 HR columns (Pharmacia, UK). 71
  2.1.3 Plating of YAC colonies and identification of recombinants. 72
  2.1.4 Yeast solution DNA preparation. 72
  2.1.5 Yeast DNA plug preparation. 73
2.2 DNA digestions with restriction endonucleases.

2.3 Fractionation of DNA.
2.3.1 Agarose gel electrophoresis.
2.3.2 Pulse Field Gel Electrophoresis (PFGE).

2.4 Transfer of DNA to nylon membranes.
2.4.1 Southern transfer.
2.4.2 Screening of bacterial libraries.
2.4.3 Screening of Phage libraries.

2.5 Labelling of DNA probes.
2.5.1 Random priming labelling of probes.
2.5.2 End labelling of probes.

2.6 Hybridisation of labelled probes to DNA immobilised on Nylon filters.
2.6.1 Identification of YAC DNA from pulse field blots.
2.6.2 Identification of single copy sequences from complex backgrounds.
2.6.3 Post-hybridisation washes.

2.7 Polymerase chain reaction (PCR).
2.7.1 Standard PCR protocol.
2.7.2 PCR from phage or colonies.
2.7.3 Isolation of YAC terminal sequences: Alu-vector arm PCR.
2.7.4 PCR screening YAC libraries.

2.8 Cloning.
2.8.1 Ligation of digested fragments into plasmids.
2.8.2 Ligation of PCR products into plasmids.
2.8.3 Transformation of E. coli by heat shock treatment.
2.8.4 Transformation of E. coli by electroporation.
2.8.5 Plating transformed bacterial clones and selection of recombinants.

2.9 DNA Sequencing.
2.9.1 Manual sequencing of cloned DNA.
2.9.2 Direct sequencing of PCR products.
2.9.3 Denaturing polyacrylamide gel electrophoresis.
2.9.4 Automated fluorescent sequencing.
2.10 Phage Library Manipulations.

2.10.1 Preparation of plating bacteria culture for infection.

2.10.2 Titrating and plating phage plaques onto solid media.

2.10.3 Amplification of Phage libraries.

2.10.4 Preparation of Bacteriophage DNA.

2.10.5 Transfer of Charon BS library to pBluescript Phagemid.

2.10.6 Preparation of single stranded DNA.

2.11 RNA manipulation.

2.11.1 RNA conduct.

2.11.2 In Vitro transcriptions.

2.11.3 Denaturing gel electrophoresis of RNA.

2.12 Sandwich selection strategy (Yan and Swaroop 1994; chapter 5).

2.12.1 cDNA selection system.

2.12.2 Capture and washing.

2.13 Sequence similarity searches.

2.14 Reagents and Suppliers.

2.14.1 General reagents.

2.14.2 Solutions for DNA preparations.

2.14.3 Electrophoresis analysis:

2.14.4 Southern transfer and hybridisation solutions.

2.14.5 Host cells and Growth media.

2.14.6 RNA reagents.

2.14.7 Suppliers.

2.14.8 Oligonucleotides.

Chapter 3.  Physical Mapping of the RP9 Critical Region

3.1 Introduction.

3.1.1 Genetic order of markers around the RP9 disease region.

3.1.2 Yeast artificial chromosomes: a bridge from genetic to physical maps.

3.1.3 Human YAC libraries.

3.1.4 Contig assembly.

3.1.5 Aims.
3.2 Results

3.2.1 Incorporation of seed YACs.

3.2.2 *Alu*-PCR

3.2.2.1 *Alu*-PCR with D7S526, D7S632 and D7S435 YACs.

3.2.2.2 *Alu*-PCR with D7S690 YACs.

3.2.3 Development of a novel STS, A1.

3.2.4 Isolation of additional YACs.

3.2.5 Isolation of YAC terminal sequences: *Alu*-vector arm PCR.

3.2.6 STS content mapping and development of a transcript map.

3.2.7 Estimation of the physical extent of the proximal RP9 region contig.

3.3 Discussion.

Chapter 4. Regional Assignment of Chromosome 7 ESTs

4.1 Introduction.

4.1.1 Tools for assignment.

4.1.2 Aims

4.2 Results

4.2.1 Assignment of dinucleotide markers

4.2.2 Source of ESTs.

4.2.3 EST assignment.

4.2.4 Homologies of ESTs to sequences deposited in databases.

4.2.5 The search for positional candidates for RP9.

4.2.5.1 Identification of a positional candidate EST for RP9.

4.2.5.2 Testing chromosome 7 genes as positional candidates for RP9.

4.3 Discussion.

Chapter 5. Sandwich Selection Hybridisation

5.1 Introduction

5.1.1 The sandwich selection strategy.

5.1.2 Hybridisation and selection procedure.

5.1.3 Model sandwich selection experiments.

5.1.4 Adult Retina Library in Charon BS.

5.1.5 Choice of genomic clones for use in selection.

5.1.6 The blue cone pigment gene as a model for sandwich selection.

5.1.7 Aims.
5.2 Results.

5.2.1 Amplification of AR2 library. 183
5.2.2 Library characterisation. 183
5.2.3 Conversion of AR2 library to pBluescript. 187
   5.2.3.1 Insert analysis and confirmation of circularisation 189
5.2.4 Single stranded DNA rescue. 190
5.2.5 Construction of model sandwich selection experiments. 193
5.2.6 Choice of genomic clones from the RP9 region. 194
   5.2.6.1 Vector assay for cross-contamination. 194
5.2.7 Cloning genomic DNA into pSV9zf9 vector. 195
   5.2.7.1 Genomic sub-library evaluation. 196
5.2.8 In vitro synthesis of “genomic RNA”. 196
5.2.9 In vitro synthesis of biotinylated “capture RNA”. 200
5.2.10 Selection procedure. 201
5.2.11 Analysis of selected clones. 202
   5.2.11.1 BCP controls. 204
   5.2.11.2 Sequence Analysis of random clones from the RP9 selected library. 208
      5.2.11.2.1 E2-1. 208
      5.2.11.2.2 E2-19. 215
   5.2.11.3 Sequence analysis of clone E1-10. 219

5.3 Discussion. 225

Chapter 6. General Discussion 231

6.1 Overview of work presented. 231
6.2 The future for positional cloning. 235
6.3 The future in understanding the disease mechanisms involved in retinal degenerations. 238

Publications 241

References 242
## List of Figures.

<table>
<thead>
<tr>
<th>Figure No.</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The key features involved in positional cloning.</td>
<td>21</td>
</tr>
<tr>
<td>1.2</td>
<td>The multiple levels of human chromosome mapping.</td>
<td>26</td>
</tr>
<tr>
<td>1.3</td>
<td>The general strategy for cDNA hybrid selection methods.</td>
<td>40</td>
</tr>
<tr>
<td>1.4</td>
<td>Cross-section of the human retina.</td>
<td>49</td>
</tr>
<tr>
<td>1.5</td>
<td>Schematic representation of mammalian rod and cone cells.</td>
<td>52</td>
</tr>
<tr>
<td>1.6</td>
<td>The phototransduction cascade in mammalian rod cells.</td>
<td>54</td>
</tr>
<tr>
<td>1.7</td>
<td>Fundus photographs of a) a normal retina and b) a retina of a patient with the advanced stages of RP.</td>
<td>58</td>
</tr>
<tr>
<td>1.8</td>
<td>Genetic localisations of chromosome 7p linked diseases.</td>
<td>68</td>
</tr>
<tr>
<td>3.1</td>
<td>Locations of genetic markers around the RP9 disease region.</td>
<td>104</td>
</tr>
<tr>
<td>3.2</td>
<td>Illustration of the seed clones used as starting material for contig assembly.</td>
<td>111</td>
</tr>
<tr>
<td>3.3</td>
<td><em>Alu</em>-PCR fingerprint generated by ALE1 and ALE3 primers on D7S526, D7S632 and D7S435 containing YACs.</td>
<td>113</td>
</tr>
<tr>
<td>3.4</td>
<td><em>Alu</em>-PCR fingerprint generated by the ALE1 primer on D7S690 containing YACs.</td>
<td>115</td>
</tr>
<tr>
<td>3.5</td>
<td><em>Alu</em>-PCR fingerprints generated by a) ALE1 and ALE3 primers and b) the ALE3 primer only on D7S690 containing YACs.</td>
<td>116</td>
</tr>
<tr>
<td>3.6</td>
<td>Autoradiograph showing the partial sequence generated from an <em>Alu</em>-PCR product derived from a D7S690 containing clone.</td>
<td>118</td>
</tr>
<tr>
<td>3.7</td>
<td>Nucleotide sequence from an <em>Alu</em>-PCR product from which the STS A1 was chosen.</td>
<td>119</td>
</tr>
<tr>
<td>3.8</td>
<td>Pulse field gel electrophoresis and Southern blot analysis for the size estimation of a selection of YACs.</td>
<td>122</td>
</tr>
<tr>
<td>3.9</td>
<td>a) <em>Alu</em>-vector PCR products obtained from YAC yWSS1545 b) Nucleotide sequence of the <em>Alu</em>-vector PCR product from which the STS 1545LA was chosen.</td>
<td>125</td>
</tr>
<tr>
<td>3.10</td>
<td>Nucleotide sequence of the <em>Alu</em>-vector PCR product from which the STS 34-DA5RA was chosen.</td>
<td>126</td>
</tr>
<tr>
<td>3.11</td>
<td>Nucleotide sequence of the <em>Alu</em>-vector PCR product from which the STS 27I-A5LA was chosen.</td>
<td>128</td>
</tr>
</tbody>
</table>
3.12 Chromosomal localisation of 27I-A5LA by PCR analysis on cell lines of the HGMP monochromosomal hybrid panel. 129

3.13 Diagram of the YAC contig which extends distal from the RP9 critical region. 130

3.14 Agarose gels showing the YAC STS content mapping results for A1 and 1545LA. 133

3.15 Amplification of EST00601 from cDNAs which were derived from human brain, liver and muscle tissues. 135

3.16 Selected clones from the RP9 contig from which the size of the region was estimated. 137

4.1 Regional assignment of EST06238 and AAAFENH by PCR on individual cell lines of the chromosome 7 hybrid panel. 150

4.2 Schematic representation of the regional assignment of each EST on chromosome 7. 153

4.3 BLAST printout showing the nucleotide identity of EST01644 with part of the PRKAR1B gene. 156

4.4 BLAST printout showing the nucleotide identity of EST06146 with part of the amphiphysin gene. 156

4.5 BLAST printout showing the nucleotide identity of EST06275 with a) part of the mGluR3 gene and b) BAC clone RG331C24. 158

4.6 BLAST printout showing the nucleotide identity of EST00272 with part of the human Krit1 gene. 159

4.7 BLAST printout showing the nucleotide identity of EST00167 with part of the putative endothelin receptor type B-like mRNA. 159

4.8 Diagram showing the physical map location of EST02120 in the RP9 disease region. 162

4.9 BLAST printout showing the nucleotide identity of EST02120 with part of the human Tigger1 transposable element. 164

5.1 Summary of the sandwich hybridisation selection technique. 173

5.2 Diagram depicting pSV9Zf9 and pSV7Zf3 vectors and the hybridising structure of their RNA products. 174

5.3 Diagram depicting the Charon BS(-) vector and directionally cloned cDNA. 176
5.4 The size determination of random clones from the AR2 retinal library. 185
5.5 A representative section of the AR2 phage library screen by hybridisation with a BCP specific probe. 186
5.6 Not I digested DNA samples of the AR2 library. 188
5.7 Amplification of random clones from the pBluescript form of the AR2 retinal library using M13U and M13R primers. 191
5.8 Electrophoretic comparison of single and double stranded DNA preparations on a non-denaturing agarose gel. 192
5.9 PCR amplification of random clones from the genomic shotgun cloned libraries in pSV9Zf9. 197
5.10 In vitro transcribed RNA products separated on a denaturing formamide gel. 199
5.11 A representative sample of positive BCP clones in the starting spiked library, identified by Southern blot hybridisation. 203
5.12 Southern blot hybridisation of a BCP cDNA specific probe onto duplicate filters of the selected material from a) elution 1 and b) elution 2. 206
5.13 Nucleotide sequence of clone E2-1. 209
5.14 FASTA printout showing the nucleotide identity of E2-1 with human ESTs zp62a10.r1 and zp62a10.sl. 211
5.15 FASTA printout showing the nucleotide identity of E2-1 with mouse EST v94g09.r1. 212
5.16 The position of the top scoring sequence alignments for E2-1. 214
5.17 Partial nucleotide sequence of E2-19. 216
5.18 Nucleotide identity of E2-19 with a) THC212875 and b) the THC structure and content. 217
5.19 The primer positions and orientation of sequences in the E1-10 insert. 221
5.20 Partial nucleotide sequence from the T7 end of clone E1-10. 221
5.21 Partial nucleotide sequence from the T3 end of clone E1-10. 222
5.22 FASTA printout showing the nucleotide identity of the T3 sequence of E1-10 with part of an L1 repeat. 224
### List of Tables

<table>
<thead>
<tr>
<th>Table number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Identity and size of YACs identified in this study or obtained from databases.</td>
<td>123</td>
</tr>
<tr>
<td>4.1</td>
<td>List of EST name, locus symbol, source library and the group who first reported and chromosomally assigned the EST.</td>
<td>148</td>
</tr>
<tr>
<td>4.2</td>
<td>Table displaying the individual results from PCR amplification of each EST through the human-mouse somatic cell hybrid panel.</td>
<td>151 &amp; 152</td>
</tr>
<tr>
<td>5.1</td>
<td>Increase of NRL cDNA after selection.</td>
<td>178</td>
</tr>
<tr>
<td>5.2</td>
<td>Negative selection: reduction of NRL cDNA clones with unrelated genomic RNA.</td>
<td>179</td>
</tr>
<tr>
<td>5.3</td>
<td>Positive selection: selection of NRL cDNA from a spiked retinal cDNA library using NRL genomic RNA.</td>
<td>180</td>
</tr>
<tr>
<td>5.4</td>
<td>Selection of BCP cDNA from a spiked retinal cDNA library using BCP containing “genomic RNA”.</td>
<td>205</td>
</tr>
</tbody>
</table>
Chapter 1  

Introduction

1.1 The Human Genome Project.

The human genome is estimated to extend three billion base pairs in length. Its vast size presents a great challenge in unravelling its structure, organisation, and in identifying the estimated 60,000-80,000 genes residing within it (Antequera and Bird 1994; Fields et al. 1994). The Human Genome Project (HGP) is a major international research program initiated by the National Institutes of Health (USA) and the Department of Energy (USA) to undertake this immense task. A set of goals were proposed in 1990 to produce high resolution genetic and physical maps in five years and to ultimately determine the sequence of the entire human genome and the genomes of several model organisms by the year 2005. To take on this work, large groups were organised to focus on different projects and to co-ordinate “human chromosome workshops”. The work had progressed at such a rapid rate that most of the targets were accomplished ahead of schedule and the goals had to be updated. A new objective of identifying and mapping all genes was then included (Collins and Galas 1993). The speed of these efforts has become feasible with the development of the polymerase chain reaction (PCR, Saiki et al. 1988), robotic technology and information databases which are accessible via the internet. The main function of the HGP is to supply shared resources and information to the scientific community. By obtaining the sequence of the human genome, we increase our understanding of the development and function of the human body, and also accelerate advances in discovering the molecular pathology of diseases.

1.2 Paths to disease gene identification.

Single gene disorders are identified through characteristic patterns of inheritance through families. The route chosen for the identification of the disease gene rests on the type of information known about the disease. The first disease genes that were recognised had been discovered because prior information was known about the defective protein product or function of the implicated gene. This “functional cloning” approach uses the amino acid sequence, antibodies or functional complementation assays to isolate the implicated gene, then molecular analysis to
identify the mutations causing the disease. The disease gene for phenylketonuria was discovered in this manner (Robson et al. 1982).

A second method is the “candidate gene” approach. As functional information about individual genes is constantly being unravelled, they can be considered to be candidates for involvement in a disease if they bear a close functional relationship with the underlying metabolic defect. These “candidate genes” can be homologues of counterparts that cause a similar phenotype in other organisms, or are members of a gene family which are known to be involved in the disease. For example, the gene for the $\alpha$ subunit of rod cGMP phosphodiesterase (PDE) was chosen for mutational analysis in patients with retinitis pigmentosa (RP) as the protein was known to be involved in the phototransduction pathway, the main biochemical process in the affected rod photoreceptor cells. Also, the $\beta$-subunit of rod PDE was already implicated in RP. Mutations were detected that cause some cases of recessive RP (Huang et al. 1995).

These approaches are becoming less common as the number of identified genes with a known function represents only a small fraction of the estimated total number of genes in the human genome. In addition, there are a large number of inherited single gene disorders where the underlying cause is not well understood. In these circumstances, a different approach known as “positional cloning” is adopted (section 1.3). This relies solely upon identifying the disease gene by its location in the genome rather than by its functional properties. This method, although laborious and time consuming, has recently been increasingly successful due to the efforts from the HGP initiative, namely the production of high resolution genetic and physical maps which aid localisation of disease genes.

1.3 Positional cloning.

This strategy aims to identify the disease causing gene for a genetic disorder with no foreknowledge of function, by pinpointing its position to a sub-chromosomal region in the genome. Once the disease locus is sub-chromosomally localised and the critical interval is refined to the smallest defined area possible, physical mapping is undertaken to clone the interval. This provides a resource with which to map and
identify genes. The genomic organisation of candidate genes is determined and these are then tested for mutations segregating with the disorder (figure 1.1). The time and effort taken in such pursuits is extensive and exhaustive. It took a decade, with the collaboration of several research groups, to identify the genetic defect in Huntington’s disease (The Huntington’s disease Collaborative Research Group 1993). Nevertheless, in recent years the results of such hunts have proved fruitful with the identification of many disease causing genes.

For most successful disease locus mapping projects to date, the association of a disease phenotype with a sub-chromosomal region has been achieved as a result of the presence of deletions or rearrangements in affected patients which interrupt or delete the affected gene. These are relatively easy to detect by cytogenetic or Southern blot analysis. This greatly aided the search process for the genes involved in chronic granulomatous disease (Roya-Pokora et al. 1986), Duchenne muscular dystrophy (Monaco et al. 1986) and retinoblastoma (Friend et al. 1986), the first reported successful positional cloning efforts. In some instances, the presence of certain factors have aided gene searches. Triplet repeat sequences present within, or in the vicinity of a disease gene, have been found expanded to hundreds or thousands of copies in affected individuals in several different neurological diseases such as fragile X syndrome and myotonic dystrophy (Ross et al. 1993). The majority of single gene defects however do not possess such features and therefore mapping the disease gene depends on genetic linkage analysis (section 1.3.1).

1.3.1 Genetic linkage mapping.

Genetic mapping observes the segregation of two or more non-allelic loci (traits, genes or markers) during meiosis. During meiotic cell division, each pair of homologous chromosomes undergoes at least one recombination (crossover) between non-sister chromatids. Only homologous regions of chromosomes are exchanged so that each chromosome has the same set of genetic loci. Even though each copy of the same chromosome comes from a different parent, they are more than 99% identical in DNA sequence. However, each copy contains many small variations (DNA polymorphisms), the sum of which gives each individual a unique genetic makeup and
Family inheritance pattern

Genetic linkage and finer genetic mapping

Cytogenetic abnormalities

Physical mapping

Gene identification and isolation

Mutation search

![ATTGCTCCGAGTTCGA](Normal)

![ATTGCTCCGACTTTCGA](Mutation)

**Figure 1.1** Schematic illustration of the key features involved in positional cloning.
which can be individually used as markers to determine the parental origin of the loci. If two loci are physically located close together on the same chromosome, there is more chance that they will not be separated by random crossover events and are therefore "linked". To detect linkage, the representative markers for that locus should be sufficiently polymorphic in the human population, giving a reasonable likelihood that a marker will be different in any two copies of a chromosome. This is so they can detect recombination events when typed through members of the family to display inheritance patterns. Loci that are nearby show similar inheritance patterns, thereby permitting proximity to be inferred (reviewed in Ott 1991).

The extent of genetic linkage between loci is measured by the recombination fraction (θ), which is the proportion of recombinations between two loci, in a given number of meiosis (opportunities for recombination). As the distance between two loci decreases there will be less chance of crossing over and the number of recombinants will become fewer, i.e. the recombination fraction falls. A θ value of 0 demonstrates tight linkage and a θ value of 0.5 is symptomatic of unlinked loci that segregate independently. This is the maximum value that can be produced as each sister chromatid has only a 50% chance of being involved in any one crossover. As the recombination fraction is a measure of the recombination frequency between loci, it can therefore serve as a measure of distance between them. The unit of genetic distance is the Morgan (M) and is defined as the length of a chromosomal segment which on average undergoes one exchange per individual chromatid strand. It has been estimated that the haploid genome is approximately 33 Morgans, and the average human chromosome is 1.5 Morgans i.e. it undergoes an average of 1.5 crossovers during meiosis. However in reality this number is largely dependent on the length of the chromosome. In practice, genetic distances are quoted in centimorgans (cM), whereby two loci are 1 cM apart if they recombine once in every 100 meioses. Over short chromosomal regions, θ is directly proportional to the genetic map distance, so that a θ value of 0.01 corresponds to a genetic map distance of 1 cM. However, over long distances this linear relationship breaks down due to multiple crossovers between the two loci, and the effect that a crossover has of reducing the probability of a second crossover in its vicinity. Over small areas, the physical distances can be approximated based on genetic distances. As the genetic length of the genome is 33 M and the
physical length of a haploid genome is $3 \times 10^9$ bp, 1 cM corresponds to $\sim 1$ Mb of DNA. However, as different regions of the genome vary in recombinogenicity, a 1 cM interval may in fact represent a physical distance of 0.5-3 Mb.

The simplest method of estimating the distances between closely linked loci of known order, is to add together the recombination fractions in each interval of adjacent loci. The map distance between two or more distant loci can be obtained as the sum of the map distances in the intervals between these loci. For larger distances, the recombination fraction is not an additive distance measure as values are likely to exceed 0.5. The order of loci for disease traits and many markers are not known. Recombination fraction is therefore converted into genetic distances by the use of a mapping function e.g. Kosambi mapping function, which produces more realistic map distance values for $\theta$ values of 0.1-0.3. The estimate of the recombination interval ($\theta$) between two loci is obtained by determining the value of $\theta$ which gives the maximum probability for the observed data. The probability of linkage over an assumed range of recombination fractions can be measured quantitatively and is expressed on a logarithmic scale to produce a lod ($\log_{10}$ of odds) score. This is the ratio of the probability that the data would have arisen if the loci are linked ($\theta<0.5$) to the probability that the data would have arisen if the loci are unlinked ($\theta=0.5$). The $\theta$ value that gives the maximum lod score ($Z_{\text{max}}$) is the best estimate of the degree of linkage between the two loci. The acceptable threshold for proof of linkage is with a lod score of 3 (odds=1000:1), whereas evidence against linkage is taken with a score of -2 or less. This is determined by statistical analysis using computer programs.

When employing this approach to disease gene identification, genetic mapping is employed to identify linkage between a marker of known location and the disease trait, by determining if they are inherited together among affected members of pedigree(s). The family size and the informativeness of a given marker affects the lod score. Factors including late age of onset, failure of individuals to reproduce or the mode of inheritance can make it unlikely that a single family will provide sufficient evidence for linkage, and often the sum of lod scores combined from more than one family is used. Extended, multigeneration kindreds are most effective for linkage studies, which yield a high number of informative meiosis. A high density genetic map is required with markers mapped at regular intervals (10 cM or less) to provide a
greater chance of finding one close to the disease gene and give a significant lod score. Once linked, the critical region can be refined by genotyping branches of the pedigree(s) containing key recombinant individuals with further genetic markers in the vicinity of the gene, to discover the closest flanking markers by “haplotype analysis”. This involves following the segregation of alleles, to determine whether each new marker has recombined with the disease locus or resides within the critical interval bounded by the recombination events. This approach is dependent on the number of informative meioses in family pedigrees, recombinant individuals, marker density and marker heterozygosity in the region. With the availability of high resolution genetic maps covering most of the genome (section 1.3.1.1) fine mapping can now often delimit the disease region to only a few cM. This is considered small enough to commence physical mapping and gene identification strategies (discussed in sections 1.3.3 and 1.3.4).

1.3.1.1 Human genetic maps.

The most common source of informative DNA markers at present are short tandem repeats (STRs) or microsatellites. Polymorphisms in the nucleotide sequence of these units can be detected via a PCR assay. The use of these repeats as markers is advantageous as they are abundant and ubiquitous throughout the genome (Weber and May 1989; Stallings 1991). The most frequently employed microsatellite repeat markers are (CA)$_n$ repeats which are thought to occur on average once every 30 kb in the human genome and have a repeat unit length of 10-50 copies (Stallings et al. 1991). The value of a marker depends on its degree of heterozygosity among the population.

One of the major goals of the HGP initiative was to assemble a detailed and informative human genetic linkage map at a resolution of 2-5 cM. This was to assist in localising and ordering genes and disease loci to chromosomal regions. This target was swiftly reached by several international groups who produced a series of human genome maps with newly identified markers. Most of these were genotyped in a common set of reference families from CEPH (Centre d'Etude du Polymorphisme Humain; Dausset et al. 1990), to establish linkage relationships between the markers. In 1992 two groups published linkage maps with an average resolution of 5 cM. The
NIH/CEPH Collaborative Mapping Group (1992) created a linkage map with 1416 loci, 339 of which were microsatellites and 279 genes. Weissenbach and colleagues (1992) at Généthon, reported 814 novel CA repeat polymorphic markers spanning ~90% of the human genome in their "second generation linkage map of the human genome". The same group reported their 1993-94 linkage map (Gyapay et al. 1994) which consisted of 1,267 new CA repeat markers and were integrated into their previous 1992 map, presenting a total of 2,066 poly (CA) markers with ~50% of these spaced at an average distance of 1 cM or less. The final Généthon map presents 5264 CA repeat markers with an average resolving power of 1.6 cM (Dib et al. 1996). Other groups focused on integrating markers generated from several sources. The Cooperative Human Linkage Centre (CHLC; Beutow et al. 1994) assembled an integrated map using data sets from the National Centre for Human Genome Research (NCHGR) Index Map Consortium and Généthon. This set also included tri and tetra-nucleotide repeat microsatellites, which resolve more clearly on polyacrylamide gels. A second integrated map was produced in 1994 (Murray et al. 1994) which combined genotype data generated from the CEPH, Généthon and CHLC groups. This brought together 5,840 loci, most of which were microsatellite markers to give the map an average marker density of 0.7 cM.

These high density gene maps have accelerated the pace of positional cloning efforts. The heterozygous polymorphic markers can be easily typed in families by PCR assay and reactions can be multiplexed together to reduce both time and effort. These maps have also benefited those groups interested in polygenic disease traits such as juvenile-onset diabetes, schizophrenia and multiple sclerosis (King 1997). The markers can also be readily used as physical mapping reagents to provide a bridge between physical and genetic maps (see section 1.3.3).

### 1.3.2 Physical Mapping.

Physical distances are represented in terms of nucleotide base pair units and there are varying levels of resolution when analysing the physical map of the whole genome (figure 1.2). At its lowest resolution, physical mapping begins at the cytogenetic level. The shift from the megabase (Mb) to the kilobase (kb) level is achieved by rare-cutter restriction maps and the construction of clone-based physical maps.
Figure 1.2 The multiple levels of human chromosome mapping. The cytogenetic map provides the lowest resolution where the position of markers or genes can be assigned to chromosome bands or breakpoints by *in situ* hybridisation. The genetic map measures the recombination frequency between linked markers and has a similar resolution to the radiation hybrid map. Physical maps provide more accurate measurements of distance. These may be generated by pulse field gel electrophoresis or by the ordering of overlapping sets of clones. Sequence tagged sites are used as unique reference points for ordering and map integration. The highest resolution is achieved by sequencing. These maps are essential tools in positional cloning efforts and are used to progressively track the location of a marker or gene from the coarsest to the finest resolution. (Adapted from McConkey, 1993).
maps. These form the preparatory frameworks for the ultimate map, the complete nucleotide sequence of the human genome.

When the search for a disease gene is tightly focused on a relatively small region, the next stage is to reconstruct a physical map of this critical region consisting of cloned DNA fragments, which span it and faithfully reflect the genome content. These are augmented by STS content and transcription maps. This physical map is later used for assigning markers or genes to regions, and can also be a resource for the use in gene identification strategies and ultimately for sequencing.

1.3.2.1 In situ hybridisation.

This procedure assigns unique sequences to chromosomal sub-regions, single cells or tissue sections. A probe is labelled with reporter molecules which when hybridised to target chromosomes, are cytogenetically detectable. In situ hybridisation is employed in a range of applications, principally mapping loci and also to detect chromosomal abnormalities by karyotyping. It had previously been performed using isotopically labelled probes. This has since been replaced by fluorescence in situ hybridisation (FISH), which utilises different coloured fluorescent labels that are directly observable by microscopy. The resolution that can be obtained to establish the physical order of two markers which reside on the same chromosome, depends on the cell cycle phase of the chromosomes which are spread. FISH to metaphase chromosomes can resolve the order of two probes at a minimum distance of 1-2 Mb (Trask et al. 1991). Hybridisation to interphase chromatin, which is less condensed, can resolve probes at a 0.03-1 Mb distance. Recent techniques which mechanically stretch out free chromatin fibres from interphase nuclei, can achieve a resolution of 10-20 kb (Haaf and Ward 1994).

In gene mapping, FISH has been applied to locate many genes and to order YAC and cosmid clones (Chumakov et al. 1995). It has also been used in whole genome organisation analysis, for example in determining the distribution of CpG island clusters which are indicative of gene rich areas in the genome (Craig and Bickmore 1994).
1.3.2.2 Somatic cell hybrids.

These are useful tools for mapping markers or genes to chromosomal regions. Somatic cell hybrids are formed by the fusion of cultured cells from different species, usually human and rodent. Nuclear fusion results in the preferential loss of most human chromosomes in a random fashion, however a proportion are retained and individual human-rodent hybrid cells can be propagated in culture and stably maintained as cell lines. The human chromosome content in each cell line can be determined by FISH using human genomic DNA probes. It is therefore possible to build up a panel of somatic cell hybrids containing specific chromosome combinations. Localisation of a human specific marker can then be achieved by observing the concordance and discordance of positive signals obtained by PCR or Southern blot hybridisation of the probe onto the panel. Monochromosomal somatic cell hybrid panels accommodate 24 hybrid cell lines, each retaining a different single human chromosome.

Sub-chromosomal hybrid panels can be used to map probes to specific portions of the chromosome. Individual cell lines of the panel can be made by fusing cell lines which harbour human chromosomal translocations or breakpoints that retain only these fractions of human chromosomes. A selection of different cell lines that each represent different portions of a specific chromosome, can be used to build a panel for the regional assignment of probes (e.g. see section 4.1.1, Tsui and Farrall 1991). These allow markers to be “binned” to certain regions. Somatic cell lines can also be used to identify gene transcripts from the human chromosomal retained fragments (Lui et al. 1989; Corbo et al. 1990) and for the development of chromosome or chromosomal region specific genomic libraries.

More recently radiation hybrid panels which have increased resolution capabilities, have become the new tool in establishing high resolution STS maps of the human genome (Walter et al. 1994). Whole genome radiation hybrids (WG-RH) are constructed by fusing diploid human fibroblasts that were irradiated by X-rays, with recipient hamster cells. The resolution of radiation hybrid mapping is dependent on the fragment sizes of the retained human chromosomes. The higher the radiation dose, the more fragmented the chromosomes will become. The Genebridge 4 whole genome radiation hybrid panel (Cox et al. 1990, Gyapay et al. 1996) was irradiated with 3,000
rads of X rays and consists of 93 hybrids, each retaining ~32% of the human genome in random fragments of ~10 Mb. The panel has been typed with genetic markers from all over the genome to form a “framework” map. New markers are tested in individual cell lines by PCR and localisation is determined by statistical analysis using the RHMAPPER software at the Whitehead/MIT World Wide Web site (Hudson et al. 1995), which determines linkage relative to markers already on the framework map. The linkage is expressed by a distance (centi-rays) measured in breakpoint frequency, which is variable according to the radiation dose, and is accompanied by a likelihood estimate expressed as a lod score (Gyapay et al. 1996). The G3 WG-RH panel generated at the Stanford Human Genome Centre, is of a higher resolution constructed with 10,000 rads of irradiation. It consists of 83 cell lines, which each retain ~15% of the human genome and have average sized fragments of ~4 Mb. Finally, the TNG RH panel, also generated at the Stanford Human Genome centre, has been constructed by irradiation with 50,000 rads, and provides a resolution of ~50 kb (Stewart et al. 1997).

The radiation hybrid map is supplementary to genetic maps and YAC-based physical maps. It allows the localisation and ordering of markers to regions which are not sufficiently covered with yeast artificial chromosome (YAC) clones (Chumakov et al. 1995). In addition, distances in RH mapping are less likely to be distorted as no chromosomal radiosensitivities have been identified, unlike genetic maps which are distorted relative to physical distance by hot spots for genetic recombination. These panels have been used to build an STS based map of the human genome (Hudson et al. 1995; section 1.3.3.4) and a gene map of the human genome (Schuler et al. 1996; section 1.3.5).

1.3.3 Molecular methods of physical mapping.

An increased resolution in mapping has been provided by advances in both fractionation and cloning of DNA. These are used to produce genomic libraries of clones that represent several genome equivalents. The genomic clones have subsequently been employed in creating overlapping cloned sets of genomic DNA (contigs) which collectively span a particular region and reflect its genome content.


**Introduction**

1.3.3.1 Sequence tagged sites (STSs), a common language for genome mapping.

These are any probes that define a sequence landmark and are detectable with specific primer pairs by the polymerase chain reaction (PCR) (Olson et al. 1989). These single copy sequences create a universal language for mapping and can include genetic markers, gene-based markers, expressed sequence tags (ESTs, see section 1.3.5) or ordinary random loci. STSs can be readily screened against any DNA source and can therefore be used to integrate genetic, physical and transcriptional maps. The main attraction is that they utilise rapid PCR technology, and the probes are easily accessible to the scientific community as primer pair sequences, which can be distributed via the internet. The Human Genome Project proposed to build YAC contigs with at least a 2 Mb continuity across the entire genome, comprising of evenly-spaced STSs with average intervals of 100 kb. These STSs can be used to build contigs of smaller, more stable clones (e.g. cosmids and bacterial artificial chromosomes; section 1.3.3.3) for the refined analysis of specific regions.

1.3.3.2 Pulse field gel electrophoresis technology.

The development of pulsed field gel electrophoresis (PFGE; Schwartz and Cantor 1984) has filled the gap between cytogenetic analysis and cloning vectors, as it allows fractionation of DNA molecules as large as 10 Mb. Conventional gel electrophoresis is limited to resolving fragments which are less than 50 kb, because larger fragments are constrained from migrating further when their size is the same size as pores in the agarose matrix. PFGE allows larger fragments to be resolved by introducing electric fields that change strength or direction over time. This forces DNA molecules to periodically reorientate in new directions which allows molecules to slowly adjust and move between pores. Small molecules are able to orientate faster in response to the changing electric fields than larger ones and therefore migrate further. Modifications of the original PFGE system include the commonly used Contour-Clamped Homogeneous Electric Field (CHEF; Volrath and Davies 1987). This provides a homogeneous electric field across the gel resulting in uniform migration of DNA samples and allows DNA to run straight from the tracks. DNA that is to be separated is prepared in the form of agarose plugs or beads to prevent shearing.

The approximate order and the physical distance between markers can be
determined by long range restriction mapping (Brown and Bird 1986). The marker probes are hybridised to Southern blots of complete or partially digested genomic or large cloned DNA fragments which are separated on pulse field gels. Infrequently cutting restriction enzymes are used to resolve very large fragments of DNA. If two markers co-hybridise to the same band they are physically linked. The smallest restriction fragment that is shared by both probes indicates the maximum length that could separate them. Using this approach, the physical size of disease intervals can be accurately determined, providing a more precise estimate as opposed to a genetic distance. In this manner, the physical size of the Huntington’s disease region was estimated as 2.5 Mb (Bates 1991). This study also demonstrated the dissimilarity of genetic and physical distances, as two markers that were genetically separated by 3.5 cM were actually physically separated by less than 600 kb. In addition, two markers that were genetically separated by 1.2 cM were actually physically separated by almost 2 Mb.

A number of restriction enzymes preferentially cleave at CpG island sites, which are associated with the 5' ends of many genes, therefore these can provide hints about the density of genes in the region (section 1.3.4.2). When applied to Yeast Artificial Chromosome (YAC; section 1.3.3.3) cloned contigs, these restriction maps can accurately determine overlaps and identify rearrangements, deletions and chimerism that may be present. Chromosome breakpoints, deletions and translocations may also be detected by this analysis.

1.3.3.3 Yeast artificial chromosomes.

In the past, contig assembly was limited to small regions that rarely exceeded 200 kb and was restricted from extending further by the presence of large uncloned “gaps”. This was mainly due to the small cloning capacity of the available vectors such as lambda bacteriophage and cosmids. Extension from a contig is usually performed by “chromosome walking”. This is where terminal clone sequences are isolated to generate an STS or single copy probe to use in isolating new clones that stretch out further from the contig. This approach although successful was laborious and time consuming when using small clones, as the distance covered was limited by their sizes, so other methods were employed. A physical map of yeast chromosome V
(Olson et al. 1986) and most of the C. elegans genome (Coulson et al. 1986) was completed by detecting overlaps in lambda and cosmid clones using restriction digestion “fingerprinting”. Another method, chromosome jumping libraries help identify clones at some distance from the clone of interest, but do not supply sequences in between them (Poustka et al. 1987). These methods and vectors have been superseded by the yeast cloning strategy for physical mapping. Yeast artificial chromosomes (YACs, Burke et al. 1987) are capable of cloning fragments of up to 2 Mb of DNA and have allowed the construction of long range contigs by greatly reducing the number of cloned fragments which are required to obtain coverage of a particular genomic region. These exist as single copy clones and have been found to be generally stable when growing in the Saccharomyces cerevisiae host (Little et al. 1992; Green and Olson 1990a). YAC libraries are usually constructed by ligating partial or complete restriction digested total genomic DNA, or DNA from monochromosomal hybrid panels for chromosome specific libraries, into the YAC vector. The genomic DNA is usually size fractionated to a desirable size before cloning. These fragments are transformed into yeast spheroplasts and are maintained during the host’s growth and replication.

The YAC vector is constructed out of a combination of yeast-derived and pBR322-derived sequences (Schlessinger 1990). When restricted with Bam HI and Eco RI, it releases an unessential stuffer fragment and two “arms” between which the insert DNA is ligated. The standard features of YACs include inverted telomeric sequences (TEL) on both arms, an autonomously replicating sequence (ARS) and a centromeric sequence (CEN) for the maintenance of the clone in the yeast host cell. Selectable markers URA3 and TRP1 are also incorporated. Ligation into the cloning site interrupts the sup4 gene, a colour selection marker used for discriminating recombinants from non-recombinants (section 2.1.3). The most commonly employed YAC vector is pYAC4, which has been used in the construction of the most widely available human genomic YAC libraries (section 3.1.3). Modifications have been incorporated in some vectors to extend the use of YACs to other applications. For example, sequences have been incorporated to extend the use of YACs to other applications. For example, sequences have been incorporated to permit rescue of insert ends to generate end clones for chromosome walking. Some vectors have included T3 and T7 promoters to synthesise RNA probes from insert ends. Also selectable markers have
been included in some vectors to allow transfer into mammalian hosts.

Several problems are associated with this cloning system which can hinder physical mapping efforts and its use in further experiments. Firstly, the isolation of pure cloned YAC material is difficult as it co-exists in the host as an additional chromosome. One procedure is to isolate it from preparative PFGE separated samples, but the YAC must not co-migrate with the host's chromosomes. However the yield obtained is low as the YAC clone exists in the host as a single copy and there is almost always contamination from degraded yeast chromosomes. Another method is to subclone the entire yeast genome and identify the human specific clones by hybridisation, a time consuming task. Secondly, with the advantage of large insert sizes arises the problem of chimerism, where a clone contains fragments which are derived from different regions of the genome but are joined together. Chimeric clones are thought to derive from co-ligation of fragments during cloning, or by homologous recombination between two clones which have co-transformed (Green et al. 1991). Approximately 10% of YACs are believed to carry two independent clones because of co-transformation (Schlessinger 1990). In addition the instability of some clones can result in internal deletions. Unlike chimeric clones, these rearrangements are more dangerous as they are less likely to be detectable by STS content mapping or FISH.

Nevertheless, YACs have had an immense impact on physical mapping efforts. When YACs were later incorporated into the C. elegans cloning effort, the investigators found that YACs have the ability to clone regions of the genome that could not be retrieved from cosmid libraries (Coulson et al. 1986). In addition the S. cerevisiae host presents another advantage. The yeast homologous recombination activity can be exploited on overlapping YACs to produce a single YAC clone, which contains an entire gene and its regulatory elements, by recombination. This was accomplished for the cystic fibrosis disease causing gene, CFTR (Green and Olson 1990a).

More recently developed cloning systems for physical mapping include the P1 clones (Sternberg 1990), bacterial artificial chromosomes (BACs; Shizuya et al. 1992), and the P1 derived cloning system (PACs; Ioannou et al. 1994).
1.3.3.4 Physical maps of the human genome.

To accomplish the physical mapping objective of the HGP several groups have employed different approaches and mapping reagents to determine the order of STSs and orientation of clones. Chromosome 21q (Chumakov et al. 1992) was amongst the first human chromosomes to be completely mapped. The first whole genome physical mapping attempt was initiated by a CEPH/Généthon collaboration. They used a “clone-based” physical mapping strategy, which involves detecting clones that contain a specific STS by PCR screening, then assessing the presence or absence of additional STSs or common segments to define the overlap relations among the YACs. The CEPH YAC library was screened with >2,000 genetic markers which were distributed over ~90% of the genome. YACs from this library were chosen as their average insert size was very large (approximately 920 kb), making long range coverage easier. Several methods were employed to detect overlaps between clones, including THE and L1 repeat fingerprinting (Bellanné-Chantelot et al. 1992), cross hybridisation using Alu-PCR derived probes (Nelson et al. 1989) and STS content mapping. Overlaps were scored to assemble different levels of tiling paths which depict the structure of contigs. A number of YACs were also mapped by FISH. This “first generation physical map” was produced to help the scientific community construct detailed clone based maps of all human chromosomes. It has since been improved by CEPH/Généthon in collaboration with WI/MIT (Chumakov et al. 1995). This new map was supplemented with additional poly (CA), genetic markers (Gyapay et al. 1994), unpublished tetranucleotide markers and novel STSs which were isolated from end clones or random sequences. Overall, this final YAC map covered ~75% of the genome with CEPH genetic markers spaced at an average of 1.2 cM. It comprises 225 contigs with an average size of 10 Mb.

A second multilevel mapping approach was adopted and which is STS-based rather than clone based (Hudson et al. 1995) as it aims to avoid the problems of chimerism and instability that are associated with YAC clones (Green et al. 1991). Such a map should provide a foundation for isolating more manageable and stable clones for sequencing such as BACs, P1 clones and cosmids. In total, 15,086 STSs were used for
mapping, which included ~4,000 that had derived from ESTs (section 1.3.5), ~7,000 genetic markers mainly from Généthon and CHLC, and various random STSs. It differs in that it employed a radiation hybrid panel (Genebridge 4; Gyapay et al. 1996) for mapping loci to supplement the genetic map. In addition 10,850 STSs were used in isolating YAC clones. This information was combined to form an integrated radiation hybrid and genetic map that is anchored by STSs which are spaced at a density of one every ~200 kb. The estimated genome coverage is 99% on YAC maps and 94% on RH maps.

More recently a second STS-based map was constructed, using the higher resolution G3 Stanford panel (Stewart et al. 1997). It has not provided a higher distribution of STSs, as half of the 10,478 STSs located on this panel present an average spacing of only 500 kb. However, this project has produced 3063 novel STSs from human genomic DNA clones and ESTs which can be incorporated into existing maps. Only the genetic markers used for integration and less than 1,000 other STSs are common to both this and the Hudson map (1995). These integrated maps provide a global framework for increasing the STS resolution to 100 kb. This refinement is mainly being pursued by individual chromosome workshops (e.g. Bouffard et al. 1997a). Such maps will provide essential tools for mapping, positional cloning and sequencing.

1.3.4 Methods of gene identification.

Despite having achieved a framework of genetic and physical maps in the last several years, the rate limiting step in positionally cloning a gene remains at the stage of isolating and identifying transcribed sequences within a large targeted genomic interval. In positional cloning, a disease interval can usually be resolved by genetic linkage analysis and physical mapping to a region of 0.5-1 cM, which is approximately 1 Mb. Such an interval may contain 30-50 genes which are interspersed between a large amount of repetitive DNA. Many methods have been devised to isolate genes from specific regions and the main ones are discussed here. The choice of procedure for gene identification rests on the available resources in the laboratory and the extent of physical coverage across the disease interval. Often several methods are employed to increase the chance of identifying novel transcripts.
1.3.4.1 Cross-species sequence homology.

Coding sequences are much more strongly conserved during evolution than non-coding sequences, a fact which can be used to identify potential transcribed sequences by observing nucleotide conservation across species. This approach was applied in detecting the dystrophin gene for Duchenne muscular dystrophy (DMD; Monaco et al. 1986). Phage genomic clones which spanned a deletion in a patient with DMD were used as hybridisation probes on Hind III digested genomic blots from different mammalian species and from chicken ("zoo-blots"). One probe demonstrated homology between all these species. Northern blot analysis identified an RNA transcript only in muscle tissues. Screening a cDNA library with this probe isolated cDNAs that partially represented the Dystrophin transcript.

1.3.4.2 CpG island detection.

In the bulk of the genome CpG dinucleotides are highly methylated and occur at a frequency 80% lower than expected. The low abundance of these dinucleotides is thought to have arisen by deamination of methylated cytosine to thymine. However, a small fraction of CpG dinucleotides in the genome differ because they are non-methylated and occur in small clusters at the expected frequency. These form a G+C rich cluster known as CpG islands. CpG islands are associated with the 5' region of house-keeping genes and approximately 40% of tissue specific genes. These therefore make useful landmarks for identifying novel genes (Larson et al. 1992). The islands range from 200 to 1,400 bp and almost always cover the whole or part of the promoter regions and may extend into one or more exon. It is estimated that about 30,000 such islands exist in the haploid genome (Bird 1987). Craig and Bickmore (1994) used FISH to estimate the density of CpG islands in human chromosomes. Approximately half of all CpG islands tended to be clustered at a high density in T-bands (a subset of R-bands). In these regions, inter-island distances have been estimated at 15-500 kb.

Methyl sensitive enzymes such as Sac II, BssH II and Eag I preferentially cleave at islands sites as they have two CpGs in their recognition sequence (Lindsay and Bird 1987). Islands can be detected by hybridising probes to total genomic or cloned DNA which is digested with each of these enzymes. As the average CpG island has at least
one site for each of these enzymes, similar sized fragments are more likely to be produced when an island is present. However at the resolution achieved by PFGE, it will not be possible to distinguish if two islands are separated by a distance smaller than 50 kb. Identification of the transcripts that are associated with these sites will require cloning of these fragments for finer analysis.

Other methods have been developed using these landmarks to identify novel genes. A CpG island library was constructed from human genomic DNA (Cross et al. 1994). The method made use of the rat chromosomal protein MeCP2, which binds to DNA that is methylated at CpG, to separate non-methylated islands from methylated dinucleotides in the genome. The final library consisted of isolated CpG islands with an average insert size of 760 bp. Associated genes could be isolated by screening full length cDNA libraries. A second method, Island rescue PCR (Valdes et al. 1994), involves digesting YAC or cosmid DNA with Sac II, Bss HII or Eag I and then ligating the fragments to vectorette linkers. Primers are used in PCR to amplify the region between the linker and the nearest Alu site. This selects for the CpG sequences at the 5' ends of many genes. The resultant products are separated alongside Alu-PCR only products. Novel fragments are isolated and used to screen full length cDNA libraries. The authors demonstrated its success by identifying the neurofibromatosis type 1 (NF1) cDNA from a NF1 containing YAC. This approach is limited as it requires the presence of an Alu repeat at suitable distance away from the restriction site to allow for PCR amplification. CpG island detection methods have been widely employed but require time consuming screening of full length libraries and do not identify genes that are not associated with islands.

1.3.4.3 Direct screening.

This method involves screening an arrayed cDNA library directly using a large genomic clone as a probe. Elvin and co-workers demonstrated the success of this technique by using a 180 kb gel purified YAC clone containing the aldose reductase gene as a probe to isolate its respective cDNA from an arrayed fetal liver library (Elvin et al. 1990). The main problem with using such a large probe is the increased complexity. Coding sequences represent only a small percentage of the clone and there is a large fraction of repeat sequences. These authors first quenched the probe with
human placental DNA and pBR322 to suppress repeat and vector hybridisation. In addition to detecting aldose reductase cDNAs, they isolated several other cDNAs which were encoded by genes in the clone. This approach was also employed to identify part of the Neurofibromatosis (NF1) gene (Wallace et al. 1990). A YAC containing DNA from the breakpoint region in a patient with NF1, was used as a probe to screen a B lymphoblast cDNA library. This identified a 0.8 kb transcript which later proved to be part of the gene that is implicated in NF1.

The technique requires complete blocking of repetitive and vector sequences, and careful screening of positive clones. The probes used are at the limits of their sensitivity in detection of cDNAs. The YAC probe only identified 10% of the aldose reductase clones that were present in the cDNA library. Larger genomic clones may be less efficient in detection and smaller length cDNAs are likely to go undetected.

1.3.4.4 cDNA hybrid selection.

In any given cell type, approximately 10,000 gene types may be expressed. These can produce mRNA transcripts at varying levels from 1-10 molecules per cell and up to 200,000 (Bishop et al. 1974). Identifying a gene by direct selection therefore requires screening hundreds of thousands of cDNAs, especially when the gene is a rare transcript or is expressed in a complex tissue. In addition, when the genomic target of interest is large, there is a very poor signal to noise ratio, mostly due to the number of repetitive elements within the genomic source and the cDNAs. To reduce this complexity in screening, normalised libraries can be used (Pantanjali et al. 1991). Alternatively a cDNA hybrid selection approach can be employed which specifically selects and enriches for target cDNAs from large genomic clones, unlike direct screening which simply identifies them. This procedure is also referred to as “cDNA selection” or “direct selection” and was originally described simultaneously by two groups (Lovett et al. 1991; Parimoo et al. 1991).

In summary, the genomic clone of choice is digested and immobilised onto a filter. If using YAC DNA, it must first be purified away from the yeast host. The cDNA library is amplified using vector specific primers. This is hybridised to the immobilised genomic DNA in solution. Non-specific cDNAs are washed off and
specific ones are eluted and then re-amplified. These undergo a second round of
selection for specificity and the final products are cloned for further analysis. Protocols
described by both groups used total human genomic DNA and total yeast DNA as
quenching reagents. The protocols differed only in that the cDNA source was blocked
for repeats and was also competed with pBR322 DNA for vector contamination in
Lovett’s method; while in Parimoo’s version, the genomic source was quenched and
also competed with ribosomal specific clones. Lovett and colleagues applied the
selection procedure to a 550 kb YAC containing the erythropoietin (EPO) gene.
Quantitative assessment was made by hybridising specific probes onto arrayed cDNA
clones from the starting and the selected library. This resulted in a ~1,000 fold
enrichment of the EPO cDNA. Similarly, Parimoo and co-workers achieved a ~7,000
fold enrichment after two rounds of selection of a rare species cDNA from a 320 kb
YAC. Although successful, they observed problems of selection artefacts resulting
from contamination of the library, and noted the importance of sufficient quenching of
repeats. They also suggested the use of further rounds of enrichment to increase
specificity. Many refinements have since been made and novel approaches based on
this technique have emerged. The gene involved in X-linked agammaglobulinaemia
was identified by this procedure. A 640 kb YAC clone was used in one round of
selection with cDNAs from a Burkitt’s lymphoma cell line. Of 1,536 selected cDNAs,
only 104 mapped back to the target genomic source. Two of the cDNAs detected
restriction fragment alterations in XLA patients. These were used as probes to isolate
full length transcripts for further evaluation (Vetrie et al. 1993).

The use of biotin-streptavidin capture systems was introduced in this selection
process to provide more efficient selection matrices (see figure 1.3). This is where
genomic DNA is biotinylated either by photobiotinylation, nick translation or by
amplification using a biotinylated primer. This target DNA is then hybridised in
solution with amplified cDNAs at a ratio of approximately 10:1 cDNAs: genomic
template, to provide more opportunity for producing cDNA-genomic hybrids at the
capture stage. The hybridised genomic-cDNA products are captured on streptavidin
coated magnetic beads. Non-specific products are washed off and selected cDNAs are
eluted then amplified using the cDNA vector/linker specific primer. Korn and
1. Digest and ligate linkers to DNA from pools of cosmids or pulsed-field gel purified YAC genomic DNA.

2. PCR amplify the genomic segments using biotinylated primers.

3. PCR amplify the cDNA library source using vector/ linker primers.

4. Suppress repeats then hybridise the genomic source to an excess of cDNA amplified inserts at high stringency.

5. Capture the genomic-cDNA complexes using streptavidin-coated magnetic beads. Wash off unbound non-specific cDNAs.

6. Elute specific cDNAs by boiling and amplify these with the vector specific primers.

7. These can either be used as the cDNA source in a second round of selection to increase specificity, or

8. The selected cDNAs can be sub-cloned for further analysis.

The high enrichment levels achieved by this method is a result of PCR amplification on the selected material.
co-workers (1992) demonstrated an 80,000 fold enrichment of a particular cDNA from a cosmid with two variants of this system. Also, 66% of cDNAs from a selected library mapped back to their genomic targets when using DNA from two cosmid contigs spanning 900 kb.

As well as employing the biotin-streptavidin capture system, Morgan and colleagues (1992) used uncloned newly synthesised cDNAs. This was to retain the complexity of the mRNA population, which can be lost when cloning, and to reduce the possibility of vector contamination. These cDNAs were digested with four base cutters to divide them into smaller fragments and so avoid PCR bias against larger cDNAs. They were then amplified by addition of linkers. The cDNA source consisted of pooled samples of cDNAs from various developmental stages to provide the maximum chance of selecting rare species. The quenching material was Cot-I DNA, which is highly enriched in intermediate repeats. Total genomic DNA was not used as it may increase the overall sequence complexity of the hybridisation mixture. A number of novel transcripts were isolated from a 425 kb YAC, as well as the interleukin 4 and interleukin 5 control cDNAs. This gave enrichment levels up to 100,000 fold. Even though many studies have shown that selection is more efficient when using smaller insert cosmids rather than YACs, this procedure has been applied to select cDNAs using total yeast containing YACs and to whole chromosomes (Parimoo et al. 1993; Rouquier et al. 1995). Another approach, End Ligation Coincident Sequence Cloning (EL-CSC) shares similarity with the above techniques, but is more specific in that it requires an exact sequence and length match to the genomic sample, and has shown to give enrichments levels between $10^6$ and $10^7$ (Brookes et al. 1994).

The main disadvantage of the cDNA hybrid selection technique is that as short fragment cDNAs constitute the selected material, a cDNA library has to be screened to isolate a full length version. In addition, the presence of contaminants such as repeat containing or vector containing cDNAs in the selected material is a recognised problem, especially when suppression is incomplete. Some cDNAs may not be selected for as short exons can fail to hybridise. As well as selecting the target cDNAs, it will also isolate homologous gene families and pseudogenes. Nevertheless, this approach has presented well documented successes in enrichment over large genomic
Introduction

areas. In addition only a small amount of starting material is required and the method has proved to be able to enrich low abundance transcripts with an associated decrease in non-specific ubiquitous transcripts (Tagle et al. 1993).

1.3.4.5 Exon trapping.

The principal limitation of direct screening and cDNA hybrid selection is the need for the target cDNA to be represented in the library screened. When looking for novel genes encoded by large genomic regions, information about which tissue and developmental stage they are expressed in is unknown and many genes are likely to be missed because of this. The exon trapping strategy is an elaborate technique which has been developed to identify transcripts from a genomic region by recognising cis-acting signal sequences in genes that are required for splicing (splice donor, SD; splice acceptor, SA). These signals can be exploited to identify exons in a genomic source as it has been observed that chimeric introns with SD and SA sites from diverse genes can be actively spliced when transiently expressed in mammalian cells. In brief, the genomic DNA of interest, usually in plasmids and cosmids, is shot gun cloned into a specialised splicing vector. This vector has certain splicing signals flanking an intron within which the cloning site resides. The library of cloned DNA is transfected into mammalian cells, which support vector transcription and allow splicing to occur between the vector and insert sequences. If the genomic DNA is cloned in the correct orientation and has at least one exon, a splicing event can occur between the vector SD and genomic SA. This RNA is harvested, converted to cDNA and further analysed to determine whether it contains true exons.

The strategy was originally developed by Duyk and colleagues (1990). The exon trapping cloning vector in this system has just one SD site which can pair with genomic SA signals that are cloned further downstream. Only one SA site is required so the exon to the 3' end of this will be isolated. The resulting product consists of a chimeric cDNA with vector and insert exons. This system was first tested on defined exons and was successful on those cloned in the correct orientation. However when tested on a mixture of shot gun cloned genomic fragments from a HLA-A2 containing clone, only a small percentage of the trapped clones had actually resulted from true splicing of introns. The remaining trapped clones had resulted either from mutations or rearrangements in the marker gene, or from cryptic splicing events, where a splice site
joins with a sequence that resembles a splice site.

The exon amplification scheme of Buckler et al. (1991) was designed to be more specific. The splicing vector has both SD and SA sites in the vector. For this process to be successful, it requires the cloning of an entire exon flanked by intronic sequences on both sides, so they have both SA and SD sites that can pair with vector signals. Initial tests using this system identified novel exons from 15-20 kb of genomic DNA which all mapped back to the target clones. This procedure and the splicing vector have been modified to selectively eliminate vector-vector and cryptic splice products (Church et al. 1994) thereby increasing exon recovery.

The success of this technique is apparent, as a number of disease genes have been identified using this approach. This includes the Huntington's disease gene (the Huntington's Disease Collaborative Research Group, 1993); the copper transporter gene defective in individuals with Menkes disease (Vulpe et al. 1993), and the neurofibromatosis type 2 tumour suppressor gene (Trofatter et al. 1993).

The main drawbacks of these original strategies is that due to the small cloning capacity of the splice vectors (1-4 kb), multiple exons would rarely be present in one insert. Instead a multitude of single exons would have to be aligned to reconstruct their original order, or the exons could be used as probes to screen a cDNA library. This somewhat negates its advantage as being a tissue independent system. Also single exons are often too short to use as probes for screening a cDNA library or Northern blot. To overcome some of these difficulties new trapping procedures have been created. 3'-terminal exon trapping generates sequences from the last exon of a gene and its untranslated regions which can provide more sequence than all other exons (Krizman and Berget 1993). Datson and colleagues developed a new cosmid-based exon trapping vector, sCOGH1 that can take DNA fragments of 25-40 kb (Datson et al. 1996). Transcription is driven via the strong ubiquitously expressed mouse metallothionein-1 promoter. The trapping procedure is very similar to established methods. When tested on large fragments containing parts of the Duchenne muscular dystrophy gene, correct splicing was observed in each case. The large insert capacity of this vector enables its use in generating physical maps as well as for exon trapping, which can yield processed transcripts with multiple exons that are spliced in the correct order.
1.3.5 Positional candidates and expressed sequence tags.

The "positional candidate" approach combines genetic linkage information with knowledge of the map position of genes. Genes can automatically become candidates for a disorder if they are located within the critical interval for a disease. Those that possess functional features that relate them with the pathophysiology of the disease become favoured candidates and are screened first for mutations in affected individuals. Functional information is advantageous, but is not always required and no gene should be overlooked. The identification of such positional candidates is now being reduced to searching computer databases where lists of genes that map to specific regions can be obtained. Such an approach has become feasible and more popular with the construction of a "transcript map" of the human genome.

At the initial discussions concerning the Human Genome Project, Professor Sidney Brenner proposed that large scale sequencing of cDNAs should take precedence over genomic sequencing. This idea was taken up in the form of "Expressed sequence Tags" (ESTs). These are partial sequences generated from random cloned cDNAs. Transcripts from the 5' end or random areas can provide information on the protein coding region. Those derived from the 3' region are preferable for selecting STSs to be used in mapping as they are less conserved between species and less likely to be interrupted by an intron (Wilcox et al. 1991).

The first approach to a full scale generation of ESTs was by Venter and colleagues who sequenced over 8,000 cDNA clones from human brain libraries over several years (Adams et al. 1991, 1992a, 1993a and 1993b). They also generated 174,472 new ESTs from 300 libraries constructed from a range of human tissues at various stages of development (Adams et al. 1995). Their preliminary studies were aimed at sequencing random primed or 5' regions from directional libraries to maximise the number of ESTs that contain protein coding sequences. A further 18,698 ESTs were created by the Genexpress Centre (Généthon; Houlgatte et al. 1995) who assigned 2,733 of these to specific chromosomes by PCR on human-rodent somatic cell hybrids. The Washington-University-Merck EST project in collaboration with the IMAGE consortium (Integrated Molecular Analysis of Genomes and Their
Expression; Lennon et al. 1996) was the third large scale attempt at EST production. More than 300,000 ESTs derived from both 5' and 3' ends of cDNAs were generated, mostly from normalised libraries from a range of tissues (Hillier et al. 1996).

In these studies, sequences were analysed to produce an “electronic expression profile” of the mRNA distribution. The nucleotide and amino acid transcripts of these ESTs were compared to Genbank entries from a broad array of organism and tissue sequences via BLAST similarity searches (Altschul et al. 1990). Most previously characterised genes were represented by ESTs. Many ESTs were tentatively identified as new members of gene families or human homologues of genes from other organisms as they were similar but not identical to known genes. A large percentage of ESTs were novel and were subsequently analysed via the codon predicting program GRAIL to estimate the probability that the sequence encoded a protein. The expression levels of certain transcripts could also be determined by observing the measure of EST redundancy as mRNA copy numbers are reflected in the composition of cDNA libraries, unless normalised or subtracted libraries were used. All these efforts provided an estimate of the diversity and activity of expressed sequences in different tissues by cataloguing them in respect to their assumed roles in cellular biology. Most accumulated sequences are gathered in the dbEST division of the Genbank database (Boguski et al. 1993) and can be accessed publicly. Although these studies have provided an important insight into the expression profile of tissues at various developmental stages, only a limited number of these ESTs were mapped to chromosomal regions.

Many small-scale projects have mapped a small number of ESTs through hybrid panels or onto genomic clones (e.g. Polymeropoulos et al. 1993; Berry et al. 1995). Hudson and colleagues reported the results of an international effort, sponsored by the Human Genome Organisation (HUGO), which assigned >16,000 ESTs on radiation hybrid panels and ~1,000 onto YACs (Schuler et al. 1996). These were placed relative to a common framework map which is connected by ~1,000 Génétion polymorphic markers. This placed ESTs to at least the same resolution as the critical intervals of diseases (0.5-5 Mb). As well as providing positional candidates that are readily accessible by database searching, it has provided a large source of STSs to increase the density and reinforce clone continuity on the physical map.
Once ESTs or genes that reside in a region of interest are identified, redundant clones can be used for assembling contigs of overlapping sequences to yield the full length gene transcript. Several groups have reported the clustering of deposited ESTs and gene sequences and formed a non-redundant set of unique sequences as a standard for comparison for each transcript. This includes the TIGR Human cDNA collection and Tentative Human Consensus Sequences (THCs; Adams et al. 1995), the Unigene set (Schuler et al. 1996) and the Merck Gene Index (Aaronson et al. 1996).

The positional candidate gene approach has become more successful in the last two years with increasing information on the locations of diseases and genes. Many disease causing genes have been identified this way. For example, by searching the human EST databases using sequences from known members of the human ATP-binding cassette (ABC) family of genes, some of which are known to be involved in disease (e.g. CFTR with cystic fibrosis), Allikmets and colleagues (1996) identified 21 new members. Their map locations and patterns of expression were determined to serve as positional candidates for diseases mapping to those regions. One of the EST-associated genes known to map to 1p22-p21 was soon found to be pathologically implicated in patients with Stargardt’s macular dystrophy and age related macular dystrophy which also map to this locus (Allikments et al. 1997). In another example, the retinoschisis disease gene had been linked to Xp22.3-p22.1. Sauer et al. (1997) identified mutations in a novel gene which had been identified as a mapped EST in the retinoschisis critical region and which was expressed exclusively in the retina.

1.3.6 Sequencing the human genome.

The ultimate aim for the Human Genome Project is to determine the complete human DNA sequence by the year 2005. This calls for improvement in sequencing technology, to provide a rapid and cost-effective means to complete this task. An organised strategy must also be implemented to decide on the best approach for handling the immense amount of data generated. The main problem at this time is primarily the lack of contigs consisting of smaller stable clones such as BACs, PACs or cosmids in the human physical map. These are the preferred substrates for sequencing since YAC clones are too large and possess structural instabilities (Boguski et al. 1996). A second problem is that unlike lower eukaryotes, genes in the
human genome are interspersed with large quantities of "junk" repetitive DNA, hence gene discrimination is difficult. Although groups are concentrating on EST generation, a large number of genes have yet to be identified in this way. These will be rare transcripts or genes that are expressed only at certain times. As sequence data is generated, gene identification can be performed using analytical software. New data can be compared to sequences that are already deposited in databases to identify any significant homologies using sequence similarity search programs (e.g. BLAST, Altschul et al. 1990 and FASTA, Pearson and Lipman 1988). Several gene-finding computational programmes are available which can recognise patterns characteristic of coding sequences, such as the Gene Recognition and Analysis Internet Link (GRAIL; Uberbacher and Mural 1991). However all gene identification programmes that are available have limitations and there is a need for their improvement. Nevertheless this method of gene identification is cheaper and faster than other experimental strategies once the sequence is known, and will probably be favoured over other techniques as more sequence is generated.

An example of such an approach was recently published by Everett et al. (1997). In the search for the mutation causing Pendred syndrome, a common form of syndromic deafness, the investigators linked the disease to 7q22-31. A YAC based contig, then a more refined BAC-based contig, was constructed across the disease interval. Seven genes and ESTs that mapped to the region were screened for mutations that segregated in affected individuals only, but none were found. Several BAC clones were sequenced by the Washington University Genome Sequencing Centre. Analysis of these using the GRAIL exon predicting program identified the presence of several putative exons for which the full length cDNA was then isolated. This PDS gene sequence demonstrated homology to the DRA gene involved in congenital chloride diarrhoea and mutations were then identified in Pendred patients.

Complete sequence maps have been attained for several bacteria, viruses and insects. Also there has been considerable progress in completing the 12.5 Mb yeast Saccharomyces cerevisiae (Dujon 1996) and 100 Mb Caenorhabditis elegans genomes. Automated fluorescent sequencers are now in wide use, but significant developments must still be made to increase the speed of sequencing. A number of sequencing centres have been formed which are committed to map and sequence
specific chromosomal regions. DNA sequences derived by these projects are stored in databases freely accessible via the internet, mainly Genbank (at the National Centre for Biotechnology information), or the Genome Data Base (at the National Centre for Genome Resources). Reading the sequence of the human genome will unravel valuable information about development, biological functions, disease predisposition, inheritance and the evolution of the human body.

1.4 Structure of the retina.

Among a wide variety of eye disorders, retinal dystrophies constitute a major cause of blindness. Retinitis pigmentosa is a prevalent group of progressive retinopathies and is the subject of the positional cloning project in this study. Before describing the symptoms and molecular pathology of the disease it is necessary to provide some insight into the normal structure and function of the visual process.

The retina is the neural sensory layer of the eye that is involved in image formation. It consists structurally and functionally of two distinct layers: the inner nervous tissue layer and the outer non-neural pigmented epithelium (RPE) layer. Photons of incident light are absorbed by the visual pigment molecules in photoreceptor cells of the neural layer. These convert the photon stimulus into a neural signal, which is amplified via the phototransduction pathway, for signalling responses to the brain.

1.4.1 The neural retina.

The vertebrate neural layer has six major neuronal cell types, which form a connecting sequence that carries information via the optic nerve fibres leading to the brain (figure 1.4). These are photoreceptors (rods and cones; section 1.4.2), horizontal, bipolar, amacrine, interplexiform and ganglion cells. They are sub-grouped into distinct layers of cell bodies (the outer nuclear layer, inner nuclear layer and inner plexiform layers) and synaptic connections (outer and inner plexiform layers).
**Figure 1.4** The cellular organisation of the vertebrate retina. A) A histological cross-section through a human retina. B) A schematic diagram of the human retina, showing most of the cell types and layers. (Adapted from Berman, 1991)

- r: rod photoreceptor cells
- c: cone photoreceptor cells
- b: bipolar cells
- h: horizontal cells
- i: interplexiform cells
- a: amacrine cells
- g: ganglion cells
- m: Muller cells
The main path for the flow of information from visual light responses is through the photoreceptor cells, to the bipolar cells, then the ganglion cells and via the optic nerve fibres to the brain. Photoreceptor axons synapse with either the bipolar or horizontal cells in the outer plexiform layer. Cone bipolar cells synapse either directly with ganglion cells or indirectly through amacrine cells. Rod bipolar cells terminate deep in the inner plexiform layer and always synapse with intermediary amacrine cells. The horizontal and amacrine cells modify and control the impulses and form laterally directed pathways between neurones. Interplexiform cells receive input from other amacrine cells and function to transmit signals back to the distal retinal layers, forming a feedback pathway. The last neuronal cell type in the retinal layer, the ganglion cells, receive their input from amacrine and bipolar cells and have extensively spread dendrites so each ganglion cell may be influenced by the activity of a large number of rods and cones. The axons of the ganglion cells join to form the optic nerve, which relays the signal to the brain. Müller cells are glial cells whose fibres extend and span through all the retinal layers to provide mechanical support for the retina.

1.4.2 The photoreceptor cells.

Vertebrates possess two kinds of photoreceptor cells, the rods and cones. These can be distinguished by their morphology, absorption spectra and anatomical distributions (reviewed in Nathans et al. 1992). These are highly specialised cells with a sensory organelle, the outer segment, which contains the photopigments for light absorption. Cones function in bright light and are responsible for colour (photopic) vision and rods function in dim light and are responsible for night (scotopic) vision but do not perceive colour. Rod photoreceptors are more abundant with a ratio to cones of about 20:1. The distribution of rod and cone cells in the retina is not uniform. The fovea centralis contains no rods but is densely packed with ~200,000 cones, each connecting with at least as many optic nerve axons. The density of cones in the fovea is approximately 150,000 cones per mm². In the periphery, 10,000 rods may be connected in clusters to a single nerve fibre. The density of cones in this region reduces to 4,500 cones per mm². In contrast the density of rods is highest in the mid-periphery and decreases when approaching the centre.
Morphologically, rods are thin and cylindrical whereas cones have a conical shaped outer segment (figure 1.5). The outer segments accommodate a dense stack of disc-like membranes which contain the visual pigment (opsin) molecules. In rods, the discs are closed, flattened sacs which are formed by successive basal invagination of the outer segment plasma membrane. Cone discs are different in that they form a continuous serrated membrane with the plasma membrane of cone outer segment. Rod discs are continually shed from the rod apex after moving the entire length of the outer segment and are phagocytosed by the microvilli of the RPE cells. Cone discs are also shed but the cone cells remodel themselves after each event.

The visual pigment opsins are embedded in the membrane of these outer segment discs and comprise 80-90% of the total protein in the discs. These have a light sensitive chromophore, 11-cis-retinal, which is bound to the opsins. The rod visual pigment is rhodopsin. When attached to the chromophore, it has a peak absorption (λ_max) of ~500 nm and is responsible for contrast sensitivity and vision in dim light. There are three types of cone photoreceptors each containing a specific photopigment with absorption maxima in the red (560nm; long wavelength), green (530nm; medium wavelength) or blue (426nm; short wavelengths; Nathans et al. 1986).

The inner segment of photoreceptors is joined to the outer segment by a slim immotile cilium. This functions to transmit cellular components from the inner segment to the discs. The inner segment is rich in mitochondria, ribosomes and golgi complexes providing the metabolic machinery necessary for active protein synthesis. The inner fibre terminates in a synaptic ending that forms a synapse with a bipolar neurone.
**Figure 1.5** Schematic representation of mammalian rod and cone photoreceptor cells. The light sensitive photopigments are located in the stacked discs of the outer segments. (Adapted from Ali and Klyne 1985).
1.4.3 Retinal pigment epithelium.

The RPE is comprised of a single layer of cuboid cells lying between the Bruch’s membrane of the choroid and the photoreceptors (figure 1.4). The apical cell surface faces the photoreceptors and interacts by their villous processes with the outer segments. The layer of cells are tightly compacted to provide a protective barrier between the choroidal circulation and the neural retina. Other functions include phagocytosis of photoreceptor outer segments, secretion of interphotoreceptor cell matrix (IPM) and transporting nutrients to the neural retina. The RPE also contains light-absorbing pigment granules, which control light scatter for precise image formation.

1.4.4 The rod phototransduction pathway.

1.4.4.1 Excitation.

The photochemical activation of the visual pigment triggers an amplified cellular response converting quanta of light energy within a particular range of wavelengths into electrochemical changes. This affects membrane conductance and ultimately generates a nerve impulse directed to the brain where interpretation and perception occurs (reviewed in Yau 1994).

1.4.4.1.1 Photoactivation.

The primary event, photoactivation, is the light triggered isomerisation of the 11-cis-retinal of rhodopsin to its all-trans isomer. This alters the geometry of retinal, which results in the release of the all-trans isomer from the opsin. The opsin changes shape to photoactivated rhodopsin (metarhodopsin II; R*) exposing a site on its cytoplasmic side which allows GDP-associated transducin to bind (see figure 1.6).

1.4.4.1.2 Transducin activation.

The G protein transducin is a peripheral membrane protein of the rod outer segment and is composed of three structural subunits TαGDP, Tβ and Tγ. Upon
Figure 1.6  The key events of the phototransduction cascade in mammalian 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
CaM: calmodulin

GC: guanylate cyclase
RK: rhodopsin kinase
PDE: cGMP phosphodiesterase
\( \text{PDE}^* \): cGMP phosphodiesterase (\* active)
GCAP: guanylate cyclase activating protein
Arr: arrestin
photoexcitation, rhodopsin activates transducin by forming a complex with it and catalyses the exchange of GDP bound to the $\alpha$ subunit for GTP. The active form of transducin, the $\alpha$ subunit charged with GTP, dissociates from the $\beta\gamma$ subunits and rhodopsin and is released into the cytoplasm. $R^*$ is free again to bind to a new transducin molecule. The signal is amplified all along the enzymatic cascade to produce a measurable effect. One rhodopsin molecule can activate more than five hundred transducin molecules until it is inhibited by the combined action of rhodopsin kinase and arrestin (see section 1.4.4.2). In the cytoplasm the GTP charged $\alpha$ subunit activates the enzyme cGMP phosphodiesterase (PDE).

1.4.4.1.3 Phosphodiesterase and cGMP-gated cation channels.

PDE is composed of $\alpha$, $\beta$ and two $\gamma$ subunits. $\alpha$GTP molecules bind with the two inhibitory $\gamma$ subunits, thus activating one PDE$\alpha\beta$ molecule. This hydrolyses $3',5'$-cGMP to $5'$ GMP at a very high rate. Each molecule of PDE that is activated by transducin is capable of hydrolysing $\sim$1,000 cGMP molecules. At least three cGMP molecules are bound to the rod photoreceptor cGMP-gated channel protein, a transmembrane cation-specific channel that in the dark allows $\text{Na}^+$ and $\text{Ca}^{2+}$ to enter the outer segment across the plasma membrane. The cGMP-gated channel protein is composed of two subunits. The $\alpha$ subunit is functional by itself, but is modulated by the $\beta$ subunit. An $\gamma$ subunit has also been reported which forms a complex with the $\beta$ subunit, but its function is unknown. These are located in the rod outer segment and are controlled by cGMP concentrations. Upon hydrolysis, the cytosolic concentration of cGMP decreases closing the cGMP-dependant cation channels and preventing the influx of sodium and calcium ions. This subsequent hyperpolarisation of the rod cell plasma membrane generates a nerve signal by reducing the level of neurotransmitter release from the rod synapse.

1.4.4.2 Recovery and restoration of the dark state.

This requires isomerisation of all-trans rhodopsin to the 11-cis form so transducin can no longer be activated. Rhodopsin kinase is a cytosolic protein in the rod photoreceptors which initiates the deactivation of the phototransduction cascade by
phosphorylating the serine and threonine residues in the C-terminus of photoactivated rhodopsin (R*). Phosphorylated rhodopsin has a decreased ability to activate transducin and an enhanced ability to bind to the 48 kD protein arrestin. Arrestin binds specifically to photoexcited-phosphorylated rhodopsin under dark conditions and prevents further Tα-GTP molecules to bind.

Phosphodiesterase is deactivated by hydrolysis of the GTP bound to Tα by this subunit’s intrinsic GTPase activity. The Tα subunit dissociates with the PDEγ subunits, which then bind to PDEαβ, blocking its activity. Tα recombines with Tβγ and reassociates with the disc membrane. The protein phosducin may participate in the modulation of signal transduction by inhibiting the GTPase activity of Tα. At the final step, rhodopsin is regenerated by recombining with 11-cis-retinal and its bound phosphate is removed by a phosphatase.

As amplification of signals occurs at every stage, feedback and adaptation mechanisms are employed to modulate the cascade. The restoration of the dark state also requires recovery of the electrochemical environment and involves the synthesis of cGMP. Calcium plays an important role in photorecovery (Polans et al. 1996). In the dark adapted state, extracellular Ca\(^{2+}\) entry through the open cGMP-gated cation channels is balanced by their efflux through Na\(^{+}/Ca\(^{2+}\) exchangers in the plasma membrane of the rod outer segment. Upon light exposure and closure of the channels, the export of Ca\(^{2+}\) continues, resulting in a drop of intracellular Ca\(^{2+}\) from 500nM to 50nM. This induces the activation of guanylate cyclase (GC), a peripherally bound membrane protein that catalyses the resynthesis of cGMP from GTP thereby speeding the restoration of the dark level of cGMP in the cell and permits reopening of the cGMP-gated channels. GC activity is mediated by the guanylate cyclase activating protein (GCAP) in response to low Ca\(^{2+}\) levels.

Two other Ca\(^{2+}\) binding proteins are recoverin and calmodulin. Recoverin is present in both outer and inner segments. This is believed to affect recovery by inhibiting phosphorylation of activated rhodopsin at high Ca\(^{2+}\) levels by binding specifically to rhodopsin kinase, thereby prolonging photoresponse. Finally, calmodulin is a ubiquitously expressed protein and is also found in the rod outer segment. This is thought to be somehow involved in modulating the cGMP-gated channel.
A similar sequence of events is thought to take place in cone phototransduction. The only differences are that the key rod specific proteins have cone specific counterparts. Phototransduction is a complex biochemical pathway and more details regarding its regulation are still surfacing.

1.5  Retinitis Pigmentosa.

1.5.1 Clinical description.

Retinitis Pigmentosa (RP) describes a group of inherited retinal disorders characterised by progressive loss of peripheral vision and night blindness (nyctalopia) due to rod photoreceptor degeneration. This results in a ring scotoma (blind or partially blind area) in the mid periphery which eventually encroaches centrally leading to cone degeneration and often to complete blindness in later life. RP is the most common form of hereditary eye disorder affecting almost 1 in every 4,000 births in the Western world (Boughman et al. 1980).

Clinically, middle aged patients with RP have fundus changes that includes a pale optic nerve head, attenuated retinal vessels and intraretinal pigmentation (see figure 1.7). The latter results in the visual appearance of clumps and strands of black pigments which are most prominent in the mid periphery of the fundus giving a pattern that is referred to as bone spicule pigmentation. This is a consequence of atrophy of the retinal pigment epithelial layer (RPE) resulting in pigment deposition in the retinal vascular layer. This phenotype is also seen in non-RP patients who have had previous injury to the RPE layer or have been afflicted with certain non-inherited disorders. These “pseudo-RP” patients can be confused with RP patients who have inherited the disease in a Mendelian fashion. However this pigment deposition is a secondary effect of the pathological process and it is not known if it contributes to the deterioration of retinal function.

Diagnosis of RP is performed by evaluation of the visual fields to determine the amount of peripheral vision loss, and also by electroretinographic testing. This is where electrical responses of the retina are measured after stimulation by a flash of light. In affected patients the response is reduced in amplitude and delayed in response time, and will deteriorate further over time. Patients also have impaired or absent
Figure 1.7

a) A fundus photograph of a normal retina.
b) A fundus of a patient with the advanced stages of retinitis pigmentosa.
adaptation in the dark or dim light (nightblindness) and can be measured by dark adaptation threshold testing.

1.5.2 Mode of RP inheritance.

RP is extremely heterogeneous both clinically and genetically. Non-syndromic cases represent at least 22 distinct genes, 11 of which have been identified to date and there are 11 putative loci. In addition RP has been associated with other diseases such as Usher syndrome, Bardet-Biedl syndrome and Kearns-Sayre syndrome (Sullivan and Daiger 1996; RetNet: http://utsph.sph.uth.tmc.edu/www/utsph/RetNet/disease.htm). Approximately 50% of non-syndromic RP cases are inherited in either an autosomal dominant (adRP), autosomal recessive (arRP) or X-linked forms (xLRP). The remainder are isolated or sporadic cases which may have resulted either from new dominant mutations; are inherited but not identified as a result of variable penetrance; or are recessive, X-linked or digenic cases where the family history is unknown.

1.5.3 Autosomal dominant RP (adRP).

adRP accounts for approximately 25% of total retinitis pigmentosa in the UK (Jay 1982). Prior to molecular classifications, Massof and Finkelstein (1981) and Lyness and colleagues (1985) classified adRP cases based on the fundus appearance and pattern of photoreceptor degeneration. Type-1 or D-type adRP is more severe with diffuse loss of rod function which in the early stages coexists with relatively normal cone function. Type-2 or R-type RP, is more common and displays regional/patchy loss of rod photoreceptor function that is accompanied by loss of cone function. A rarer class known as sector RP is characterised by retinal atrophy that occurs only in the lower nasal quadrant of the fundus, and is present in all affected members of the family. Patients with 1/D-type disease usually exhibit night blindness before 10 years, while pigmentary changes and visual difficulties may not be detectable until 10-20 years after the onset of nightblindness. Type 2/R patients have a variable age of onset. Some type 2/R adRP pedigrees show a wide variation of severity, and the inheritance of the disease can appear to skip generations. This is referred to as variable or incomplete penetrance.
1.5.4 Molecular genetics of adRP.

1.5.4.1 Rhodopsin.

Rhodopsin was the first gene to be identified that was involved in RP. A large family was linked to the C17 (D3S47) marker on 3q21. The rhodopsin gene mapped to the same region and became an attractive candidate owing to its known role in retinal physiology. The first reported RP causing mutation in this gene was a single base change causing a pro→his substitution in codon 23 of the rhodopsin gene (Dryja et al. 1990). This codon is highly conserved among the opsins and related G-protein receptors. Since then, more than eighty-eight different mutations have been identified in this gene, causing approximately 25% of adRP in the Western world, as well as some cases of recessive RP and congenital stationary night blindness (reviewed by Gal et al. 1997). In addition there is interfamilial and intrafamilial variability in the clinical phenotypes observed in patients with these mutations. These mostly include missense mutations, a few frameshift deletions and nonsense mutations. The mechanisms by which these mutations lead to photoreceptor degeneration have not yet been established. It has been suggested that the majority of mutations 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/destabilisation into rod outer segments (Daiger et al. 1995). Rhodopsin constitutes the bulk of rod outer segment protein, therefore it has both a catalytic and structural role.

1.5.4.2 rds/Peripherin

Mutations within the peripherin/rds gene that cause adRP were also reported. This protein is localised to the outer segment disc membranes of both rods and cones, and is thought to be essential for the assembly and physical stability of outer segment discs. It has been associated with another membrane spanning protein, ROM1 (section 1.5.5.3). Mutations in the rds gene (retinal degeneration slow) causes a slow form of retinal degeneration in mice. The original rds mouse appeared to be recessive and its phenotype was thought to be restricted to rod photoreceptors only. In rds/rds
homozygous mice, the photoreceptor outer segments fail to develop, leading to a slow
degeneration of the photoreceptor cell bodies and cause blindness by six months.
However, on close examination a milder phenotype is seen in rds/+ heterozygotes,
where the outer segments develop but are shortened and disorganised, resembling the
histopathological phenotype of retinitis pigmentosa. The rds mutation is a 9.2 kb
insertion of exogenous DNA into exon 2 of the gene (Ma et al. 1995). The human
homologue of this gene mapped to chromosome 6p and was studied as a candidate for
human retinopathies. Both linkage analysis and mutational screening of this locus were
performed on adRP families. Farrar and co-workers (1991a) found linkage in an Irish
pedigree to chromosome 6p, where peripherin/rds was located. Meanwhile, mutations
in this gene were identified in some adRP afflicted patients by two groups (Farrar et al.
1991b; Kajiwara et al. 1991). In addition, peripherin mutations are implicated in
autosomal dominantly inherited macular degenerations (Wells et al. 1993; Nichols et
al. 1993). Also different diagnoses have been assigned to relatives with the same
mutation, which suggests that genetic background and environmental factors may play
important roles in disease. These results conclude that the destabilisation of the outer
segment disc membrane is very damaging to the photoreceptor.

1.5.4.3 Other unidentified adRP genes.

Many families did not show linkage to either of these loci, indicating that the
disorder is genetically heterogeneous. At least 9 other non-syndromic loci have been
implicated with adRP to date. Blanton and co-workers (1991) reported linkage to
chromosome 8q11 (RP1) in a large seven generation family exhibiting a type2/R
phenotype with variable expressivity. Further genetic heterogeneity was seen when
Inglehearn and colleagues (1993) found linkage between adRP and markers on
chromosome 7p13-15 (RP9; discussed in section 1.6). Also in the same year a new
locus on 7q was identified in a Spanish family with early onset of disease (RP10;
Jordan et al. 1993) and two other families with RP have been since reported to link to
this region. Another locus for adRP was discovered on 17p13, segregating through a
South African family of British descent with severe early onset adRP. It mapped close
to the recoverin gene but this gene was excluded by mutation screening (RP13;
Greenberg et al. 1994). Subsequently RP in two other families originating from Britain
have been linked to this region. A further locus was identified at 17q24 in a South African family (RP17) and two genes, PDEG and TIMP2 have since been excluded by linkage analysis (Bardien et al. 1997). A new locus was found on 19q13.4 and to date at least nine families with RP, have been reported to segregate with this region suggesting it is a major locus for adRP (RP11; Al-Maghtheh et al. 1994 and 1996; McGee et al. 1997). Families mapping to this locus have R-type adRP with partial penetrance. Finally, a large Dutch family with adRP has been linked to the pericentromeric region of chromosome 1 (RP18; Xu et al. 1996). Most of these loci do not have obvious candidate genes mapping to the critical regions and researchers are therefore employing a positional cloning approach to gene identification. In addition at least six other medium sized families with RP have been excluded from these loci indicating further genetic heterogeneity. The high level of genetic heterogeneity in this disease group reflects the complexity of the visual process.

1.5.5 Other inherited non-syndromic RP loci.

1.5.5.1 ArRP.

At least 10 non-syndromic loci have been implicated with arRP and 8 genes have been identified with mutations segregating in arRP families. Three mutations in the rhodopsin gene have been found to cause arRP (reviewed in Gal et al. 1997). Two of these, a nonsense and a splice site mutation, are not likely to encode a functional rhodopsin. The third is a missense mutation resulting in an amino acid change in the second cytoplasmic loop of rhodopsin.

A second implicated gene is the rod specific cGMP phosphodiesterase β subunit (cGMP-PDEβ; section 1.4.4.1.3). This was a likely candidate as the mouse gene for cGMP-PDEβ had been found to cause retinal degeneration in rd mice, which are considered to exhibit the clinical phenotype of arRP (Pittler and Baehr 1991), and because the protein product has a key role in the phototransduction pathway. Mutations in this gene were also found to cause retinal degeneration in Irish setter dogs. Subsequently four mutations, a compound heterozygote with two nonsense mutations; a frame-shift deletion; and a missense mutation, in the human homologue were discovered (McLaughlin et al. 1993). These and additional mutations later found
in cGMP-PDEβ are thought to be responsible for about 5% of arRP cases and mutations in this gene has also been implicated in dominant congenital stationary night blindness (adCSNB; McLaughlin et al. 1995). Nonsense and missense mutations in arRP patients have also been found in the α subunit of rod cGMP phosphodiesterase (cGMP-PDE α; Huang et al. 1995). These mutations are likely to result in the absence or alteration of the putative catalytic domains of these proteins, resulting in a decrease of phosphodiesterase activity. This is likely to lead to an increase in cGMP levels within the rod cell which in turn may be toxic to the photoreceptors. However, in large proportion of arRP patients, mutations have not been identified in any of these genes, indicating further heterogeneity. The PDE γ subunit gene has not been implicated in disease as yet. However, mutations in the mouse PDE γ gene that were introduced by gene-targeting, has resulted in a retinal degeneration pattern that resembles retinitis pigmentosa (Tsang et al. 1996).

Like the above genes, those coding for proteins which are involved in the phototransduction pathway or those that are expressed in retinal cells, seemed prime candidates for involvement in hereditary retinal disorders. Consequently, screening unrelated RP patients for mutations in the gene for the α subunit of the rod specific cGMP gated channel protein (section 1.4.4.1.3) identified mutant sequences co-segregating with the disease among families with arRP (Dryja et al. 1995). The mutations are likely to be null or encode a channel protein that is deleterious to rod photoreceptors.

Mutations segregating with arRP patients have also been found in the RPE65 gene (Gu et al. 1997). This encodes a 65 kD protein with an unknown function, but it is expressed exclusively in the retinal pigment epithelium layer and it has been proposed that it may play a role in vitamin A metabolism in the retina. This gene is also implicated in Lebers congenital amaurosis type 2, a severe retinopathy. Other genes known to be implicated in arRP, include the retinal ATP-binding cassette transporter (ABCR), which encodes a transmembrane protein that is located exclusively in the rod outer segment and is possibly involved in the energy dependant transport of substrates across membranes. This gene was initially implicated in Stargardt’s disease and age-related macular dystrophy. Subsequently a consanguineous arRP family carrying two putative ABCR null alleles in affected individuals was
Introduction

reported. These were caused by a 1 bp deletion early in the transcript that generates a frameshift and results in a truncated protein (Martinez-Mir et al. 1998). In addition, a missense mutation in the cellular retinaldehyde-binding protein gene (CRALBP) has been reported in affected individuals of a consanguineous arRP family (Maw et al. 1997). This gene encodes a water-soluble protein that carries 11-cis-retinaldehyde or 11-cis-retinal and is thought to modulate their transport and interaction with visual cycle enzymes in the RPE. Mutant CRALBP was found to lack the ability to bind to 11-cis-retinaldehyde, thereby presumably results in a disruption of vitamin A metabolism. Finally, mutations in the TULP1 protein have been discovered in arRP patients (Hagstrom et al. 1998). This gene is expressed exclusively in the retina but its function is unknown.

Other loci which are implicated in non-syndromic arRP by linkage analysis, include those on chromosome 1q31-q32.1 (Soest et al. 1994) and recently 16p (Gal et al. 1996 unpublished).

1.5.5.2 X-linked RP.

X-linked forms of RP are more severe in terms of onset and progression, mostly commencing in the first decade and progressing to partial or complete blindness by the third or fourth decade of life. The first RP locus to be mapped was found on Xp11.3 in five British families (RP2; Bhattacharya et al. 1984). A second locus was soon implicated as a patient (referred to as BB) suffering from RP, DMD, chronic granulomatous disease and McLeod syndrome, was found to have a deletion in band Xp21 (RP3; Franke et al. 1985). Wright and colleagues (1987) presented linkage data that supported the hypothesis that the RP3 locus was distinct from RP2. Other loci on the X chromosome have been proposed by genetic analysis but have not been proved definitively.

Recently the disease gene implicated in RP3 has been identified and is the first RP gene to be cloned using a positional cloning strategy. The deletion in patient BB upon which the RP locus was founded, actually gave a misleading physical localisation for RP3. Another deletion (75 kb) was found ~400 kb from this region in a patient with classical xIRP only. Meindl and co-workers (1996) analysed sequences from cosmids in this critical region using the GRAIL prediction program. They
identified and characterised the RPGR gene which has sequence homology to RCC1-related repeats. These repeats have been found to be associated in genes involved in GTPase regulation. Northern analysis showed it was ubiquitously expressed, highly in the retinal pigment epithelial layer, but was not present in the neural retina. Mutation screening identified nonsense mutations in affected patients giving rise to probable null alleles, and missense mutations in conserved regions of the RCC1 repeat domain, suggesting loss of function of RPGR. However, mutations in this gene have only been found in a small percentage of RP3 patients so far, which may support the evidence for a second locus near this, or could be due to unidentified mutations outside the coding regions.

1.5.5.3 Digenic RP.

Another inheritance pattern seen to cause RP involves the simultaneous segregation of heterozygous mutations in different genes have been found to give rise to RP. As well as allelic and nonallelic heterogeneity in RP cases, digenic inheritance complicates studies further. The rod outer segment protein-1 (ROM1) is a 37 kDa rod photoreceptor specific integral membrane protein which has significant homology to peripherin/rds (section 1.5.4.2), and is also localised to the rod photoreceptor disc rim. Both these proteins are known to interact together for the morphogenesis and structural maintenance of the photoreceptor disc rims. To date four families have been reported where RP occurred only in double heterozygotes for mutations in the unlinked ROM1 and peripherin/rds loci. In all cases, a Leu185Pro substitution was present in peripherin/rds. This was associated with one of three different 1 bp insertions found early in the coding region of ROM1. These insertions cause frameshifts and premature stop codons, which are likely null alleles (Kajiwara et al. 1994). The presence of either the peripherin/rds or ROM1 mutation without the other is not pathogenic in these families, as carriers of just one mutation are asymptomatic. The precise functions and interactions of these proteins are not known. These findings have presented further complexity in RP inheritance and suggest that further candidate gene analyses should be performed to determine the involvement of additional loci in other RP cases. ROM1 alleles segregating with RP in affected families which are not accompanied by peripherin/rds mutations have also been identified, but due to the small size of those
kindreds involved, these alleles cannot be definitively established as the sole cause of retinal disease (Bascom et al. 1995).

1.6 AdRP on chromosome 7p13-15 (RP9).

1.6.1 Genetic analysis of RP9.

In 1993 Ingleheam and colleagues reported a fourth adRP linked locus (RP9) on chromosome 7p13-15 in a large nine-generation English family (adRP7). The markers D7S435 and D7S460 were found to give significantly positive two point lod scores of 4.19 ($\theta = 0.046$) and 5.65 ($\theta = 0$) respectively with the disease. These map within 2 cM of each other with a lod score of 12.15. D7S435 had previously been mapped by FISH to 7p13-15.1. Multipoint analysis of these two markers and disease gave a peak lod score of 8.22 at D7S460 that placed adRP within 15 cM of these two markers and significantly excluded the disease in the immediate vicinity of D7S435.

A more accurate location of the disease was determined by using a number of additional markers in multipoint analysis on adRP7 patients and nine further affected individuals. The disease locus was placed between markers D7S484 (proximal) and D7S526 (distal) with a peak lod score of 17.8. Haplotype analysis showed two crossovers between D7S526 and the disease, and one crossover with D7S484 (Inglehearn et al. 1993). The interval between these markers had been estimated by Généthon to be approximately 4 cM in their “second generation linkage map of the human genome” (Weissenbach et al. 1992). However by recalculating the interval in a collection of their own families Inglehearn and colleagues suggested the interval maybe 1.6 cM. At the time of study, no obvious candidate genes appeared to map within the critical region. However, the disease interval was sufficiently small to commence physical mapping from the flanking markers into the disease region. The search for other genetic markers to further refine the region was also continued.

Additional markers, once published, were tested in the family and haplotype analysis with these refined the region even further to between D7S484 and D7S795 (Keen et al. 1995a). A number of markers were ordered along the critical region providing a high resolution genetic map which also serves as a framework of STSs across the region to support physical mapping.
The genetic and physical mapping effort in the RP9 positional cloning project has greatly benefited researchers working on other diseases mapping to 7p (figure 1.8). These include a Wilms' tumour suppressor locus with a breakpoint in the RP9 interval (WT; Reynolds et al. 1996); dominant cystoid macular dystrophy (DCMD; Kremer et al. 1994); non-syndromic hearing loss (DFNA5; Van Camp et al. 1995); Charcot-Marie-Tooth axonal neuropathy (CMT2D; Ionasescu et al. 1996); Blepharophimosis syndrome (BPES2; Maw et al. 1996); and spinal muscular atrophy (SMAD1; Christodoulou et al. 1995).

1.6.2 Clinical phenotype of RP9.

Affected individuals in the nine generation RP9 linked family, which has the laboratory reference number aDRP7, show the type-2/R form of aDRP. Documentation of blindness in early generations was obtained from census and clinical records (Jay et al. 1992). Affected individuals display a spectrum in their age of onset and show wide variation in the progression of the disorder. Of the 41 cases where the age of onset was known, six experienced symptoms before the first decade of life; 23 in the second with the remainder varying from 21 to over 50 years of age. Out of 73 family members with an abnormal genotype (from generation IV onwards), 13 are asymptomatic or show no signs of the disease; 26 are considered mildly affected and were independently mobile until the age of 60; 11 became blind in their 40's and 50's and the remaining 13 were severely affected and were blind within their 20's and 30's.

In one study (Moore et al. 1993), electrophysiological and psychophysical testing has shown that most apparently asymptomatic patients with aDRP do not have total non-penetrance but always show at least some detectable ERG and psychophysical mild retinal abnormalities. However another study by Kim et al. (1995) found that some asymptomatic patients show only very mild fundus changes and actually exhibit normal (or near normal) electrophysiology and psychophysical responses. In this family there is no recognisable correlation between the severity of disease with the age of onset or between siblings and parents. These feature presents difficulty for genetic counselling. The reasons for such a wide spectrum in the age of onset and in the clinical variability maybe due to unknown environmental or genetic factors.
Figure 1.8 The approximate genetic or physical locations of chromosome 7p linked diseases mapping around the RP9 locus. (Not to scale).
1.7 Aims of this thesis.

The purpose of this study was to employ a positional cloning strategy to identify the disease causing mutation for one form of retinitis pigmentosa, RP9. RP9 had already been localised to chromosome 7p13-15 by linkage analysis. The first priority was to construct a physical map of YAC clones across the disease interval. This would enable integration of genetic and physical maps within the critical region. It also would provide a valuable tool for the identification of genes residing in the area with which to build a high-resolution transcript map.

Two strategies were employed for gene identification. Chromosome 7 expressed sequence tags (ESTs) were sub-regionally assigned to provide positional candidates for RP9 and other chromosome 7 mapped diseases. Also, a novel cDNA selection procedure, the sandwich selection technique, was assessed experimentally to determine its success and specificity as a hybrid selection method. This method was applied in test experiments for isolating both control cDNAs and cDNAs from the RP9 region.
Chapter 2. Materials and Methods

Note: A list of general reagents, solutions and suppliers used in this study is at the end of this chapter (section 2.14)

2.1 DNA Isolation.

2.1.1 Plasmid and cosmID DNA preparation.

Small scale preparations of cloned DNA, were performed using the Wizard™ Miniprep columns (Promega Co., USA). This procedure is based on the alkaline lysis method for plasmid purification and employs a silica based resin to bind to plasmid DNA that is contained in the cleared lysate. Single colonies or samples from glycerol stocks of libraries were inoculated into 5 ml of LB (2.14.5.1) and were grown overnight in a 37°C shaking incubator (200 rev/min). Cells were harvested by centrifugation for 10 min at 3,000 x g. The supernatant was discarded and the pellet was resuspended in 300 μl of Cell Resuspension solution (section 2.14.2.1), then transferred to a 1.5 ml eppendorf. 300 μl of Cell Lysis solution (section 2.14.2.1) was added and mixed by inverting repeatedly until the suspension turned opaque. Protein and RNA components of the cell extract were precipitated by adding 300 μl of Neutralisation solution (section 2.14.2.1), mixing thoroughly, and placing on ice for 15 min. The precipitated cell debris was pelleted by centrifuging at 13,000 x g for 15 min, and the clear lysate supernatant was transferred to a fresh 1.5 ml tube. 500 μl of Binding resin (section 2.14.2.1) was mixed well with the clear lysate and vacuum pumped through Wizard mini-columns. This was followed by 1.5 ml of Column Wash solution (section 2.14.2.1) and the resin was dried by continuing to draw a vacuum for 30 seconds after all the wash solution has been pulled through. The column containing the DNA bound resin was then spun in a microfuge for 20 seconds to remove any remaining traces of wash solution. This was then transferred to a fresh tube and the bound DNA was released and collected by applying 25 μl preheated (80°C) dH₂O to the dried resin, leaving for 1 min, and then centrifuging the tube at 13,000 x g for 30 seconds to elute the DNA. This step was then repeated to bring the total elutant to 50 μl. The DNA was stored at -20°C.
For cosmid mini-preps the volume of overnight culture was increased to 10 ml and the quantity of cell resuspension, lysis and precipitation solutions were increased to 600 µl.

2.1.2 Purification of DNA from preparations or PCR.

2.1.2.1 Phenol chloroform extraction.
For small volumes of DNA, the extraction volume was increased to 200 µl with 1 x TE. Larger volumes were extracted directly. An equal volume of phenol chloroform: isoamyl alcohol (25:24:1) was mixed gently into the solution by inversion. The samples were then centrifuged at 6,000 rpm for 3 minutes to separate the aqueous and organic layers. The top aqueous layer was then transferred to a clean tube and the phenol extraction was repeated if necessary, depending on the protein content of the sample. An equal volume of chloroform was added, mixed and re-spun. The aqueous samples were transferred to a fresh tube for ethanol precipitation.

2.1.2.2 Ethanol precipitation.
For standard DNA precipitations, two volumes of absolute ethanol and 1/10th volume of 3 M sodium acetate were mixed and placed at -80°C for 30 min to allow the DNA to precipitate. The samples were then spun at 13,000 rpm for 15 min. The pellet of DNA was washed with 70% ethanol and the samples were centrifuged again for 5 min. The pellet was either air dried or was placed under vacuum for approximately 10 minutes. The samples were either stored in this state at -20°C or resuspended in 1 x TE.

2.1.2.3 Sephacryl-S200 and S400 HR columns (Pharmacia, UK).
S200 HR columns were used for desalting, buffer exchange, and the removal of labelled or unlabelled nucleotides, providing the DNA fragment was >100 bp. S400 HR columns serve the same purpose, but also removes primers or excess primer dimers from PCR products prior to sequencing. The Sephacryl HR resin (sephacryl equilibrated in TE buffer, pH 7.6) was resuspended in the column by gentle vortexing. The cap was loosened and the base snapped off before placing in an open topped eppendorf. This was spun at 3,000 rpm for 1 min to compact the gel matrix. The
column was transferred to a fresh tube, and 20-25 µl of the DNA sample was added to the bed of resin. The sample was collected at the bottom of the tube by centrifuging again for 1 min at 3,000 rpm.

2.1.3 Plating of YAC colonies and identification of recombinants.

Yeast artificial chromosome (YAC) vectors, cloned into Saccharomyces cerevisiae host cells, were grown in AHC selective media (2.14.5.1). This is a rich uracil -ve, tryptophan -ve medium. Yeast cells containing YACs that were identified by library screenings, were first streaked onto AHC agar plates with ampicillin and incubated at 30°C for 24-48 hours. The pYAC4 vector bears the selectable marker genes TRP1 and URA3. Large DNA inserts are ligated between the vector arms at the Eco RI cloning site (Schlessinger 1990). This is located within the sup4 gene which encodes a mutant tRNA that suppresses the ochre (nonsense) mutation in the ade2 gene of the yeast host. Cloning of DNA into this site therefore insertionally inactivates the sup4 gene, abolishing the suppression. Adenine metabolism is therefore disrupted and results in the accumulation of a phenotypically visible red adenine pre-metabolite. Recombinant clones produce red/pink colonies, and non-recombinants appear white.

2.1.4 Yeast solution DNA preparation.

This protocol yields DNA that is fragmented but is sufficient for PCR analysis. Single yeast colonies were inoculated into 10 ml AHC broth (2.14.5.1) supplemented with ampicillin (35 µg/ml), and incubated overnight in a 30°C shaking incubator (200 rev/min). 0.2 ml of the overnight culture was used to seed 100 ml AHC broth, then left to grow for a further 24 hours or until cells reached a density of 3.3 x 10^6 cells/ml (O.D.600 = 1.5-2.0). Cells were pelleted at 4,000 x g for 5-10 min, and the supernatant was discarded. These cells were resuspended in 5 ml solution I (section 2.14.2.2) and 20 µl of 10 mg/ml Lyticase. These were then incubated, at 37°C for 1 hour without shaking, and the resulting spheroplasts were collected by gentle spinning at 1,000 x g for 10 min. The supernatant was subsequently discarded and the cells were then resuspended in 5 ml of solution II (section 2.14.2.2). These were first incubated at 65°C for 10 min, then cooled to room temperature. An equal volume of ethanol was
Materials & Methods

added, and the precipitated nucleic acids were collected by centrifugation at 2,000 \( \times \) g for 10 minutes. The pellet was resuspended in 2 ml of 1 \( \times \) TE pH 7.4, with 100 \( \mu \)g/ml RNase and incubated at 37°C for 30 minutes. 200 \( \mu \)g/ml proteinase K was added, and the samples were further incubated at 65°C for 1 hour. These were then chilled at 4°C for 30 minutes prior to phenol extraction. The aqueous DNA solution was extracted twice with one volume phenol:chloroform (1:1), then once with one volume of chloroform:isoamyl-alcohol (24:1) (section 2.1.2.1). The nucleic acids were subsequently precipitated with 2.5 volumes of absolute ethanol in the presence of 0.3 M sodium acetate. After pelleting the precipitates, several 70% ethanol washes were performed, and the vacuum dried DNA pellet was resuspended in 250 \( \mu \)l
\( \text{dH}_{2}\text{O} \). Samples were stored at -20°C. 1/50\(^{th}\) dilutions were used for PCR analysis.

2.1.5 Yeast DNA plug preparation.

Due to the high molecular weight of YAC DNA in comparison to other cloned molecules, the intact strand of DNA is subject to increased breakage and damage. In this method, yeast cells are converted into spheroplasts, and are then embedded in agarose. These are treated with Proteinase K, leaving the DNA intact and retaining its stability. The agarose plugs are used for sizing YACs by pulse field gel electrophoresis (PFGE) and for restriction enzyme digestion. This is a modification of the method by Volrath and Davies (1987).

Single yeast colonies were inoculated into 10 ml AHC broth (section 2.14.5.1) with ampicillin (35 \( \mu \)g/ml) and grown overnight in a 30°C shaking incubator (200 rev/min). 0.2 ml of this was used to seed 100 ml of AHC, and left to grow for a further 36-48 hours or until the cells grew to stationary phase. Cells were harvested by centrifugation at 4,000 rpm for 10 min, and the supernatant was discarded. The pellet was then washed in 15 ml SCE solution (section 2.14.2.3) and the centrifugation was repeated. The resultant pellet was resuspended in 800 \( \mu \)l SCEM solution (section 2.14.2.3) to approximate a cell density of 2.5 \( \times \) \( 10^9 \) cells/ml (O.D. \(_{600} = 1.5\)). Lyticase (2 mg/ml in 0.01 M sodium phosphate containing 50% glycerol; Sigma) was added to the cell suspension, to a final concentration of 0.5 \( \mu \)g/ml and was incubated for 1 hour at 30°C without agitation. 1 volume of 1% molten low melting point agarose in 1 M sorbitol (55°C) was mixed with the cell suspension and immediately dispensed into
pre-cooled 200 µl volume plastic well moulds. The agarose plugs were left to set in the moulds on ice for at least 30 min. These were then carefully removed and incubated in 15 ml proteinase K solution (section 2.14.2.3) at 50°C for 2 days. The plugs were subsequently rinsed 4-5 times in 1 x TE, before storing at 4°C in 1 x TE with 10 mM EDTA.

2.2 DNA digestions with restriction endonucleases.

Genomic and cloned DNA were digested with the appropriate restriction enzymes (NEB and Promega) and the manufacturer's recommended buffers. Reaction volumes and incubation times varied according to the amount and nature of DNA being digested. In general, the reaction composed of a 1 x concentration of the recommended buffer, with 1 unit of restriction enzyme per µg of template DNA for digestion. Digestion was performed following the manufacturer's recommendations, and at the appropriate incubation temperature specified for the enzyme. 20 µl standard reaction volumes were used when digesting 1-2 µg of DNA, but the volume was increased when digesting more template DNA or for total genomic digests. Double digests that required the use of enzymes which are active in incompatible buffers were carried out sequentially. Digests using the lowest salt buffer were performed first, then the salt concentration and the volume was adjusted for the second enzyme. In cases where reactions had high salt buffer concentrations, spermidine was added to a final concentration of 5 mM to prevent partial digestion. The volume of enzyme in the reaction was kept to less than 1/10th of the total volume to avoid altering or reducing the specificity of target sites due to the presence of glycerol. The reaction was arrested by heating the samples at 65-85°C to denature the restriction endonucleases. Complete digestion was confirmed by electrophoresis of a fraction of the reaction on a 1% agarose gel.
2.3 Fractionation of DNA.

*Note:* Fractionation of DNA in polyacrylamide gels is described in section 2.9.3, and fractionation of RNA in denaturing formamide gels is described in section 2.11.3.

2.3.1 Agarose gel electrophoresis.

Products of digestion reactions or PCR were routinely size fractionated by electrophoresis through 1% or 2% agarose gels (w/v) in 1 x TAE buffer (2.14.3), depending on the size resolution required (Sambrook *et al.* 1989). DNA was loaded in 1 x Ficoll buffer (2.14.3) with size standards in adjacent lanes (Promega φX174/Hae III, λ/Hind III or Gibco/Promega 1 kb ladder markers). These were electrophoresed at 80 V for 30 min for check gels, or 30 V for 16 hours when higher size resolution was required. DNA fragments were visualised by staining the gel in ethidium bromide solution (0.5 μg/ml) for 15 min, then placing on a UV transilluminator for photography. In cases where fragments needed to be isolated, the DNA was generally run in low melting point agarose. Bands were excised using sterile scalpels.

2.3.2 Pulse Field Gel Electrophoresis (PFGE).

The development of Pulse Field Gel Electrophoresis (PFGE; Schwartz and Cantor 1984; Carle and Olson 1984) has made megabase molecular studies possible. This technique enables direct resolution of molecules of up to 10 Mb, including chromosomes of lower organisms. Separation of large DNA fragments is achieved through molecular reorientation by exposure to alternating electrical fields. Various size ranges can be resolved depending on differing variables, most importantly the duration of the alternating electric fields (pulse/switch time). The Bio-Rad CHEF-DR II or the Pharmacia-LKB Gene Navigator™ apparatus was used, and operates the CHEF (clamped homogeneous electric field) technique.

A 1% agarose (molecular biology certified, Biorad) gel in 0.5 x TBE (14 cm x 12.7 cm, 5 mm thick), was set on a level surface, in the casting tray with the comb provided by the manufacturer. Agarose plugs containing embedded yeast DNA, were cut to the size of the wells or to the size required for loading 1-2 μg, with a scalpel. The plugs were gently slid into the wells, ensuring they were placed against the front
and bottom of the well. These were sealed in with 1% low melting point agarose (Sigma) in 0.5 x TBE.

Gels were run in 0.5 x TBE buffer at 16°C, with continual but gentle buffer recircularisation. Pulse times (switch times), field strength (voltage), and length of run varied according to the resolution desired. In general, for providing good separation of the *Saccharomyces cerevisiae* chromosomes, pulse field conditions of 60 second to 90 second ramping switch time, at 175 volts for 24 hours was applied. The conditions varied slightly between each machine.

Following electrophoresis, the gel was stained in 0.5 µg/ml ethidium bromide solution and placed on a UV transilluminator for photography. For identification of YAC clones within the *Saccharomyces cerevisiae* chromosome background, the DNA was transferred to a nylon membrane by Southern blotting (section 2.4.1), and hybridised with labelled Cot-1 DNA probes.

### 2.4 Transfer of DNA to nylon membranes.

#### 2.4.1 Southern transfer.

Transfer of electrophoretically separated DNA by capillary action, from an agarose gel to the surface of membranes, was first described by Southern (1975). The nucleic acids were denatured prior to transfer onto Nylon membranes. The DNA was initially fragmented by immersion of the gel into depurinating solution (section 2.14.4.1), for 15 min. It was then transferred to denaturing solution (section 2.14.4.1) for 30 min. After rinsing with dH₂O, the gels were finally placed in neutralising solution (2.14.4.1) for at least 30 min.

The gels were inverted and placed onto a wick of 3MM Whatmann paper assembled over a reservoir of 20 x SSC. Hybond-N+ or Hybond-N membranes were cut to the exact size of the gel and marked for orientation of the lanes. The membrane was placed onto the gel in one action and the assembly was covered with a further two pieces of 3MM Whatmann pieces soaked in transfer buffer. At all times, care was taken not to introduce any air bubbles between the layers. This was finally overlaid with dry filter paper, and a uniform weight was placed on top. Capillary transfer could be achieved in eight hours, but was normally left to proceed overnight. After the stack
was dismantled, the membrane was rinsed in 2 x SSC, placed onto dry filter paper DNA side up, and nucleic acids were permanently immobilised on the membrane either by baking at 80°C for 2 hours, or by UV crosslinking (70,000 μJ/cm²).

Alternatively, smaller fragments (< 20 kb) were transferred by using 0.4 M NaOH as the transfer agent, and the use of only Hybond N+ membranes. Gels were soaked in 0.4 M NaOH for just 20 minutes prior to transfer, and the assembly was set up as described above. This avoids the necessity of the long winded initial denaturation process.

2.4.2 Screening of bacterial libraries

Nylon filters (Hybond-N+, Amersham) cut to the size of the plate to be screened, were placed on the surface of the LB agar, supplemented with the required antibiotic. Bacterial culture was plated onto the surface of the filter, by spreading with a flamed glass rod, providing even distribution. This master filter was incubated at 37°C overnight until single colonies were approximately 2 mm in diameter.

Two replica filters were prepared from each master plate. The master filter was placed colony side up on 3MM Whatmann filter paper. Replica filters were pre-wetted by placing onto fresh LB plates, then carefully overlaying them on top of the master filter, ensuring no air bubbles were introduced between the filters. These were covered with a second piece of Whatmann paper, and equal pressure was enforced with a glass plate and rolling pin to transfer all colonies from the master filter. Asymmetric needle holes were introduced at the edges for orientation. Replica filters were gently peeled off and placed onto LB agar plates with antibiotic. A second replica filter was constructed by repeating the procedure. Colonies were grown to 2 mm by incubation overnight in the 37°C oven.

The filters were carefully removed and placed colony side up, on 3MM Whatmann paper soaked in denaturing solution for 10 minutes, then in neutralising solution for 10 minutes (section 2.14.4.1), to lyse the cells, and promote binding of nucleic acids to the filter. These were then placed in 2 x SSC for 5 minutes. After air drying, the filter was baked at 80°C for 1 hour to fix the DNA onto the nylon membrane. These were then ready for hybridisation (section 2.6.2).
2.4.3 Screening of Phage libraries.

Plates of bacteriophage plaques (section 2.10.2) were cooled at 4°C to harden the top agarose. Bacteriophage DNA was transferred to nitrocellulose filters (Amersham) by gently placing the filter on the surface of the agarose for three minutes. Orientation marks on the filter were made with a sterile needle and dye. The filter was carefully removed and placed plaque side up, on 3MM Whatmann soaked in denaturing solution for 5 minutes, followed by neutralising solution for 5 minutes (2.14.4.1), and finally 2 x SSC for 2 minutes (2.14.1). After air drying, the filter was baked at 80°C for 1 hour. Duplicate filters were made by repeating the above, but the second filters were left on the top agarose for 6 minutes. Hybridisation of probes to these was filters were carried out as in section 2.6.2.

2.5 Labelling of DNA probes.

2.5.1 Random priming labelling of probes.

The random primed labelling technique is a primer extension method developed by Feinberg and Volgelstein (1983). Hexanucleotides of random sequence are annealed to denatured probe DNA. These act as primers for incorporation of [α-3²P]dCTP into newly synthesised strands by the action of Klenow polymerase. A commercial kit (Ready to Go -CTP labelling Kit, Pharmacia) was used following the manufacturer’s instructions. PCR probes were spun through S-200 columns prior to labelling (section 2.1.2.3). 25-50 ng of probe DNA was denatured in 27 μl dH₂O, and immediately quenched on ice. 20 μl of dH₂O was used to resuspend a vitrified mix containing dATP, dGTP, dTTP and random hexanucleotides in 1 x reaction buffer. 3 μl (1.1 Mbq) of [α-3²P]dCTP and the denatured DNA was added to the mix and the reaction was incubated at 37°C for 15-30 min. 0.5 μl was removed and the degree of incorporation was tested by the precipitation of the labelled probe with 5% trichloroacetic acid onto a glass fibre disc (Whatmann GF/B). Unincorporated dCTP was removed by passing the labelled mix through Sephadryl S-200 spin columns (Pharmacia). The labelled probe was denatured at 95°C for 3 min immediately before adding it to pre-heated hybridisation solution.
2.5.2 End labelling of probes.

20 pmol of the oligonucleotide was end-labelled in 10 μl volumes containing 1 x reaction buffer (One-Phor-All buffer, Pharmacia), 3.6 Mbq [γ³²P]ATP and 5 units of T4 polynucleotide kinase (Pharmacia). The reactions were incubated at 37°C for 45 minutes and 2 μl of this was used directly in a PCR reaction with one labelled and one unlabelled primer. 6 μl of formamide loading buffer (2.14.3) was added to the reaction. Samples were denatured at 95°C immediately before loading onto acrylamide gels.

2.6 Hybridisation of labelled probes to DNA immobilised on Nylon filters.

Hybridisation was performed in sealed sandwich boxes in a shaking incubator, or in glass bottles in a Hybaid rotational oven. Filters containing the immobilised DNA were moistened with 2 x SSC then placed in the container of choice, with nylon meshes separating individual filters.

2.6.1 Identification of YAC DNA from pulse field blots.

50 ng of human Cot-1 DNA was labelled and purified accordingly (section 2.5.1). Filters were prehybridised in 20 ml Church’s buffer (Church and Gilbert 1984; section 2.14.4.2) for 1 hour at 65°C. Hybridisation was carried out in fresh Church’s solution at 65°C for approximately 16 hours. Filters were then washed to the appropriate stringency (section 2.6.3).

2.6.2 Identification of single copy sequences from complex backgrounds.

Filters were prehybridised with 25 ml Denhardt’s hybridisation solution with denatured 75 μg/ml sonicated salmon sperm (section 2.14.4.2) at 65°C for at least 3 hours. The denatured labelled probe was added to fresh pre-warmed Denhardt’s hybridisation solution with the filters, and left to hybridise at 65°C for at least 16 hours. Filters were then washed to the appropriate stringency (section 2.6.3).
2.6.3 Post-hybridisation washes.

Filters were washed once in 2 x SSC (section 2.14.1) and 0.1% SDS initially, then twice in 1 x SSC and 0.1% SDS for 20 minutes at room temperature. If necessary, they were also washed at a higher stringency, in 1 x SSC and 0.1% SDS at 65°C, and in some cases with 0.5 x SSC and 0.1% SDS. All filters were sealed in plastic film to retain moisture, and were exposed to autoradiograph film (Kodak, XAR-5 or X-OMAT AR) for varying times depending on signal strength. In some cases, films were exposed with intensifying screens at -80°C, if the signals were poor.

When filters were required for repeat hybridisations, they were stripped for the removal of probes from the previous hybridisations. The filters were washed in 0.4 M NaOH solution at 50°C for 30 min, with constant shaking. These were subsequently replaced with a 0.2 M Tris (pH 7.5), 0.1 x SSC and 0.01% SDS solution, and washed for a further 30 minutes at 50°C. Alternatively, the filters were immersed in boiling dH₂O for three minute intervals. This was repeated several times. Removal of the probe was confirmed by autoradiography.

2.7 Polymerase chain reaction (PCR).

2.7.1 Standard PCR protocol.

Standard PCR reactions were performed using Bioline UK products. A typical PCR reaction consisted of 1 x concentration of (NH₄)₂SO₄ PCR buffer (1.6 mM (NH₄)₂SO₄, 67 mM Tris-HCl and 0.1% Tween-20), 200 μM each dNTP, 7.5 pmol each oligonucleotide primer, 1.5 mM MgCl₂, and approximately 100 ng template DNA in a total volume of 25 μl. 0.2 units of BIOTAQ™ taq polymerase was added to the reaction either before or following an initial denaturation step of 94°C for 5 min. MgCl₂ concentrations were adjusted to optimum conditions for each primer pair. The reaction was overlaid with mineral oil, unless performed in equipment with heated lids. Thermocycler models available were Hybaid OmniGene and Perkin Elmer GeneAmp PCR system 2400 and system 9600.

Typical parameters for cycling consisted of an initial denaturation step for 5 min, followed by 35 cycles at 94°C for 20 seconds, an annealing step for 20 seconds, and extension at 72°C for 30 seconds per kb. A final extension step at 72°C for 5
minutes was sometimes included. Annealing temperatures were calculated by estimating the dissociation temperature (T_d) using the formula: \( (4(G+C) + 2(A+T)= T_d) \), and subtracting 2-5°C.

2.7.2 PCR from phage or colonies.

Plasmid or yeast colonies were subjected to PCR after incubation in a solution which assists in the disruption of intact cells to release the DNA into the surrounding media, which eliminates the need for purified DNA for PCR (Ling et al. 1995). Isolated colonies where picked with sterile tooth picks or disposable pipette tips and were washed in 10 µl of an enzyme solution consisting of 1.2 M sorbitol, 100 mM sodium phosphate pH 7.4, and 2.5 mg/ml Lyticase. This was incubated at 37°C for 5 min and 2 µl of this was used for PCR.

Individual bacteriophage plaques were isolated by stabbing the surrounding area of the plaque and transferring the plug of agar to 20 µl of SM buffer. Phage particles were allowed to diffuse out into the surrounding media at room temperature for several hours. 2-4 µl aliquots of this were used in a PCR reaction. The samples were treated to an initial heat denaturation in the PCR reaction for 5-10 min to destroy the bacteriophage coat, or were incubated in the enzyme solution prior to addition in the PCR reaction, as described above.

2.7.3 Isolation of YAC terminal sequences: Alu-vector arm PCR.

This method was employed to generate YAC terminal sequences for the production of STSs to use in chromosome walking and is a modification of Nelson et al. (1989). It employs primers chosen from Alu consensus sequences (ALE1 and ALE3, section 2.14.8) and YAC vector oligonucleotides (pYAC4-LA(a/b) and pYAC4-RA (a/b), section 2.14.8) to amplify inter-Alu products which reside between the YAC vector termini and the nearest Alu repeat in the insert. Standard PCR reactions were carried out in 50 µl volumes containing ~200 ng of individual YAC DNA in solution, with the following combinations of primers:

(i) ALE1 + pYAC4-RA(a)   (ii) ALE1 + pYAC4-LA(a)
(iii) ALE3 + pYAC4-RA(a)   (iv) ALE3 + pYAC4-LA(a)
(v) ALE1 only             (vi) ALE3 only
The cycling profile consisted of 35 cycles at 94°C for 1 min, 60°C for 1 min and 72°C for 2 minutes. Products were visualised by agarose gel electrophoresis and unique products generated within the ALE-vector arm samples (which represent YAC insert termini) were directly sequenced using nested pYAC4 primers (pYAC4-RA(b) and pYAC4-LA(b), section 2.14.8).

2.7.4 PCR screening YAC libraries.

The ICI human YAC library (Anand et al. 1990) was screened by PCR using a variety of STSs from the RP9 region on 7p14-15 to identify corresponding YAC clones. The library was screened as hierarchical pools of decreasing complexity until a single positive clone was identified. The library consists of 34,500 clones with an average insert size of 350 kb and a >3.5 x coverage of the human genome. It was made from a human lymphoblastoid 48XXXX cell line and approximately 10% of the clones are thought to be chimeric. All library pools were received as agarose plugs in 5 mM EDTA and were rinsed several times in TE and in dH₂O prior to use. The plugs were melted at 65°C for 10 minutes and 2-3 µl was used per PCR reaction. All YAC clones were ordered from the HGMP resource centre.

PCR pools: 40 primary pools (numbered 1-40) which consist of ~860 YAC clones each. Each primary pool has 9 corresponding secondary pools (A-I), which represent the 9 plates comprising each primary pool (i.e. 95 YAC clones per secondary pool). Each of the 20 tertiary pools correspond to the row (pools 1-8) and columns (pools 9-20) of each plate (i.e. secondary pool) to establish the plate co-ordinates for the positive clone(s).

2.8 Cloning.

2.8.1 Ligation of digested fragments into plasmids.

Prior to ligation, the DNA products were digested with chosen restriction enzymes which ensured they possessed compatible ends with the vector. In instances where there was no appropriate restriction site, ligation was performed by blunt end cloning. 5' protruding ends of the vector and/or insert were filled in with Klenow polymerase creating blunt termini by addition of 1 unit of DNA Klenow enzyme.
(Promega) per μg vector DNA and 40 μM each dNTP into the digestion buffer. This was incubated at room temperature for 30 min, and then the enzyme was inactivated by heating at 70°C for 10 min. The DNA was then purified by phenol:chloroform extraction and precipitated in 2.5 volumes 100% ethanol in the presence of 1/10 volume 3M sodium acetate (section 2.1.2.1 and 2.1.2.2).

To prevent self-ligation of the vector, the 5' ends were dephosphorylated by the addition of 0.1 units of calf intestinal alkaline phosphatase (CIAP) per pmol of ends, directly to the digestion mixture. This was incubated first at 37°C for 15 min, then at 56°C for 30 min. The incubation steps were repeated after adding the same amount of enzyme. The reaction was stopped by adding 2 μl of 0.5 M EDTA, and heating at 65°C for 45 min. The DNA was then purified and precipitated as above.

The vector was ligated to an appropriate amount of insert, such that the ratio of molar ends of the vector to insert DNA was in the range of 1:1 to 1:3, unless specified. Ligation was carried out using Pharmacia Ready-To-Go™ T4 DNA ligase kit, following the manufacturer’s protocol, and was left to incubate overnight in a 12°C water bath. An aliquot of the ligation mixture was either used directly for transformation into E. coli by heat shock, or was salt precipitated (Ready-to-Go T4 DNA ligation protocol) prior to electroporation. The remainder was stored at -20°C.

2.8.2 Ligation of PCR products into plasmids.

Two systems were used for cloning PCR derived templates. The TA Cloning kit (Invitrogen) provides a quick one-step cloning strategy which takes advantage of the non-template dependant activity of taq thermostable polymerase used in PCR, which adds single deoxyadenosine to the 3' end of all duplex molecules. The pCRII vector (TA cloning kit) and the pTAg vector (R&D systems) have a compatible 3' dT overhang at each termini which allow easy insertion of these PCR products. This method allowed cloning of the PCR products without prior purification. The protocol was followed according to the manufacturer's recommendations.

The second system was used when the vector of choice did not posses compatible 3’ dT overhangs. PCR products were amplified using Ventγ DNA polymerase (New England Biolabs). This is a thermophillic DNA polymerase, with a 3'→5' proof-reading exonuclease, thereby removing 3' overhangs and result in blunt
Materials & Methods

ended templates. The PCR products were spun through Sephacryl S-400 columns (Pharmacia) for buffer exchange and removal of residual primers and dNTPs. The 5' termini of the fragments were phosphorylated prior to ligation by treatment with polynucleotide kinase. 5 μl of 10 x kinase buffer and 1 μl of ATP (100 mM) was added to 20 μl of the cleaned PCR product and the total volume was brought up to 50 μl with dH2O. The samples were heated to 70°C for 5 min and cooled on ice. 2 μl of T4 polynucleotide kinase was then added, and the sample was incubated at 37°C for 30 minutes. These were subsequently extracted with phenol:chloroform and precipitated prior to ligation. The Pharmacia Ready-To-Go T4 ligation kit was used to ligate the PCR product into blunt ended vector treated with CIAP as described in section 2.8.1.

2.8.3 Transformation of *E. coli* by heat shock treatment.

In most instances, DH5α or XL1-Blue cells were used for heat shock transformation. Competent cells were made by a simple one-step procedure developed by Chung et al. (1989). A 20 ml culture of cells was grown to early log phase (O.D._600 = 0.3-0.4), and cells were harvested by centrifugation at 2,000 rpm for 5 min at 4°C. The pellet was gently resuspended in 1/10th volume of filter sterilised ice cold TSB solution (10% (w/v) PEG 6000, 5% DMSO, 50 mM MgCl₂ and 50 mM MgSO₄ in LB broth). 50 μl aliquots were either used immediately or frozen for storage at -80°C.

Transformation was accomplished by gently mixing approximately 50 ng or 1-2 μl of a ligation reaction to 50 μl thawed competent cells. After incubating on ice for 20 min, the cells were subjected to heat shock at 42°C for 1 min, and were quickly transferred to ice for 5 min. 450 μl of SOC (section 2.14.5.1) was added to the cells, and was placed in a 37°C shaking incubator (200 rev/min) for 1 hour. Generally, 200 μl aliquots were used for spreading onto LB agar plates with the appropriate antibiotic (section 2.8.5).

2.8.4 Transformation of *E. coli* by electroporation.

XL1-Blue MRF' *E. coli* cells were the preferred hosts in all instances. Cells for electroporation were either purchased (Stratagene), or were prepared in the laboratory:

200 ml of LB broth (section 2.14.5.1) supplemented with 12.5 μg/ml
tetracycline was inoculated with a single colony of bacteria and grown until mid log phase (O.D\textsubscript{600} = 0.5-1). After cooling on ice, cells were harvested by centrifugation at 4,000 rpm for 10 min at 4°C and the supernatant was subsequently removed. The cells were then washed in several steps to lower the ionic strength in the surrounding media, in order to eliminate the problem of arcing during electroporation. After each wash the cells were spun again at 4,000 rpm for 10 min and the supernatant was discarded. The pellet was first washed with 1 volume of cold dH\textsubscript{2}O, then with 0.5 volume cold dH\textsubscript{2}O, then in 0.02 volume cold 10% glycerol, and were finally resuspended in 0.002-0.003 volumes cold 10% glycerol. 50 µl aliquots were dispensed into 1.5 ml eppendorfs, flash frozen in liquid nitrogen and stored at -80°C until use.

1-2 µl salt precipitated ligation or DNA sample was mixed gently with 50 µl of thawed cells on ice. This mixture was transferred to a chilled 1 mm electroporation cuvette, tapping gently so cells settled at the bottom of the chamber, and left to chill for 1 min on ice. The cuvette was then transferred to the electroporation chamber (Bio-Rad Gene Pulser II) and subjected to 4.5-5 msec pulse (time constant), using voltage settings of 2.5kV, 200Ω and 25µF. 0.5 ml or 1 ml of SOC (section 2.14.5.1) was immediately added to the cuvette to resuspend the cells and was then transferred to a 15 ml Falcon polypropylene tube. After incubation at 37°C shaking (200 rev/min) for 1 hour, 10 µl or 50 µl aliquots were taken for spreading onto antibiotic-containing LB agar plates.

2.8.5 Plating transformed bacterial clones and selection of recombinants.

Most plasmid vectors used in this study permit identification of recombinant clones by exploiting the intragenic complementation system of the \textit{E. coli} β-galactosidase Lac Z gene. This is where the plasmid of choice contains a portion of the Lac Z gene that codes for the α-peptide portion of the β-galactosidase protein (termed Lac Z'). When transformed into \textit{E. coli} host cells which have a modified or deleted section of the α-peptide coding region (e.g. DH5α or XL1-Blue), the β-galactosidase enzyme can be produced by the expression of the bacterial and plasmid genes to produce a complete molecule. The plasmid cloning sites are located in the Lac Z’
gene. Recombinant clones therefore disrupt this, hindering the production of a complete β-galactosidase molecule. This can be detected in a colour assay, as the enzyme causes the breakdown of the lactose analogue X-gal to a substance that is deep blue. Recombinant clones can therefore be identified as white clones, and non-recombinant clones as blue when plated onto media containing X-gal and its inducer IPTG. Standard selective LB agar plates were used for this analysis.

50 µl of X-gal (20 mg/ml in dimethylformamide) and 100 µl of IPTG (24 mg/ml in dH₂O, filter sterilised) were plated onto 25 ml 100 mm diameter LB agar plates containing the appropriate antibiotic (section 2.14.5.2). The plates were inverted and left in a 37°C oven for 30 min with their lids slightly ajar to permit evaporation. Aliquots of the culture containing transformed clones were plated, and incubated overnight at 37°C. The blue white colour assay determined identification of recombinants.

Cosmid clones which were obtained from previously constructed libraries were grown in the same manner as plasmids, in LB broth or streaked onto LB agar plates which are supplemented with the appropriate antibiotic. Individual plasmid or cosmid colonies or total library clones were grown overnight in 10 ml LB broth supplemented with the appropriate antibiotic (section 2.14.5.2). Glycerol stocks for all cultures were made by freezing 1 ml aliquots of culture at -80°C to a final concentration of 20% glycerol.

2.9 DNA Sequencing.

All sequencing methods used here were adapted from the chain-termination sequencing method developed by Sanger et al. (1977). This is based on enzymatic synthesis of a template strand by polymerase action and termination of the extended products by the incorporation of dideoxynucleotides, which lack hydroxyl residues at the 3' position of the deoxyribose. This prevents the formation of a phosphodiester bond with free nucleotides, thereby leading to termination of chain synthesis. The incorporation of labelled nucleotides during strand synthesis provides easily detectable products corresponding to each nucleotide in the sequence, which are separated on high resolution acrylamide gels for reading.
2.9.1 **Manual sequencing of cloned DNA.**

The Pharmacia T7 polymerase kit was used. This employs a standard $^{35}$S-dATP incorporation labelling method. Approximately 1-2 μg of double stranded DNA was alkali denatured in 2 mM NaOH and 20 μM EDTA at room temperature for 10 min in a total volume of 40 μl. The DNA was then precipitated in 100 μl of absolute ethanol in the presence of 1/10th volume 3 M sodium acetate at -20°C for 30 minutes. The DNA was recovered by centrifugation at 13,000 rpm for 20 minutes. The resulting pellet was washed in 70% ethanol, vacuum dried and resuspended in 10 μl dH2O.

To the denatured DNA, 2 μl of sequencing primer (10 pmol) and 2 μl of annealing buffer was added, and the reaction was incubated at 65°C for 5 min, then 37°C for 10 min to allow annealing of the primer to the template. 3 μl of labelling mix A (containing dCTP, dGTP, dTTP), 1 μl of $^{35}$S-dATP (0.36 M bq) and 2 μl of diluted T7 polymerase was added, mixed and incubated at room temperature for 5 min. 2.5 μl of each of the 4 dideoxy termination mixes (ddATP, ddCTP, ddGTP and ddTTP) in separate tubes, were heated to 37°C. 4.5 μl of the $^{35}$S-dATP reaction mix was added to each of the termination mixes, and strand synthesis occurred at 37°C for 5 min before the reaction was terminated by the addition of 5 μl Formamide stop buffer (2.14.3). 2-3 μl of the reaction was then run on a polyacrylamide gel after denaturation.

2.9.2 **Direct sequencing of PCR products.**

PCR products were purified by spinning through sephacryl S-400 microspin columns (Pharmacia, section 2.1.2.3). 2.5 pmoles of sequence specific primer was labelled at the 5' end using [γ-$^{32}$P]dATP in a final volume of 10 μl by adding 1 μl T4 polynucleotide kinase (5-10 units), 3 μl [γ-$^{32}$P]dATP (3,000 Ci/mM) and 1 μl One-Phor-All buffer. All components except the isotope were purchased from Pharmacia. The reaction mixture was incubated at 37°C for 45 minutes. 2 μl of the labelled primer was then added to 10 μl of purified DNA (~100 ng). This was denatured and placed on ice to prevent strand reassociation. 2 μl annealing buffer (DTT), 2 μl reaction buffer dATP (label mix A) and 2 μl diluted T7 polymerase were then combined with the denatured template and primer. After mixing carefully, 4 μl of this was added to each
Materials & Methods

tube of 2.5 µl pre-warmed (37°C) ddNTPs, as described in section 2.9.1, then incubated for 5 min at 37°C. The reaction was terminated with 4 µl of stop solution. Samples were denatured prior to loading on a denaturing polyacrylamide gel.

2.9.3 Denaturing polyacrylamide gel electrophoresis.

To resolve small differences in size between DNA fragments, a high resolution denaturing gel was prepared to fractionate single stranded DNA fragments. Samples were run in 6%, 0.4 mm thick denaturing polyacrylamide gels, on 50 cm length Bio-Rad electrophoresis apparatus. Prior to assembly, the back sequencing plate (which was attached to the buffer reservoir) was silanised with Sigmacote (Sigma). Acrylamide concentrate (19:1 acrylamide:bisacrylamide) in 8.3 M urea solution (Sequagel, National Diagnostics), was mixed with dilutent (8.3 M urea) and buffer (10 x TBE in 8.3 M urea) to give a 6% gel solution in 8.3 M urea and 1 x TBE. A 50 ml acrylamide plug was set at the bottom of the plate, by the addition of 6 µl/ml TEMED (BDH) and 6 µl/ml 25% (w/v) ammonium persulphate. For polymerisation of the gel solution, 2.5 µl/ml of 25%(w/v) ammonium persulphate, and 1 µl/ml TEMED was added, mixed and quickly poured into the apparatus, taking care to prevent the introduction of any air bubbles.

Prior to loading, the samples were denatured by heating at 95°C for 3 min, and were then immediately loaded into a pre-warmed (50-55°C) 6% gel and electrophoresed in 1 x TBE buffer at a constant power (50 W for 25 cm wide gels, and 90 W for 40 cm wide gels). After running for the required time, the gels were fixed in a 10% methanol-10% acetic acid solution for 5 min, transferred to 3MM Whatmann paper and wrapped in cling film before vacuum drying. The gel was then exposed to autoradiograph film.

2.9.4 Automated fluorescent sequencing.

Automated sequencing was performed on an ABI 373A DNA sequencer (Perkin Elmer), using the ABI PRISM™ Amplitaq® DNA polymerase FS, Dye Terminator Cycle Sequencing Ready reaction Kit. Automated cycle sequencing has several advantages over manual radioactive incorporation techniques. Foremost, it is
Materials & Methods

safer as it eliminates the use of radioactive nucleotides. The technique is based on using 4 fluorescent dye-labelled terminators which are detected by lasers that analyse and convert the DNA sequence to a graphical image on a computer. Reactions are also carried out in single tubes, and do not have to be separated for each terminator nucleotide. Less starting material is required and the sequencing of larger templates is possible. Only one lane on the denaturing gel per reaction is required, as individual labelled nucleotides can be distinguished by unique fluorescent labels. Also the results can be seen immediately in a computer analysis file.

Both cloned material and PCR products could be used for cycle sequencing. PCR products were directly sequenced after purifying the template to remove unincorporated primers and dNTPs by passing through Centricon-100 spin columns (Centricon), according to manufacturer's recommendations. A fraction of the purified product was quantified upon an agarose gel. The cycle sequencing reaction for both cloned and PCR fragments consisted of 8 µl terminator ready reaction mix (includes labelled dye terminators, buffer, and dNTPs), 1 µl sequence specific primer (3.2 pmole) and 0.5 µg template DNA (cloned DNA or PCR product) in a total volume of 20 µl. This was subjected to 25 cycles of 96°C for 10 seconds, 50°C for 5 seconds and 60°C for 4 minutes in a Perkin Elmer GeneAmp 9600 or 2400 thermal cycler. Excess unincorporated fluorescent dye was removed by ethanol precipitation of the DNA with 2.5 x volume of 95% ethanol in the presence of 0.3 M sodium acetate. This was then washed with 70% ethanol and vacuum dried. The resultant pellet was resuspended in 5 µl of ABI loading buffer (section 2.14.3), and the sample was denatured before loading onto the denaturing acrylamide gel. The ABI 373A DNA sequencer was set up and run in accordance to the manufacturer’s instructions. The gel was generally run for 13 hours or overnight. The output data was converted to a text file and an analysis file on an Apple Macintosh computer.
2.10 Phage Library Manipulations.

2.10.1 Preparation of plating bacteria culture for infection.

For most titrations and for amplification, the host bacterium XL1-Blue MRF' (Stratagene) was used. An individual colony of the bacterial host was inoculated into 10 ml LB broth which was supplemented with 0.2% maltose (w/v) and the appropriate antibiotic. After overnight growth in a 37°C shaking incubator, 1 ml was seeded into 50 ml LB broth with 0.2% maltose, in a sterile flat bottom conical flask and grown to mid log phase (O.D.₆₀₀ = 0.5) The culture was cooled on ice and the cells were pelleted by centrifugation at 3,500 rpm for 10 minutes at 4°C. The supernatant was discarded, and the cells were resuspended in ice cold 10 mM MgSO₄ to a density of approximately 1.6 x 10⁹ cells/ml (O.D.₆₀₀ = 2). These cells were viable for up to a week when stored at 4°C. Maltose was not added when amplifying the phage library.

2.10.2 Titrating and plating phage plaques onto solid media.

To estimate the number of plaque forming units per millilitre, appropriate dilutions were made from the bacteriophage stocks in SM. In general, 10⁻³ and 10⁻⁶ dilutions were used. 1 µl of a dilution was mixed with 100 µl of plating bacterial culture. These were incubated for 30 minutes at 37°C to allow the phage to adsorb into the cells. A negative control, of host cells uninfected by phage, was always carried out in addition. 3 ml of molten LB top agarose (50°C) was added, mixed and quickly poured onto pre-warmed (37°C) 90 mm LB plates. After the top agarose layer hardened, the plates were inverted and incubated in a 37°C oven overnight. The following day, the number of plaques were scored and the titre of phage in the original bacterial stocks were calculated.

The titre estimate was used to determine the dilution needed to obtain the required number of plaques for amplification (section 2.10.3) or screening of the library (section 2.4.3). The desired dilution was then plated onto E. coli hosts on LB agar plates, as described above.

2.10.3 Amplification of Phage libraries.

Amplification of phage in solution may lead to significant under-representation or loss of some recombinant phage that grow poorly due to competition from faster growing ones. Therefore amplification of the phage library was carried out by plating
onto solid media, and eluting off the progeny phage to ensure that individual plaques have the same chance of growing.

The preparation of high titre phage lysates on solid media was performed similarly as with the titrations, although on a larger scale. A dilution of the library stock was calculated and estimated to provide approximately 100,000 plaques when plated on a 506.25 cm² (22.5 cm x 22.5 cm) Nunc plate. The dilution of phage library stock was added to 250 μl of host bacterial plating cells, without growth in maltose, and incubated at 37°C for 30 minutes for adsorption. 6.5 ml of molten top agarose (50°C) was added, inverted to mix, and immediately poured onto slightly dried and pre-warmed (37°C) 200 ml LB agarose plates with tetracycline (12.5 μg/ml). After leaving at room temperature to allow the top agarose to set, the plates were sealed with parafilm, inverted and incubated in a 37°C oven for approximately 8 hours, or until the plaques were visible and around ~0.5 mm in diameter.

12 ml of sterile SM solution was added to each plate and swirled to ensure full cover. The plates were left flat at 4°C overnight to allow the phage particles to diffuse out. These were then left on a shaker at slow speed for a further 2 hours at room temperature. Under sterile conditions, the SM from each plate was collected and transferred to falcon tubes using large pasteur pipettes, and all the eluted samples were combined. 200 μl chloroform was added and mixed to the collected sample to lyse contaminating bacterial host cells. To remove any top agarose debris that may have co-transferred with the eluted SM, the samples were spun at 4,000 rpm for 5 minutes and the supernatant was recovered. 1-2 drops of chloroform was added, and the amplified phage lysate stock was stored at 4°C. The titre of the amplified library was later determined.

2.10.4 Preparation of Bacteriophage DNA.

This procedure is designed primarily to avoid contamination of host cell DNA and RNA by the addition of the nucleases to the phage lysate solution. The phage DNA is protected in its protein capsid during this stage, and nucleases are subsequently removed by proteinase K incubation, SDS, heat inactivation and phenol extraction (Chisholm 1989).

100 μl of chloroform was mixed well into 40 ml of amplified phage lysate.
Materials & Methods

400 μl of nuclease solution (50 mg DNase I (Sigma), 50 mg RNase A (Sigma), in 10 ml 50% glycerol, 30 mM sodium acetate, pH 6.8) was added and incubated for 30 min at 37°C. 2.1 g NaCl was then gently dissolved, and the mixture was centrifuged for 20 minutes at 7,000 rpm at 4°C. The supernatant was then decanted into a new falcon tube. 3.7 g of polyethylene glycol MW 8,000 was gently dissolved and the bacteriophage particles were left to precipitate on ice for 60 minutes. After centrifugation at 7,000 rpm for 20 minutes (4°C), the supernatant was removed, the pellet was resuspended gently in 500 μl of SM and then transferred to 1.5 ml microfuge tubes. 500 μl of chloroform was carefully mixed with this, then spun in a microfuge for 5 min, and the supernatant was transferred to a fresh tube. 20 μl 0.5M EDTA, 5 μl 20% SDS and 10 μl Proteinase K (2.5 mg/ml) were added and left to incubate at 65°C for 30 min. The DNA was then extracted twice by phenol:chloroform (1:1), then once with chloroform:isoamyl alcohol (24:1) and was subsequently precipitated with ammonium acetate and isopropanol. After centrifugation for 15 min in a microfuge, the pellet was washed with 70% ethanol, and dried briefly under vacuum. The pellet was resuspended in 250 μl of TE and the yield was determined by spectrophotometry.

2.10.5 Transfer of Charon BS library to pBluescript Phagemid.

The conversion of a Charon BS library to pBluescript is accomplished by digesting the recombinant phage with Not I to release the plasmid which contains cDNA inserts. These are subsequently recircularised by self ligation, and later transformed into E. coli (Swaroop 1993).

Approximately 20 μg of recombinant phage library DNA was digested to completion with Not I restriction enzyme (section 2.2), which was subsequently inactivated by incubating at 65°C for 30 minutes. An aliquot was separated on a 0.8% agarose gel by electrophoresis to ensure total digestion was achieved and to verify that the correct sized products were observed. The volume of the digested products was brought up to 898 μl with dH₂O. 100 μl (10 x) ligation buffer and 2 μl of T4 DNA ligase was added and then left to incubate at 12°C overnight. The DNA was then precipitated for 1 hour at -80°C after the addition of 2.5 x volume 100% ethanol in the presence of 0.3 M sodium acetate. Following centrifugation for 15 min at 13,000 x g,
Materials & Methods

the pellet was washed in 70% ethanol, vacuum dried and then resuspended in 40 μl of TE.

2 μl aliquots of this were used in transformations by electroporation (section 2.8.4) into XL-1 Blue MRF' cells (Stratagene). The entire ligation reaction was transformed to ensure that as complete as possible a representation of library clones was achieved. After 1 hour of incubation in SOC, 10 μl and 50 μl aliquots were taken for plating and further analysis. 4 ml LB broth supplemented with ampicillin (40 μg/ml) and tetracycline (12.5 μg/ml) was added with the remaining transformation mixture, and were left to grow in the 37°C shaking incubator overnight. Cultures from all transformations were combined and mixed, and glycerol stocks (final concentration of 20% glycerol) were made for long-term storage at -80°C to serve as a phagemid library stock. The remaining cultures were used to prepare single-stranded DNA.

2.10.6 Preparation of single stranded DNA.

The MI3 based phagemid pBluescript is a derivative of the pUC19 vector. This contains the intergenic region of the f1 filamentous phage which accommodates the cis-acting functions for phage DNA replication and morphogenesis. Single stranded DNA can be recovered when the phagemid is co-infected with a helper phage in E. coli cells that contain an F' episome. The superinfecting bacteriophage provides the material that is required for packaging. The R408 helper phage (Stratagene) used in this instance, is a modified f1 phage which preferentially packages single stranded DNA of phagemids at typically a 30-40 fold higher rate than its own (Russel & Kelly, 1986). The single stranded rescue protocol described here is modified from the Promega Protocols and Applications Guide.

500 μl of an overnight culture of the pBluescript plasmid cDNA library within XL1-blue MRF' host cells was inoculated into 25 ml of 2 x YT media with 12.5 μg/ml tetracycline (for selection for F' episome containing hosts) and 100 μg/ml ampicillin. This was incubated in a 37°C shaker for 30 min or until growth to O.D._600= 0.3, to which R408 helper phage was subsequently inoculated at a multiplicity of infection of 20 (helper phage particles to cells). This was left to incubate for 8 hours with continuous vigorous shaking. The culture was transferred to 1.5 ml eppendorfs for centrifugation at 13,000 x g to pellet the cells. After transferring the supernatant to
fresh tubes, the cells were pelleted and the supernatent transferred again to ensure the level of contaminating cellular DNA was at a minimum. 0.25 x volume of phage precipitation solution (3.75 M ammonium acetate, 20% PEG) was mixed to the resultant supernatant and left at -80°C for 2 hours. After collecting the precipitates by centrifuging at 13,000 x g for 15 min, the supernatant was discarded. The PEG precipitates were resuspended in 30 µl of TE, and all the samples were pooled. 1 volume of chloroform:isoamyl alcohol (24:1) was added and vortexed for 1 minute to lyse the phage particles. After spinning at 13,000 x g for 5 min, the aqueous phase was transferred to a new tube. The single stranded DNA was then extracted with phenol:chloroform (1:1) several times, and finally with chloroform:isoamyl alcohol (24:1). The DNA was precipitated with 0.5 x volume of 7.5 M ammonium acetate and 2 volumes 100% ethanol. The resulting pellet was washed in 70% ethanol before drying briefly under vacuum (sections 2.1.2.1 and 2.1.2.2). This was resuspended in 20 µl Nuclease-Free dH₂O. The quantity of DNA was determined by spectrophotometry and by comparison of band intensities with the control samples.

2.11 RNA manipulation.

2.11.1 RNA conduct.

All equipment, and work areas were kept thoroughly clean and free of commercial ribonucleases during manipulations with RNA. All glassware, and solutions (except Tris-based solutions) were pre-treated by filling with 0.01% DEPC dH₂O, leaving overnight at room temperature, then autoclaving for 30 min to remove all traces of DEPC. Non-disposable plasticware were rinsed with chloroform. Tris based solutions were prepared with DEPC pre-treated water. Where possible, the chemicals used were reserved separately from other uses. Filter tips and sterile disposable plasticware were utilised and gloves were used at all times.

2.11.2 In Vitro transcriptions.

To linearise nucleic acid templates prior to transcription, DNA from the sublibrary of pSV9Zf9 was digested with Sal I, and DNA from the pSV7Zf3 vector was digested with Sst I and Pvu II to completion. 50 µg/ml proteinase K was added to the digested products in the restriction buffer, and incubated for 30 min at 37°C. These
were extracted twice with phenol:chloroform (1:1), precipitated with ethanol and resuspended in nuclease free dH₂O (sections 2.1.2.1 and 2.1.2.2). Transcriptions were performed using Promega Riboprobe in vitro Transcription System kit, and the protocol was modified to obtain high yields of RNA.

Standard transcription reactions consisted of: 10 µl 5x transcription buffer, 5 µl of 100 mM DTT, 50 units of RNasin Ribonuclease Inhibitor, and 10 µl ribonucleotide mix (2.5 mM of each Bio-UTP or UTP, and ATP, GTP, CTP) with 30-40 units of RNA polymerase to approximately 5 µg of linearised template. The reaction was brought up to 50 µl total volume with nuclease free dH₂O. This was incubated for 90-120 min at 37-40°C. For the preparation of “genomic RNA”, ordinary UTP was used, with SP6 RNA polymerase. Bio-11-UTP (Enzo diagnostics) was substituted for UTP, and T7 RNA polymerase was used when synthesising biotinylated “capture RNA”.

To remove the DNA template, 1 u/µg Promega RQ1 RNase free DNase was added to the transcriptions and incubated for a further 15 min at 37°C. The “genomic RNA” samples were extracted once with phenol:chloroform (1:1) and once with chloroform:isoamyl alcohol (24:1). These were subsequently precipitated on dry ice for 2 hours or overnight at -80°C with 2.5 volumes 100% ethanol in the presence of 2 M ammonium acetate. After spinning at 13,000 x g for 15 min, the pellet was washed with 70% ethanol, dried, and resuspended in 20 µl of RNase free dH₂O. The biotinylated RNA samples were purified by passing through a Sephadex G-50 column (Boehringer Mannheim) to remove free nucleotides, after the addition of 0.5 M EDTA to stop the reaction. RNA concentration was determined by O.D.₂₆₀ measurement and by gel electrophoresis. The samples were stored at -80°C until use.

2.11.3 Denaturing gel electrophoresis of RNA.

Gel tanks, combs, and trays were cleaned by washing in 2% SDS and rinsing in DEPC dH₂O prior to use. 5-10 µg of transcribed RNA or size marker was denatured in 25 µl RNA loading buffer (section 2.14.3) and 1 µl 1 mg/ml ethidium bromide, by heating 10 min at 65°C and was immediately transferred to ice. The RNA was size fractionated through a 1% (w/v) agarose gel containing 5% formaldehyde, in 1 x MOPS buffer. The gel was run in 1 x MOPS buffer for 2 hours at high voltage (100v).
2.12 Sandwich selection strategy (Yan and Swaroop 1994; chapter 5).

2.12.1 cDNA selection system.

This selection procedure involves solution hybridisation of cDNAs to target genomic DNA which has been *in vitro* transcribed. Specific cDNAs are captured onto an avidin matrix via a biotinylated "capture RNA" transcript, which has a sequence tag homologous to one on the genomic RNA transcript (see figure 5.1 and section 5.1.1). Preblocking was performed to prevent non-specific interactions with repeat sequences. Non-specific products are washed away and selected cDNAs are isolated.

Approximately 20 µg of genomic RNA was mixed with 40 µg of Cot-1 DNA and precipitated in the presence of 2 M ammonium acetate and 2.5 x volume 100% ethanol at -20°C overnight. After centrifugation for 30 min at 14,000 x g, the precipitated DNA was washed in 70% ethanol, and vacuum dried. The sample was resuspended in 30 µl of 0.5 M sodium phosphate pH 7.2, 10 mM EDTA and transferred to a 0.2 ml eppendorf tube. This was denatured in boiling water for 10 min and then incubated immediately at 60°C in a Perkin Elmer thermal cycler with heated lid, for 2 hours to preanneal. 1-1.5 µg of single stranded cDNA was mixed with approximately 40 µg of biotinylated capture RNA. After precipitation, the samples were resuspended in 20 µl of 0.5 M sodium phosphate pH 7.2, 10 mM EDTA and 0.5% SDS. After brief denaturation, this mixture was added to the preannealed genomic RNA, and hybridisation was carried out for 20 hours at 60°C.

2.12.2 Capture and washing.

Vectrex Avidin D matrix (Vector laboratories) was used to capture the sandwich complexes. 200 mg of the matrix was rehydrated prior to use, with 10 ml binding buffer (100 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.5% SDS), and sonicated denatured salmon sperm (100 µg/ml). This was mixed by gentle rotation at room temperature for 30 min. The matrix swells to 1.25 x on rehydration. This was spun for 10 min at 3,000 rpm, and the supernatant was discarded. The matrix was resuspended in 10 ml fresh binding buffer, omitting salmon sperm DNA. This was mixed by vortexing and left to equilibrate at room temperature for 10 min. The hybridisation samples were added to the matrix, and left at room temperature for 60 min with
Materials & Methods

continuous shaking, to allow capture of biotinylated molecules. The samples were centrifuged at 3,000 rpm for 10 min, and the supernatant was decanted. The matrix was first washed twice with 10 ml of binding buffer, then twice with 10 ml of buffer A (10 mM Tris HCl pH 7.5, 2 mM EDTA, 0.5% SDS, and 100 mM NaCl), and finally twice with 10 ml of buffer B (10 mM Tris HCl pH 7.5, 0.5% SDS, 2 mM EDTA, and 50 mM NaCl). At each step, the matrix was washed with the solutions by continuous shaking for 10-20 min at room temp, then spun at 3,000 rpm for 10 min to pellet the matrix prior to removing the supernatant. After the final wash step the matrix was equilibrated for 10 min in 10 ml of pre-elution buffer (10 mM Tris-HCl pH 7.5, 2 mM 0.5M EDTA, and 80 mM 5 M NaCl). Captured cDNAs bound to the matrix are recovered by incubation in 600 µl of elution buffer (pre-elution buffer containing 125 µg/ml RNase A). After vortexing and subsequent incubation in a 37°C shaker for 30 min, the matrix was pelleted by centrifugation at 13,000 x g for 5 minutes, and the supernatant carefully transferred to a 1.5 ml eppendorf. The elution step was repeated to isolate any remaining cDNAs.

The captured cDNAs were then extracted once with phenol:chloroform, once with chloroform: isoamyl alcohol, and then precipitated overnight at -80°C with 2 volumes ethanol in the presence of 1/10th volume 3 M sodium acetate and 20 µg of glycogen (Boehringer Mannheim). The samples were then centrifuged for 30 min at 14,000 x g at 4°C, washed with 70% ethanol, vacuum dried and resuspended in 10 µl of dH2O. The selected single stranded cDNAs were subsequently transformed into E. coli cells by electroporation (section 2.8.4).

2.13 Sequence similarity searches.

The BLAST program (Altschul et al. 1990) and FASTA program (Pearson and Lipman, 1988) are methods which utilise rapid database searching algorithms for finding optimal sequence alignments between query and database comparisons. The BLAST program is a fast yet less sensitive mode of similarity searching as it disregards gaps to improve the alignment, whereas FASTA does take into consideration gaps in sequences. They both determine the smallest sum probability of gaining the resultant alignment from the query and database sequences by chance. DNA sequence databases can be searched with the query sequence using BLASTN, or
the query sequence can be translated in either three or six reading frames to search protein databases using BLASTX.

2.14 Reagents and Suppliers.

2.14.1 General reagents.

TE buffer: 0.5 ml Tris+HCl (2M) pH 7.5 and 20µl EDTA (0.5M) pH 8.0 in 100 ml dH₂O.

20 x SSC: 175.3g NaCl (3M), 88.2g sodium citrate. Dissolve in dH₂O to a total volume of 1 litre.

2.14.2 Solutions for DNA preparations.

2.14.2.1 Plasmid/cosmid preparations:

Cell Resuspension solution: 50 mM Tris- HCl pH 7.5, 10 mM EDTA, 100 µg/ml RNaseA.
Cell lysis solution: 0.2 M NaOH, 1% SDS.
Neutralisation solution: 1.32 M potassium acetate, pH 4.8.
Binding Resin: 5 g celite Analytical Filter Aid (BDH Chemicals), rehydrated in 250 ml dH₂O for 3 hours, drained off remaining dH₂O and resuspended in 250 ml of 6 M guanidine hydrochloride, 50 mM Tris HCl and 20 mM EDTA.
Column Wash: 80 mM NaCl, 8 mM Tris HCl, 2 mM EDTA, 55% ethanol.

2.14.2.2 Yeast solution DNA preparation:

Solution I: 0.9 M sorbitol, 20 mM EDTA and 14 mM β-mercaptoethanol pH 7.5.
Solution II: 4.5 M GuHCl, 0.1 M EDTA, 0.15 M NaCl and 0.05% Sarkosyl, pH 8.0.

2.14.2.3 Yeast DNA plug preparation:

SCE solution: 1 M Sorbitol, 0.1 M Sodium Citrate pH 5.8 and 10 mM EDTA.
SCEM solution: 2.4 µl/ml β-mercaptoethanol in SCE.
Protease K solution: 2 mg/ml protease K in 48 ml 0.5 M EDTA, 0.5 g SLS (N-lauryl sarcosine).

2.14.3 Electrophoresis analysis:

TAE buffer (10 x): 0.4 M Tris-acetate, 10 mM sodium EDTA, pH 8.
TBE buffer (10 x): 1 M Tris base, 0.83 M Boric Acid, 10 mM sodium EDTA, pH 8.3.
Ficoll Loading buffers (10 x): 25% (w/v) Ficoll, 0.25 EDTA, 0.25% (w/v) Orange G or (for agarose gels) Bromophenol blue, in 1 x TAE buffer.
Formamide loading/stop buffer (10 x): 95 ml formamide, 50 mg Bromophenol blue, 50 mg xylene cyanol (for radioactively labelled samples) and 4 ml 0.5 M EDTA.
ABI loading buffer: 5:1 formamide: 50 mM EDTA with 50 mg/ml Dextran Blue. (for loading fluorescent products)
Materials & Methods

RNA loading buffer: 0.75 ml formamide, 0.24 ml formaldehyde, 0.1 ml DEPC H₂O, 0.1 ml glycerol, 0.08 ml 10% Bromophenol blue and 0.15 ml 10 x MOPS buffer (0.2 M MOPS (MW 209.26), 50 mM sodium acetate (trihydrate MW 136.08) and 10 mM EDTA. Dissolve in 800 ml H₂O pH 7.0, make up to 1 litre).

2.14.4 Southern transfer and hybridisation solutions.

2.14.4.1 Southern transfer:

Depurinating solution: 0.5 M HCl.
Denaturing solution: 1.5 M NaCl, 0.5 M NaOH.
Neutralising solution: 3 M NaCl, 0.3 M sodium citrate, pH 5.5.

2.14.4.2 Hybridisation solutions:

Church’s solution: 0.5 M sodium phosphate buffer, pH 7.2, 7% SDS, 1 mM EDTA.
Denhardt’s hybridisation solution: 0.05 M sodium phosphate buffer, pH 7.2, 4 x SSC, 5 x Denhardt’s (1% (w/v) BSA, 1% (w/v) Ficoll and 1% (w/v) PVP), 0.3% SDS and 0.15% tetra-sodium pyrophosphate. 7.5 μg/ml denatured sonicated salmon sperm was added to the pre-warmed hybridisation mix for prehybridisation and hybridisation.

2.14.5 Host cells and Growth media.

E. coli host cells for plasmid, cosmid and phage vectors: DH5α (Gibco BRL) and XL1-Blue MRF’ (Stratagene).

Yeast hosts containing YAC clones: AB1380 (ICI).

2.14.5.1 Growth media:

Luria-Bertani (LB) broth and agar: 10 g NaCl, 10 g bacto-tryptone, 5 g bacto-yeast extract to 1 litre dH₂O. pH to 7.0 with 5 M NaOH. For LB agar add 20 g agar/litre of LB. Autoclave.

AHC Broth and agar: 1.7 g Yeast nitrogen base (without amino acids), 10 g casein acid hydrolysate and 20 mg adenine hemi-sulphate to 1 litre dH₂O. For AHC agar, 20 g agar/ litre was added. Autoclave.

SOC Media: 20 g bacto-tryptone, 5 g yeast extract and 0.5 g NaCl brought up to a volume of 900 ml with dH₂O, then autoclaved. 2.03 g MgCl₂, 1.2 g MgSO₄ and 3.6 g glucose were dissolved in 100 ml dH₂O, filter sterilised and added to the cooled media.

2 X YT media: 16 g bacto-tryptone, 10 g bacto-yeast extract, 5 g NaCl, pH 7.0.
SM buffer: 0.1 M NaCl, 50 mM Tris-HCl pH 7.5, 0.2% (w/v) MgSO₄·7H₂O, 0.001% (w/v) gelatin.

2.14.5.2 Antibiotics:

Ampicillin: Stock solution: 25 mg/ml of the sodium salt in dH₂O.
(most Plasmids). Working concentration: 35-50 μg/ml.
Kanamycin: Stock solution: 25 mg/ml in dH₂O.
(Cosmids) Working concentration: 50 μg/ml.
Tetracycline: Stock solution 12.5 mg/ml tetracycline hydrochloride in 50% ethanol (v/v).
(XL1-Blue MRF' cells). Stored in the dark.
Working concentration: 12.5-15.0 μg/ml.

All antibiotics were filter sterilised, and stored in aliquots at -20°C.

2.14.6 RNA reagents.

RNA loading buffer: 750 μl formamide, 240 μl formaldehyde, 100 μl DEPC dH₂O, 100 μl glycerol, 80 μl 1% orange G and 150 μl 10 x MOPS buffer.
10 x MOPS buffer pH 7.0: 0.2 M MOPS, 50 mM sodium acetate, 10 mM EDTA dissolved in DEPC treated dH₂O and pH to 7.0 with NaOH.

2.14.7 Suppliers.

Amersham International plc. UK. Bio-Rad laboratories Ltd. UK.
BDH Chemicals. Boehringer Mannheim Corporation, USA.
Clontech laboratories Inc. UK. Cruachem
Flowgen Ltd. UK. ICI (Zeneca), Cambridge Biosciences (suppliers for Enzo Diagnostics), UK.
Invitrogen Corporation, USA. New England Biolabs (NEB) Ltd. UK
Pharmacia Ltd. UK. Promega Corporation, USA.
Scotlab, UK. Stratagene Ltd. UK.
United States Biochemicals (USB), USA. Vector Laboratories, USA.

Chemicals not specified were purchased from:
BDH Chemicals Ltd. Gibco BRL.
Pharmacia Biotech. Sigma Chemicals, USA.
2.14.8 Oligonucleotides.

Primer pairs were either purchased from Cruachem or Amersham. Most oligonucleotides for the EST work was kindly synthesised by the HGMP resource centre.

EST primers:
All STSs used in this study are available from the dbSTS division of NCBI at http://www.ncbi.nlm.nih.gov/dbSTS, and can be accessed using their EST name or “D” locus number.

BCP primers:
BCP 1+ 5'-ATGAGAAAAATGTCGGAGGAAG-3'  BCP 2- 5'-GGCTCCAGCCAAAGAACGAGGTGG-3'  BCP 4+ 5'-TTGCAGCTCAGCAGCAGGAG-3'  BCP5- 5'GGTGGGCGCACTTGGGTAGA-3'

pYAC4 primers:
pYAC4-RA(a) 5'-ATATAGGCAGCGCAACCGACACTGTCGGC-3'  pYAC4-LA(a) 5'-CACCCTTCTGAGGACACTGTCGGACCGC-3'  pYAC4-RA(b) 5'-CTTGCAAGTCTGGGAAATGGAATGGAGACAAA-3'  pYAC4-LA(b) 5'-GTTGGGTAAAAGGCAGCAAG-3'

Alu repeat specific primers:
ALE1 5'-GCCTCCCAAGGTGCTGGAGTACA-3'  ALE3 5'-CCAT/CTGCACTCCAGCCTGGG-3'

Plasmid vector primers:
M13 5'-GTAAACGACCGCCGACGT-3'  reverse  M13 5'-GGAAACAGCTATGACCATG-3'
universal  T3 5'-GTAATACGACTCACTATAGGGC-3'  pTAG5 5'-GGATATGACCATGATTACGCCA-3'
T7 5'-GTAATACGACTCACTATAGGGC-3'  T3 5'-AATTAACCCTCACTAAAGGG-3'
SK 5'-CGCTCTAGAAGTCTGGATC-3'  pTAG5 5'-GCTATGACCATGATTACGCCA-3'

End clone primers:
A1: product size-0.136 kb  
A1 forward 5'-TGTAACAGCGCTGAAGTACCAC-3'  A1 reverse 5'-CCAATAATTAGTACCACTTC-3'

1545 LA: product size- 0.116 kb  
1545 LA 5'-GGGCTATCTCCTTGGATTCTC-3'  reverse  1545 LA 5'-GCAGCTAACATCACATTTACC-3'

27I-A5 LA: product size- 0.136 kb  
27I-A5 LA 5'-AAGTAATATACGCTGGATC-3'  reverse  27I-A5 LA 5'-CCTTAGGATGTCTTCTATGAG-3'
Materials & Methods

34D-A5 RA: product size- 0.223 kb

34D-A5RA 5’-TAGCCTAGGAAAGTCTCATTGC-3’  34D-A5RA 5’-GTACAGACTTTAGGATCACAGTC-3’
forward reverse

Oligonucleotides for gene specific STSs:

Amphiphysin: product size- 0.502 kb

AMPHF 5’-CCTCCCACCAATTCTAAAGT-3’ AMPHR 5’-TCGTGCTATCTGTATCTCA-3’

Clone E2-1:

   product size-0.500 kb

E2-1A 5’-CTTAGTGAGACTCATTGTAC-3’  E2-1B 5’-CACAGAAAGAATGAACACGTC-3’

   product size-0.156 kb

E2-1C 5’-GACGTGTTCATTCTTCTGTG-3’  E2-1D 5’-ACTGATACAGAAAACATGGGCC-3’

Clone E2-19:

A001W23F 5’-GAGTTTACATATACCTGAAGATGA-3’  A001W23R 5’-GTAATCATATATCTCTGCAATC-3’

Clone E1-10:

1 5’-CACTACACTCTATCTACTGCTGCTC-3’  2 5’-GTTTGCTAGTTGTCTCTGTGACAG-3’
3 5’-GGGAACACTTATACCTCTGGGGGTG-3’  4 5’-CCGACATGAAATAGGCTGG-3’
5 5’-CCAGCCTATTTTCATGTGG-3’
Chapter 3. Physical Mapping of the RP9 Critical Region

3.1 Introduction.

The first step in identifying a disease causing gene by positional cloning is to localise it to a sub-chromosomal region (section 1.3). Cytogenetic rearrangements or deletions in affected patients can be resolved to ~10 Mb, while in the absence of karyotype abnormalities, genetic analysis can delimit the critical region to 1-2 cM, the equivalent of 1-2 Mb in physical map terms. Once linked, the disease gene region must be defined by identifying flanking markers and determining the order of genetic loci within the region. Sequence tagged sites (STSs; 1.3.3.1) can be used in the second step, as anchor points for the construction of a contig of overlapping cloned genomic fragments that reflect the genomic organisation of this area and physically link the flanking markers. Contig assembly is fortified with additional STSs which are produced from the clones, and by fingerprinting analysis (section 3.1.4). The contig is a powerful resource with which to determine the physical distance of the disease region and to identify expressed sequences residing therein.

3.1.1 Genetic order of markers around the RP9 disease region.

Following genetic linkage of RP9 to microsatellites D7S435 and D7S460 (section 1.6.1; Inglehearn et al. 1993), new markers from a second generation microsatellite genome map (Weissenbach et al. 1992) were used in the construction of a genetic map of the region, and to identify flanking markers which define the disease interval. The genetic distance and order between the markers were determined by the analysis of recombinant individuals in several families (see figure 3.1; Inglehearn et al. 1994; Keen et al. 1995a). However, the relative positions of some markers could not be established genetically. Multipoint analysis of adRP with these markers placed the disease in an interval between D7S526 (distal) and D7S484 (proximal) with a peak lod score of 17.8 (Inglehearn et al. 1994). The interval was estimated at approximately 4 cM by Weissenbach and co-workers, while estimates based on segregation analysis in the RP9 linked family and other families suggested that interval may be 1.6 cM.
Figure 3.1 Locations of genetic markers in and around the RP9 disease region (adapted from Keen et al. 1995a). Centimorgan distance between the markers are shown on the left. Those markers on the right hand side could not be uniquely placed genetically with respect to other markers. Markers in the bracket show no recombination with each other and their physical order was unknown prior to contig assembly. D7S526 and D7S484 were the original flanking markers for RP9 (Inglehearn et al. 1994). The disease region was later refined between D7S795 and D7S484 (Keen et al. 1995a). In this study, contig assembly was initiated at the distal region with markers D7S526, D7S632 and D7S435.
No obvious candidate genes had been located to the RP9 region and the disease interval was now felt to be sufficiently small to commence physical mapping. The work presented in this study concentrated on contig assembly and chromosome walking from the distal end of the RP9 critical region, and was initiated at D7S526 and D7S435. Chromosome walking from the proximal end, into the critical region, was performed simultaneously by a co-worker. Midway into this study, the region was genetically refined even further, between D7S795 (distal) and D7S484 (proximal). New markers obtained from detailed genetic maps were ordered along the region, providing new STS points for contig assembly (see figure 3.1).

3.1.2 Yeast artificial chromosomes: a bridge from genetic to physical maps.

The small cloning capacity of many phage and cosmid vectors was for many years the limiting factor in the transition from genetic to physical mapping. This is due to the immense time and effort that is required to assemble a full contig in a region that spans only a few megabases, because numerous phage or cosmid clones are required. The development of the yeast artificial chromosome (YAC) cloning system (Burke et al. 1987; section 1.3.3.3) overcame this problem, and to date is the most commonly employed vector for assembling contigs over of large genomic intervals. YAC vectors are capable of cloning DNA fragments in the range of 0.1 to up to 2.0 Mb, which allow long-range continuous coverage over extensive regions. These were therefore the choice of clones for constructing a contig that spans the RP9 interval. However, the large clonal ability of these vectors has been accompanied by the problem of instability. Chimerism of insert DNA is not uncommon and rearrangements or deletions can occur during amplification. Nevertheless, tests have indicated that the integrity of the DNA is generally stable (Little et al. 1992). The pYAC4 vector has been the preferred vector used in the production of the main human genomic YAC libraries (section 3.1.3), which are employed in most physical mapping ventures.
3.1.3 Human YAC libraries.

Construction of YAC libraries is often difficult and labor intensive. Several laboratories have constructed YAC libraries for communal use, to serve the increasing number of positional cloning projects being pursued. These shared resources are beneficial as information from the analysis of these clones, by different scientific groups, can be gathered and compared to produce integrated genetic and physical maps for individual chromosomes and the entire genome (Chumakov et al. 1992, Cohen et al. 1993, Chumakov et al. 1995; section 1.3.3.4). These libraries are organised into a hierarchical pooling or arrangement system for easier screening using either PCR or hybridisation. Hybridisation screening utilises filters with a high density array of YAC clones which are gridded onto them by a robotic device. In PCR screening, clones are pooled into organised groups, which can identify individual positive YACs by an address that corresponds to a microtitre plate position (Green and Olson 1990b). PCR screening is generally more rapid with only a few pools to screen, but by hybridisation all the positive clones can be identified at once.

The CEPH library (Albertsen et al. 1990, and Chumakov et al. 1992) represents 7-8 times coverage of the human genome and has an average insert size of 920 kb. However, the large insert size results in an increased amount of chimerism (approximately 40-60%). The ICRF library (Larin et al. 1991) combines three different libraries which are made from cell lines 48XXXX, 49XYYYY and 46XX. Its average insert size is 600 kb. The ICI library (Anand et al. 1990) represents 3.5 times coverage of the human genome and has an average insert size of 350 kb. All the above libraries have been organised into a pooling system for PCR screening and are available freely from the Human Genome Mapping Centre (Hinxton, Cambridgeshire).

The construction of a physical map around the RP9 region was greatly augmented by the chromosome 7 YAC resource. This is a highly enriched collection of chromosome 7 YACs from three sources (Green et al. 1994 and Green et al. 1995). The first, is the Hybrid Cell Line library, which was made especially for this resource, and was constructed from DNA of the monochromosomal human-hamster hybrid cell line GM10791. The average YAC size is approximately 310 kb, and it had been estimated that 5-15% of the YACs are chimeric based on data generated by PFGE, FISH analysis and YAC end clone STS mapping. The second, were chromosome 7
specific YACs that had been isolated by PCR screening the Washington University total human genomic DNA library (Brownstein et al. 1989). YACs from this library have an average YAC size of approximately 250-300 kb and 40-60% of these YACs are estimated to be chimeric. The remaining were chromosome 7 specific clones from the CEPH YAC library (Albertsen et al. 1990), that had been identified by chromosome 7 specific Alu-PCR-based hybridisation analysis. Some clones in the resource have been allocated a YAC identity (which are referred to in this study) separate from their individual plate addresses.

3.1.4 Contig assembly.

Contig assembly relies upon identifying overlapping clones across a region by linking them with one or more reference points, and bridging gaps by isolating new clones that extend further. The strategy depends on the density of available markers within the region and the size and complexity of the library screened. Genetic markers serve as anchors for identifying seed YACs as starting material for the development of large cloned contigs. These markers also provide positional information for the order and orientation of other STSs which are placed on the YAC contig and is a means to integrate physical with genetic maps. To save time-consuming screening for clones, it is beneficial to exhaust all the resources available to the scientific community. YACs can be identified which have already been mapped to the region of interest and have been placed on an integrated map (Cohen et al. 1993, Chumakov et al. 1995).

However, contig assembly is rarely completed with these clones and it is usually necessary to augment the map with other clones by screening the publicly available libraries and to facilitate the ordering of markers across the region.

One approach to determine overlaps between clones, which was employed here, is STS content mapping (Kere et al. 1992). This is where PCR or hybridisation is used to determine the presence of unique markers or probes in clones. Overlapping clones are identified when an STS is present in both of them, and the order of STS markers can be determined by deciphering a pattern from which clones they accommodate. The PCR-based approach is more favourable as it is faster and only the primer sets have to be unique and specific. In hybridisation based mapping, if a probe includes repeats such as microsatellites, it can hinder analysis. The STSs on the other
hand can be polymorphic probes, cDNA probes or others generated from sequences in the region and must be site specific. At the start of contig construction in the RP9 region the density of STSs were low, mostly including poly (CA), markers which had been ordered by genetic analysis.

A second approach to determine overlaps between clones is by fingerprint analysis. This method is usually performed in probe-poor regions or when handling a large number of clones. Unpurified YAC DNA can be digested with one or more restriction enzymes, then separated by Pulse Field Gel Electrophoresis (PFGE), and common bands are identified by Southern blot analysis using a human specific repeat probe (e.g. Alu or LINE-1 DNA) (Bellanne-Chantelot et al. 1992). A more rapid alternative is to generate a fingerprint by inter-Alu PCR (Nelson et al. 1989). Alu repeats represent the major family of short interspersed repeats (SINE) in mammalian genomes and are spaced at an average of one every 3-6 kb (Britten et al. 1988, Moyzis et al. 1989). Human specific oligonucleotides designed from the terminal ends of the repeat and which direct away from it, are used to amplify inter-Alu products in the YAC clones, providing the distance between the adjacent repeats are amenable to PCR. The samples are fractionated by gel electrophoresis, producing a fingerprint consisting of PCR products of characteristic length. The orientation of a proportion of the repeats are inverted, which therefore allows the use of single primers to produce different fingerprints. Inter-Alu PCR products can also be used as a source for the production of additional probes or STSs from less well characterised regions of the genome (Nelson et al. 1991 and Cole et al. 1991).

In cases where there is an insufficient number of probes to assemble complete contigs directly, terminal sequences from a clone can be isolated and used as a probe to screen a library for the identification and isolation of overlapping clones. This approach, termed chromosome walking, is usually employed to bridge gaps between anchored clones that are physically separated. The walk is usually performed bi-directionally unless the orientation of the clone is known. Many methods for the isolation of end sequences have been reported. The recovered end clone products can be used directly as probes, or are more commonly sequenced to generate PCR-based STSs to isolate overlapping YACs.
BACs, PACs and cosmids can be incorporated into the contig at a later stage, which can be used for resolving STS order and to provide a more manageable and stable source for further manipulations. The size of individual YACs can be determined by Pulse Field Gel Electrophoresis (PFGE) (Schwartz and Cantor 1984). This method can be used in a range of further manipulations to characterise YAC clones.

3.1.5 Aims.

At the commencement of this study, the RP9 disease interval had been defined by the flanking markers D7S484 and D7S526 (Inglehearn et al. 1994), both of which exhibited one crossover with the disease allele from the affected members of the adRP7 family. The construction of a complete contig consisting of ordered overlapping clones was the next stage in the positional cloning of the disease gene. This study describes contig assembly in the distal part of the region. The first step was to order and assign YACs with STSs to create a contig. Secondly, known genes and ESTs were mapped and ordered onto the contig to assemble a transcriptional map and identify any genes that may be candidates for RP9. The contig would also provide an estimation of the physical distance of the region. Most importantly it would provide a valuable resource to identify genes residing in this area.

3.2 Results

Contig assembly was a continuous process and as new PCR-based markers or STSs were assigned to the region, they were incorporated into the existing map. New YAC clones were isolated to increase clone redundancy at critical points and to help resolve the STS order. Both Alu fingerprinting and STS content mapping strategies were applied, as they provide a rapid method of determining overlaps between clones. The results are presented as a non-chronological summary of the work. The resulting contig provided a useful resource both to compare the genetic and physical distances and as a foundation for a transcript map of the 7p14-15 region.
3.2.1 Incorporation of seed YACs.

At the onset of contig assembly a large number of YACs, which map to the 7p14-p15 region, were donated by Dr. E. Green (NHGRI at Washington University). These were isolated from the chromosome 7 YAC resource (section 3.1.3; Green et al. 1995) by PCR assay with known microsatellite markers from the region of interest (Gyapay et al. 1994). They had been isolated from large scale screening efforts and had not been characterised at all upon receipt. YAC DNA in solution and plug form were prepared for each clone (section 2.1.4 and 2.1.5). An initial PCR assay was performed to ensure that the YACs were positive for the original STSs with which they were isolated. Only four out of ten YACs that were assumed to contain D7S632 were truly positive for this STS. This was later confirmed in a subsequent publication (Green et al. 1994). All other YACs were verified positive for the STSs that they were isolated with (see figure 3.2).

Some clones were sized by Pulse Field Gel Electrophoresis (PFGE) and the YAC insert DNA was identified from the yeast chromosome background by hybridising with Cot-1 DNA (sections 2.3.2, 2.4.1 and 2.6.1). Two clones of ~1000 kb and ~ 900 kb were present when yWSS3263 was analysed, which may represent two unrelated clones or one clone coexisting with a deleted version. All clones were also tested for the presence of adjacent polymorphic STSs to detect any physical links by PCR-based STS content mapping. No overlap between D7S690, D7S526/D7S632 or D7S435 containing YACs was detected, hence presenting three apparently unconnected contigs (see figure 3.2).

3.2.2 Alu-PCR

At this stage in the analysis, the density of STSs available in the region was poor. The Alu-PCR approach (section 3.1.4, Nelson et al. 1991) was therefore adopted to generate comparable fingerprints, which could detect connections between these individual contigs and determine the degree of overlap between redundant clones.
Figure 3.2  Illustration depicting the seed clones from the chromosome 7 YAC resource (Green et al. 1994 and 1995) producing three apparently unconnected contigs in the 7p14-15 region. The STSs used to isolate them are cited at the top. Sizes are given where determined. Horizontal lines depict YACs, and only one is represented for the D7S690 cluster. Those denoted with # were subsequently used in Alu-PCR fingerprint analysis (section 3.2.2). *Indicates more than one YAC present when analysed by PFGE (see section 3.2.1). YACs for the proximal markers are not illustrated. HCL-Hybrid cell line library; WU-Washington University library; CEPH-CEPH mega YAC library.
3.2.2.1 *Alu*-PCR with D7S526, D7S632 and D7S435 YACs.

D7S526 and D7S632 are physically linked by YAC yWSS1545 and yWSS922 and their genetic order and orientation was known (figure 3.1 and 3.2). *Alu*-PCR was used to determine the extent of overlap between these clones and to detect any links with D7S435 or D7S690 YACs. The *Alu*-specific primers utilised in this study were ALE1 and ALE3 (Cole *et al.* 1991; section 2.14.8) which are designed from generally conserved regions from the terminal portions of the repeat. Amplification is directed out from the ends of the repeat units. PCR was performed using both ALE1 and ALE3 primers in a total volume of 25 μl on YAC DNA solution preparations. This was size fractionated through a 3% agarose gel alongside a size marker. Control PCRs using total yeast DNA and no DNA templates were also performed.

PCR with ALE1 and ALE3 revealed three bands in common with clones yWSS1545, 922 and 2458 (Figure 3.3). The 1.8 kb product was also shared by yWSS2056, which only contains D7S526. No overlaps were seen with YACs containing D7S435.

The D7S435 containing YACs yWSS29 and 2057 shared three bands (figure 3.3). The fingerprint suggested that yWSS29 is a larger clone as it produced additional unshared products. Alternatively these could have resulted from chimeric ends. *Alu*-PCR using single ALE primers did not give clear fingerprints and shared bands were hard to distinguish, therefore PCR was also performed with an end labeled primer and the samples were separated on a 6% acrylamide denaturing gel for better resolution (data not shown). This did not provide any other significant information. *Alu*-PCR between YACs containing D7S526 and D7S690 failed to show overlaps.
Figure 3.3 Alu-PCR fingerprint generated by both the ALE1 and ALE3 primers on D7S526, D7S632 and D7S435 containing YACs. The products were separated on a 3% agarose gel. White marks indicate shared bands between YACs. The STSs contained in the YACs are noted above. No overlap was identified between D7S526/632 and D7S435 containing YACs.
3.2.2.2 *Alu*-PCR with D7S690 YACs.

*Alu*-PCR was performed on a set of 14 redundant YACs (figure 3.2) that contain D7S690 to establish which YACs extended outwards from the cluster. This information would facilitate generation of a new STS with which to potentially link up with either D7S526 or D7S484/497/460 containing clones (in the proximal region) by chromosome walking (section 3.1.4). Different *Alu*-PCR fingerprints were generated by using both ALE1 and ALE3 primers together and each primer individually. When analysing the fingerprints generated, it was possible guess the amount of overlap between clones and to determine which extends the furthest. However, bands produced from chimeric or rearranged inserts can be misleading. Only the most prominent bands were scored to avoid confusion from less distinct bands.

The ALE1 primer only produced a simple fingerprint of four distinct bands and the level of ghost background was low (figure 3.4). A band of ~850 bp was present in all but clones yWSS291 and 2597. yWSS291 did not share any bands with the other YACs. A faint double band of ~820 bp and 800 bp was common in clones yWSS3215 and yWSS2597 which indicates that these may extend further out in one direction. The latter clone possessed two additional bands which suggests it may extend even beyond this. This is confirmed by the later assembled STS-based YAC contig across the RP9 interval (Keen *et al.* 1995a; figure 3.16).

Fingerprints generated from both ALE1 plus ALE3 primers were more complex. The number of bands had increased and were not of equal intensity in all YACs, which made scoring of bands more difficult. yWSS291 does share bands with this fingerprint, but no new products are seen in clones yWSS3215 and yWSS2597 (figure 3.5a). It is possible that there is a PCR bias in preferentially amplifying smaller inter-*Alu* products rather than longer ones, also a large amount of non-specific bands are evident.

Fingerprints generated from the ALE3 primer alone generated at least eighteen products of varying intensity in different YACs. A lack of overlapping bands was seen in YACs yWSS1474, 1787, 4037, 294, 293 and 292 (figure 3.5b). When comparing this with both the single ALE1 and the ALE1 plus ALE3 fingerprints, the relatively few bands shared among these fingerprints indicated that these YACs are small. Extra less intense bands of ~850 bp and 700 bp are seen in yWSS3215 and yWSS2597,
Physical Mapping of the RP9 Critical Region

**Figure 3.4** *Alu*-PCR fingerprint generated by the ALE1 primer only, on D7S690 containing YACs. The ~820 bp and ~800 bp faint bands generated in YACs yWSS3215 and yWSS2597 suggests that these YACs may extend out further in one direction. yWSS2597 contains two additional bands at ~1,100 bp and ~950 bp, indicating this may extend even further. This was confirmed by STS mapping (Keen *et al.* 1995a and figure 3.16). The ~600 bp *Alu* fragment was isolated to generate an STS for chromosome walking.
Physical Mapping of the RP9 Critical Region

Figure 3.5 Alu-PCR fingerprints generated by a) both ALE1 and ALE3 and b) ALE3 only primers on D7S690 containing YACs. YACs yWSS1474, 1787, 4037, 294, 293, and 292 produced fewer bands in comparison to the others and are probably smaller in size. YACs yWSS3215 and 2597 generated additional products (~850 bp and ~750 bp) in the ALE3 only fingerprint. yWSS298 also produced extra bands (~350 bp and ~310 bp). STS content mapping confirmed that these clones do extend further from the others.
suggesting that these extend to one side of the cluster. Interestingly, this is supported by STS content mapping analysis which was performed later (Keen et al. 1995a and see figure 3.16) YAC yWSS298 produced less intense extra double bands of ~310 bp and 350 bp. The subsequent STS content mapping data also confirmed that this clone extends distal to the others.

### 3.2.3 Development of a novel STS, A1.

A ~600 bp Alu fragment from the single ALE1 fingerprint (figure 3.4) was chosen to generate an STS that may join these clones to individual contigs on either side. This fingerprint generated distinct products that could be purified easily. The chosen product amplified from only four of the fourteen clones, and appeared to extend out from the YAC cluster. The extra products from clones yWSS3215 and 2597 were too indistinct and too close to other bands to isolate and purify. The ~600 bp fragment was isolated from the gel and re-amplified with the ALE1 primer alone to yield this product only. After confirmation on an agarose gel, a sample of the reaction was used to clone into the pCRII vector (Invitrogen, sections 2.8.2 and 2.8.3). Several white recombinant colonies were directly checked for the correct insert by PCR amplification. Two clones were chosen and DNA was prepared from these for sequencing (section 2.9.1). The SP6 primer from the pCRII vector gave ~300 bp of reliable sequence from each clone (figure 3.6 and 3.7). The sequence displayed no significant homology to sequences in Genbank, EMBL and Repeat databases when searched using the FASTA program (section 2.13). A 136 bp STS named A1 was designed and tested on the D7S690 containing YACs. This STS was confirmed positive for only those clones which amplified the ~600 bp fragment from ALE1 only fingerprints. It did not amplify in YACs from an unattached contig of D7S460 containing clones (clones that were obtained from the chromosome 7 YAC resource), or in YACs from the D7S526 cluster. However, it was later found to be present on a CEPH “mega” YAC, which provided a physical link to the D7S460 containing YACs and orientated the clones from the D7S690 contig (see section 3.2.6 and figure 3.13).
Figure 5.6 Autoradiograph showing the partial sequence of an Alu-PCR product that derived from D7S690 containing YAC clones. The product was cloned into a pCRII vector for sequencing. An STS (A1) was chosen for content mapping. The position of A1 forward and reverse oligonucleotides in this sequence are shown above.
**Figure 3.7** Partial sequence of the ~600 bp *Alu*-PCR product from D7S690 containing clones. A 136 bp STS was chosen (A1). Sequences labelled in red and green represent the forward and reverse primers respectively.
3.2.4 Isolation of additional YACs.

As Alu-PCR had not identified any overlaps between D7S690, D7S632/D7S526 and D7S435 containing YACs (see figure 3.2), additional YACs were isolated by PCR screening the ICI library, which was readily available in the laboratory. The ICI library (Anand et al. 1990) is ordered into a pooled system with 40 primary, 9 secondary and 20 tertiary pools, where the combination of positive signals identifies a clones microtitre plate, column and row YAC address (section 2.7.4). Human DNA, no DNA and yeast DNA controls were included. Clones were supplied in agar stabs from the HGMP resource centre (Hinxton, Cambridge). Single colonies were amplified in selective media and DNA in solution and plug form were prepared for each clone. Each clone was tested by PCR assay to directly confirm the identity of the clone and was then assessed for size and purity by PFGE.

Screening of the primary pools with an STS for D7S632 (AFM198ze5, GDB: G00-199-168) identified two positive signals. However, secondary and tertiary screenings resulted in identification of only clone, 33B-E5 (~250 kb). An STS for D7S526 was also used to screen this library (AFM248vc9, GDB: G00-188-526). This identified clone 2F-F10. However upon inspection, this clone proved to be negative for D7S526. Screening efforts which resulted in additional incomplete YAC addresses were not resolved due to the subsequent availability of CEPH clones mapping to the region.

As new microsatellite markers were generated by the scientific community, those that were assigned to the 7p14-15 region were incorporated into the genetic map (section 3.1.1 and figure 3.1). D7S1514 and D7S474 could not be placed genetically with respect to D7S526 and D7S690 (figure 3.1) but were both found to be present in yWSS298. An STS for D7S1514 (GDB: G00-314-801) was used to screen the ICI library. This resulted in the identification of clones 27I-A5 (460 kb), 34D-A4 (365 kb) and 34D-A5 (460 kb). 34D-A4 and 34D-A5 were isolated from the same positive primary and secondary screenings, but produced different tertiary signals.

EST00601 (D7S552E) is one of thirty ESTs that were assigned to sub-regions of chromosome 7 by PCR analysis (chapter 4). After mapping to the 7p14-15 area it was detected in clone yWSS2458 only, placing it between D7S632 and D7S435. This EST was therefore used to isolate more clones to join and extend the individual contigs.
of D7S632 and D7S435. Clones 30H-A8 (285 kb) and 6B-F3 (~200 kb) were identified as positives from complete screenings with EST00601. All clones (listed in table 3.1) were tested for adjacent markers by PCR-based STS content mapping (see section 3.2.6). Selected clones were chosen to isolate YAC terminal ends to establish further links (section 3.2.5).

As integrated genetic and physical maps became available (Cohen et al. 1993 and Chumakov et al. 1995), databases were searched for clones which contained markers that are in the RP9 critical region. CEPH YACs 959_B_3, 908_D_2 and 823_C_11 were originally found to contain both D7S526 and D7S632. These clones were obtained from the HGMP resource centre and DNA was prepared. Upon STS confirmation, 823_C_11 was found negative for D7S526, suggesting this clone might be deleted. These YACs were sized by PFGE to verify clone integrity. 823_C_11 revealed the presence of three clones of ~365, 500 and 610 kb. This clone was regrown from single colonies, prepared and sized again in case contamination had occurred during growth, but PFGE analysis revealed the same number of additional YACs. The CEPH database indicates that 823_C_11 has a 1220 kb insert, so the multiple artificial chromosomes found here may reflect deleted versions of the original YAC. Nevertheless, this was also incorporated into the RP9 contig and subjected to PCR-based STS-content mapping (section 3.2.6). Figure 3.8 shows PFGE results for a selection of these clones.
Figure 3.8 a) Size estimation of a selection of YACs by Pulse Field Gel Electrophoresis (PFGE). b) Southern blot hybridisation with Cot-1 DNA to identify the YACs which are segregating with the yeast chromosomes. These YACs were isolated by screening a YAC library using PCR and were subsequently incorporated into the distal contig. 823_C_11 contains three YACs of ~365, 500 and 610 kb. YACs 6B-F3, 33B-E5 and the 500 kb YAC of 823_C_11 are clearly visible on the agarose gel. This gel was run at conditions which favour the separation of the smaller yeast chromosomes.
Table 3.1
Identity of additional YACs isolated or obtained from databases and subsequently incorporated into the contig. * Indicates those clones identified by PCR screening in this study.

<table>
<thead>
<tr>
<th>YAC identity</th>
<th>STS used in isolation of YAC</th>
<th>Approximate Size</th>
<th>Source library</th>
</tr>
</thead>
<tbody>
<tr>
<td>823_C_11</td>
<td>D7S632 &amp; D7S526</td>
<td>365, 500 and 610 kb</td>
<td>CEPH</td>
</tr>
<tr>
<td>959_B_3</td>
<td>D7S632 &amp; D7S526</td>
<td>1500 kb</td>
<td>CEPH</td>
</tr>
<tr>
<td>908_D_2</td>
<td>D7S526 &amp; D7S632</td>
<td>1200 kb</td>
<td>CEPH</td>
</tr>
<tr>
<td>33B-E5*</td>
<td>D7S632</td>
<td>250 kb</td>
<td>ICI</td>
</tr>
<tr>
<td>27I-A5*</td>
<td>D7S1514</td>
<td>460 kb</td>
<td>ICI</td>
</tr>
<tr>
<td>34D-A4*</td>
<td>D7S1514</td>
<td>365 kb</td>
<td>ICI</td>
</tr>
<tr>
<td>34D-A5*</td>
<td>D7S1514</td>
<td>460 kb</td>
<td>ICI</td>
</tr>
<tr>
<td>30H-A8*</td>
<td>EST00601</td>
<td>285 kb</td>
<td>ICI</td>
</tr>
<tr>
<td>6B-F3*</td>
<td>EST00601</td>
<td>200 kb</td>
<td>ICI</td>
</tr>
</tbody>
</table>

3.2.5 Isolation of YAC terminal sequences: Alu-vector arm PCR.

This procedure takes advantage of the frequent distribution of Alu repeats in the genome and relies on the presence of a repeat unit within a suitable distance from the pYAC4 cloning site to allow amplification of the insert end by PCR (Nelson et al. 1991). It was performed by using a combination of one of the Alu specific primers against either the left or right arm pYAC4 oligonucleotides. In parallel, PCR with just the single Alu primer was also done and separated on an agarose gel adjacent to the Alu-vector products (section 2.7.3). Comparison of the resultant banding patterns allowed identification of unique products which are only generated in the Alu-vector arm sample and which therefore represent the terminal ends of the YAC clone plus part of the pYAC4 vector sequence that is adjacent to the primer binding site. Only insert-vector junction products of >300 bp, inclusive of the vector sequence, were chosen for sequencing to provide an adequate insert length for choosing an STS. The sequences generated were first compared against the Genbank, EMBL and Repeat databases using the BLAST programs (section 2.13; Altschul et al. 1990) to detect any similarities with known sequences. Oligonucleotides were chosen to generate
unique products of 100-300 bp.

yWSS1545 was initially chosen for isolation of end sequences in the hope of designing an STS that would join the D7S690 and D7S526 containing YACs (see figure 3.2). *Alu*-vector arm PCR with ALE1 and a pYAC4 right arm primer (pYAC4-RA(a); see section 2.14.8 for primer sequences) resulted in the production of a unique product of ~185 bp. This was not analysed further as 170 bp of this was pYAC4 sequence after the primer binding site. Amplification with ALE3 and a pYAC4 left arm primer (pYAC4LA-(a)) resulted in a unique band of ~500 bp (figure 3.9a). The pYAC4 left arm primer is 285 bp from the *Eco* R1 cloning site, thereby giving ~200 bp of YAC insert sequence. This fragment was isolated from the additional *Alu* products by re-amplifying this product after isolating the band from an agarose gel. After confirmation of the single product on an agarose gel, the remaining PCR reaction was purified through an S-400 column and used as a template for direct sequencing (sections 2.1.2.3, 2.9.2 and 2.9.3). As this proved unsuccessful, the product was cloned into the pTAg vector (R&D systems) for cycle sequencing. Two clones were chosen and were successfully sequenced from both ends. Database comparisons against the sequence revealed no significant homology to known sequences. Figure 3.9b presents the sequence that was obtained, from which a 116 bp STS, 1545LA, was generated (section 2.14.8).

*Alu*-vector PCR was also attempted on clones 27I-A5 and 34D-A5 to generate an STS that could link the D7S690 contig with the D7S526 contig. PCR with the right arm vector primer and the ALE3 primer on clone 34D-A5 resulted in a unique fragment of ~475 bp, 170 bp of which is vector sequence. This product was isolated as described above. 282 bp insert end sequence was obtained by direct automated fluorescent sequencing using a nested internal right arm primer (pYAC4RA-(b); 2.9.4 and 2.14.8). Database comparison of the YAC insert sequence revealed no significant homology to known sequences. A 223 bp STS (34D-A5RA) was designed from this (section 2.14.8; figure 3.10).

PCR using a left arm vector primer and the ALE3 primer on 27I-A5 gave a unique product of ~900 bp. Isolation and direct sequencing using a nested left arm primer (pYAC4LA-(b)) generated 246 bp of insert sequence which was partially homologous to the L1 repeat. However a 136 bp unique STS (27I-A5LA; section
Figure 3.9  

a) *Alu*-vector PCR for the isolation of the YAC end clone sequence from YWSS1545. A unique product of ~500 bp was generated using the pYAC4 left arm (LAa) and ALE3 primer combination. This product was cloned into pTAG for sequencing. b) Nucleotide sequence generated from the left arm of YWSS1545. The YAC end insert sequence was isolated and a 116 bp novel STS (1545LA) was chosen. Sequences labelled in red and green represent the forward and reverse primers respectively.
Figure 3.10  Sequence generated from the right arm of 34D-A5. The end clone was isolated by Alu-vector PCR and the product was directly sequenced. A 223 bp STS was chosen. Sequences labelled in red and green represent the forward and reverse primers respectively.
2.14.8) could still be selected (figure 3.11).

Both 34D-A5RA and 27I-A5LA successfully amplified from their respective clones. However neither amplified on redundant and overlapping clones in the vicinity, suggesting that these clones have chimeric termini. To determine their true localisation, the STSs were mapped by PCR in the HGMP monochromosomal somatic cell hybrid panel (Kelsall et al. 1995). This revealed that 27I-A5LA is actually located on chromosome 9, while 34D-A5RA in fact maps to chromosome 12 (figure 3.12).

3.2.6 STS content mapping and development of a transcript map.

PCR assays were performed on solution DNA. All clones were tested for the presence or absence of STS markers mapping to the region. This enabled clones to be ordered into overlapping arrays which can physically connect STSs. The search for genes mapping to the region was an ongoing process. Newly assigned chromosome 7 genes were tested for their presence onto these clones. Those that were found to be positive were incorporated into the contig as gene-based STSs, to assemble a transcript map in the region. The completed contig showing the physical locations of these genes and markers is illustrated in figure 3.13.

The Alu-PCR generated STS, A1 (section 3.2.3), was found to be present on the D7S690 containing YACs yWSS3043, 3255, 3215 and 2597. It was also found on YACs yWSS2550 and 961-F-5, part of a proximal contig assembled by a co-worker. The latter clone contains D7S656 and D7S683, allowing the D7S690 containing YACs to be orientated relative to other genetic markers.

D7S795 was present on all D7S690 containing YACs. The physical order of these markers is based on STS content mapping by Washington University (E. Green, personal communication). D7S1514 was found to be present in the D7S690 containing YAC yWSS298 and CEPH clones 959_B_3, 908_D_2 and 823_C_11, therefore immediately providing a physical link between the genetic markers D7S690 and D7S526.

D7S474 was present on clones yWSS298, 959_B_3 and 908_D_2. This places it proximal to D7S1514 as it not present on 823_C_11. However, this location
**Figure 3.11**  Sequence generated from the left arm of 271-A5. The end clone was isolated from Alu-vector PCR and the product was sequenced directly. A 136 bp novel STS was chosen. Sequences labelled in red and green represent the forward and reverse primers respectively.
| Hybrid | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | X | Y |
| GM10611 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| GM10628G | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| GM10695 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| HHH118 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| GM106914 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| HCM0161A | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| CL2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| GM10061 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| SI2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| JEL4 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| AJAM0062+ | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| 289 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| GM10479 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| HCM61L1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| 2409H13 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| PUTRI.8 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| HCM870 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| GM10612 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| GM1072A | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| TVY11 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| T295F2X5X1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| HOX10 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| 853 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |

* indicates presence; - indicates absence; / indicates chromosome translocation, extra chromosomes or other modifications.

**Figure 3.12** The HGMP monochromosomal somatic cell hybrid panel (Kelsell *et al.* 1995). The individual hybrid name and human content are listed in the table. PCR analysis of the 271-A5LA STS (136 bp) on cell lines of the panel shows amplification from hybrid GM10611 only. This contains chromosome 9 as its only human component. Similarly, 34D-ASRA was found to map to chromosome 12.
Figure 3.13 The YAC contig which extends distal from the RP9 critical region. The disease region was originally defined by the distal flanking marker D7S526 and D7S484. The critical region was subsequently refined between D7S795 and D7S484, therefore excluding this area. The horizontal thin lines depict the YAC clones with their identity and sizes above. Only a few D7S690 YACs are illustrated. Dashed horizontal lines indicate regions of internal deletions in YACs. The dashed vertical lines represent the positions that the STSs occupy in the YACs. Thick blocks depict insert ends which were isolated in this study. Those from clones 271-A5 and 34D-A5 are from chimeric portions of the clone.
is considered tentative as 823_C_11 was found to be deleted and unstable, therefore D7S474 could still be indistinguishable from D7S1514.

DRES34 (Drosophila related EST 34) is one of the sixty-six human ESTs which were reported to have exhibited sequence homology to Drosophila genes that cause mutant phenotypes (Banfi et al. 1996). This EST showed certain similarity to the Drosophila atonal (ato) gene, a proneural gene for olfactory sensilla which encodes a transcription factor. However, upon examination, the level of identity between these transcripts is too low to suggest that it is a true homologue of atonal. The sequence identity was limited to a portion of the EST, possibly due to the presence of a conserved domain. Nevertheless it represents a novel human transcript. DRES34 was located to 7p15, by radiation hybrid mapping and FISH, by these investigators. An STS for this (GDB: 3906429) was used in PCR analysis in this study and located it to clones yWSS298, 959_B-3 and 908_D_2, placing it close to D7S474.

ADCYAP1R is one of three types of receptors that mediates the actions of the pituitary adenylate cyclase activating polypeptide (PACAP) hormone, which affects secretion of many hormones e.g. growth hormone, adrenocorticotropin hormone and insulin. It is expressed in the hypothalamus, brain stem, pituitary, adrenal gland, pancreas and testis (Stoffel et al. 1994). The human ADCYAP1R gene was previously mapped to chromosome 7 by PCR on a panel of human-rodent somatic cell hybrids. Localisation to 7p14 was achieved by FISH and it was found to be present on the CEPH mega YAC 959_B_3 (Stoffel et al. 1994). The ADCYAP1R gene specific STS primers (GDB:456105) were tested on the other YACs in this region and was also found in clones 908_D_2 and 823_C_11. The position of this gene was deduced to lie between D7S1514 and D7S526 as it was not present on any YACs of the D7S526/632 contig, and the positive clone 908_D_2 does not extend further past this contig. This tentative placement depends on clone 908_D_2 being unaltered. Recent high density STS content mapping ventures (Hudson et al. 1995) has suggested that this clone is not deleted or rearranged. It also smaller in size than the other CEPH clones that overlap it considerably, and the size estimate determined in this study is in agreement with that determined by Chumakov et al. (1995).

The end clone STS 1545LA was found to be present on clones 959_B_3, 908_D_2 and 823_C_11. Its absence from clones yWSS922, 2056 and all those distal
Physical Mapping of the RP9 Critical Region

to it indicates it is located between ADCYAP1R and GHRHR. This STS did not link up with any clones from D7S1514, but it was not used to isolate further YACs as the RP9 disease region had subsequently been refined and the inclusion of the CEPH clones had already provided the necessary link. As an example of PCR-based STS content mapping, figure 3.14 presents the results for the mapping of A1 and 1545LA STSs.

The growth hormone releasing hormone (GHRH) plays an important role in the regulation of growth hormone synthesis and secretion, and acts via the GHRH receptor (GHRHR). The gene for the receptor was localised to 7p14 by in situ hybridisation (Gaylinn et al. 1994) and by PCR analysis of somatic cell hybrids (Vamvakopoulos et al. 1994). When tested on clones in this region, the STS for GHRHR (oligonucleotides: Pf and Pr, Vamvakopoulos et al. 1994) was found to be on YAC yWSS1545. Simultaneously a publication describing its location to this YAC was released (Wanjnrajch et al. 1994). This gene was later also found on 959_B_3 and 908_D_2, but not on 823_C_11. As clone 823_C_11 is known to be deleted, it can be assumed that GHRHR is therefore located between 1545LA and D7S526 (figure 3.13).

Aquaporin (AQP1) functions as a water-selective transmembrane channel, and is expressed in various secretory and absorptive epithelia, and is very abundant in the anterior segment of the eye. This gene was mapped to chromosome 7p14-15 by in situ hybridisation (Deen et al. 1994). STS content mapping placed it on clones yWSS2056, 1545 and 922 (EMBL Accession No: Z21985; Keen et al. 1995b). It was later also found in clones 959_B_3 and 908_D_2 but not in 823_C_11, thus further supporting the presence of a deletion in the latter clone. The localisation for AQP1 could not be distinguished physically from D7S526 by STS content mapping. However it is in a 375 kb vicinity around this marker, defined by its presence on clone yWSS2056 which contains only D7S526 and AQP1 (figure 3.13).

Glycyl-tRNA synthetase (GARS) is one of the aminoacyl-tRNA synthetases that catalyses the attachment of an amino acid to its respective tRNA. The GARS cDNA was assigned to chromosome 7 by PCR-based screening of somatic cell hybrid panels and was further localised to 7p15 by FISH (Nichols et al. 1995). The
Figure 3.14 STS content mapping results in a selection of YACs from the RP9 distal contig for A1 and 1545LA. YACs yWSS3029, 798 and 3E-G9 are proximal contig YACs (see figure 3.16).
STS for GARs (GDB: 3750172) was used to test a pool of YACs in the region and was then more specifically localised to YACs yWSS1545, 922, 2458, 33BE5, 959_B_3 and 823_C_11. This STS was negative for 908_D_2, which indicates the clone's boundary, thereby placing the gene distal to D7S632 but proximal to EST00601 (figure 3.13).

EST00601 (D7S552E) was first placed on yWSS2458 (section 4.2.5.1). Further analysis demonstrated it was also present in clones 33B-E5, 959_B_3 and 823_C_11. This positioned EST00601 between GARs and D7S435. It was not present in any D7S435 containing YACs, therefore leaving an uncloned break between these contigs. This could be resolved by producing further end clone STSs from EST00601 containing YACs to bridge this gap. However this was not pursued as the RP9 disease region had since been refined and the completion of this contig was not a priority. As the partial sequence for the cDNA clone for EST00601 did not demonstrate any homology to known sequences, the EST was tested by PCR amplification on a small set of tissue cDNAs (Clontech) to provide some idea of its expression. It was shown to be abundant in brain, less in liver and even lower in muscle (figure 3.15). This is by no means a quantitative measure, but suggests a pattern of its expression.

The location of these polymorphic STSs and genes on the YACs allows the determination of the maximum physical distances between them. The 1,500 kb YAC 959_B_3 physically connects 11 STSs, 6 of which are from transcripts. yWSS1545 places six STSs, three of which are genes, within 400 kb. D7S632, GARs and EST00601 are contained in both yWSS2458 (1300 kb) and 33B-E5 (250 kb). The isolation of the latter clone was therefore beneficial as it has enabled a more accurate estimation of the physical distance between these STSs. The incorporation of smaller YACs has also allowed the assignment and order of these STSs relative to one another. Five YACs had singular hits with just one STS. The D7S690 YACs were subjected to further STS content mapping with new markers by a co-worker (figure 3.16). The total contig presented here includes 21 YAC clones (excluding some redundant D7S690 containing clones and the D7S435 containing YACs) with six microsatellite markers, one end clone STS, one random STS and six gene-based STSs (figure 3.13).
Figure 3.15 Amplification of EST00601 (141 bp) from cDNAs which were derived from human brain, liver and muscle tissues (Clontech).
3.2.7 Estimation of the physical extent of the proximal RP9 region contig.

These data together with that of a colleague led to the creation of a well defined contig spanning the RP9 region and extending distally to it. The disease interval was also further refined during the course of this work between markers D7S484 and D7S795 and excluded all of the genes placed on the contig (Keen et al. 1995a). To determine the physical extent of the RP9 critical region the sizes of most of the clones comprising this contig were measured in this study. YAC DNA was prepared in plug form from colony purified clones in the region (section 2.1.5). These were sized by PFGE and YAC inserts were identified from yeast chromosomes by Southern blot and hybridisation analysis using a human Cot-1 DNA probe. Sizes were estimated by comparing the migration of the YAC clones with yeast host chromosomes of known size.

Selected clones were chosen to estimate the maximum and minimum extent of both the contig and the RP9 interval. The maximum size is defined by the shortest tiling path of overlapping YACs and the minimum size is defined by the total length of a set of non-overlapping YACs from the contig which do not quite extend to each end. This provides an approximation and assumes no chimerism, rearrangements or deletions. Clone sizes and tiling paths for the whole contig are illustrated in figure 3.16.

Whole contig.

Maximum tiling path.
Clones yWSS3011, 3586, 1564, 1397, 1211, 961_F_5, 298 and 959_B_3 define a 7,300 kb maximum size of the whole contig.

Minimum tiling path.
Clones yWSS3011, 1564, 3029, 298, 1545 and 33B-E5 define a 4,575 kb minimum size of the whole contig.
Figure 3.16  Selected clones from the contig that encompasses the RP9 region and extends out distally (Keen et al. 1995a). The sizes of clones were determined by PFGE and hybridisation with Cot-1 DNA. The estimated physical size can be determined by the average size of the two tiling paths, which has been estimated as 5,937 kb for the whole contig and 3,220 kb for the RP9 interval.

Represent clones forming the maximum tiling path of the whole contig (clones that cover the least distance with overlapping clones).

Represent clones forming the minimum tiling path of the whole contig (clones which span the interval but do not overlap).

Represent clones forming the maximum tiling path spanning the RP9 interval.

Represent clones forming the minimum tiling path spanning the RP9 interval.
RP9 interval.

Maximum tiling path.
YACs yWSS3586, 1564, 1397, 3029 and 2597 define a 3,840 kb maximum size estimation for the refined RP9 critical region.

Minimum tiling path.
Clones yWSS1564, 3029 and 2550 define a minimum size estimation of 2,600 kb for the RP9 interval.

Taking into consideration both the maximum and minimum sizes, this gives an average estimate for the whole contig as 6 Mb, and 3 Mb for the RP9 interval.

FISH analysis of certain YACs in the region has demonstrated that these lie on the 7p14-15 boundary. A D7S484 containing clone, yWSS3586, has been mapped to the 7p14-15 border, marking the proximal boundary for RP9. Also D7S632 and D7S526 containing clones yWSS922 and 1545, representing the distal region, were mapped to 7p15 (Green et al. 1994). 7p14 is a giesma positive “dark staining” band, indicative of a gene poor region, while 7p15 is giesma negative “light staining” band, indicative of a gene rich region (Bickmore and Sumner 1989). As yet, no genes have been identified which reside in the RP9 critical region, yet six transcripts lie immediately distal to this. It can therefore be suggested that the contig spanning the distance between D7S484 to EST00601 defines the 7p14-15 region.

3.3 Discussion.

At the initiation of this work, the RP9 critical interval was defined by the flanking markers D7S484 and D7S526, a region of approximately 1.6 cM (Inglehearn et al. 1993). Contig assembly was concentrated at the distal end of this area in this study, which was subsequently excluded from the RP9 disease region when it was refined (Keen et al. 1995a). Genetically ordered markers and gene-based STSs identified YACs to assemble an overlapping array of clones that cover this region and ordered these STSs relative to one another.

As demonstrated here, Alu-PCR provides a rapid alternative to the general fingerprinting procedure of restriction digest analysis of YAC DNA. It can be performed using smaller quantities of DNA samples and even directly on yeast
colonies. These fingerprints allowed determination of the measure of overlap between the clones in the initial unlinked contigs, and identified which extended furthest. Clones yWSS3215 and 2597 appeared to extend further out, as suggested by the presence of shared bands which were not present in other clones. Also yWSS298 produced a unique additional band, suggestive of further extension beyond the other YACs. This was actually confirmed by later STS mapping in this laboratory (figure 3.16) and in the recently published NHGRI chromosome 7 physical map (Bouffard et al. 1997b). However, Alu-PCR does pose problems such as producing fragments of unequal intensity between shared YACs. For example, smaller inter-Alu products are preferentially amplified due to PCR bias over larger ones, which causes confusion and can lead to discrepancies in analysis. When employing Alu-PCR to generate STSs for chromosome walking, it maybe important to choose products represented in at least two YACs, as a solitary band may derive from a chimeric portion of the YAC. Nevertheless, the Alu-PCR approach is a powerful method for the physical characterisation of sequences in clones, and has been adopted to generate clone specific probes to identify overlapping YACs in the construction of a whole genome YAC physical map (Cohen et al. 1993 and Chumakov et al. 1995).

Additional clones were isolated in the hope of bridging gaps between the unconnected individual contigs in the hope of providing a link to neighboring STSs. When this did not succeed, key YACs were chosen for the isolation of end clone termini to generate STSs for chromosome walking or to detect unrecognised overlaps. The end clone STS 1545LA did not connect with the D7S1514 containing YACs, and no new YACs were isolated using this STS, as CEPH clones which connected the separate contigs were then incorporated. However, this STS provides an additional landmark for further resolution of this region when assembling a contig of smaller clones. End clone sequences from D7S1514-containing YACs 271-A5 and 34D-A5 were both from different chromosomes, suggesting these YACs are chimeric. These clones are comparatively small, but nevertheless demonstrate the possibility of chimerism in any library. It is believed that chimeric clones arise either from co-ligation of two fragments of insert DNA or by co-transformation of two individual clones into one host cell followed by homologous recombination between insert sequences (Green et al. 1991). Chimeric clones hinder contig assembly when
Physical Mapping of the RP9 Critical Region

chromosome walking is necessary. STS content mapping is not affected by this, but is sensitive to internal deletions and rearrangements especially if clone redundancy in that region is low.

The order of some markers and all YACs on this contig are in general agreement with the recently assembled NHGRI chromosome 7 physical map and the Whitehead STS-based map (Bouffard et al. 1997b, Hudson et al. 1995). The only discrepancy is D7S1514, which has been placed by NHGRI proximal to its map position in this study. However the presence of this STS in the four clones that are known to span this region confirms its map position here (see figure 3.13). This STS may in fact amplify multiple duplicated sites, possibly accounting for the assumed chimerism of YACs which were isolated with D7S1514 in this study. The true localisation of D7S1514 can be determined by FISH. D7S474, DRES34, AQP1, GARS and EST00601 have not yet been placed relative to each other by NHGRI or any other groups and has been ordered here for the first time. The relative locations of ADCYAP1R1 and GHRHR between D7S690 and D7S526 were confirmed by NHGRI. The clear distinction between the number of gene transcripts that have been identified to date in the RP9 region and those distal to it suggests that this region is on the 7p14-15 boundary, a localisation supported by FISH analysis of a selection of YACs from the contig.

The inclusion of clones from the chromosome 7 YAC resource and the CEPH "mega" YACs (Green et al. 1994 and Chumakov et al. 1995) were of great value to this positional cloning project. Clone 959_B_3 is 1.5 Mb and although it was not used for ordering STSs, it encompasses the six gene-based STSs. This clone is likely to be a faithful copy of the genomic DNA, as high density STS content mapping results on this YAC show no signs of large deletions or rearrangements (Hudson et al. 1995). Also FISH analysis, which placed it on 7p15, gave no additional signals which suggests it is not chimeric (Chumakov et al. 1995 and Bray-Ward et al. 1996). This YAC will therefore prove valuable if further investigation of this region is required. Deletions found in 823_C_11 has also been reported by high density STS content mapping endeavors (Hudson et al. 1995).

D7S1514, D7S474 and DRES34 provided a physical link between the CEPH mega YACs and the D7S690 containing clones. The STS A1, designed from an
isolated Alu-PCR generated fragment, was instrumental in physically linking the D7S690 containing clones to 961_F_5, a CEPH clone containing three genetic markers that are non-recombinant with the disease. A gap remains between EST00601 and D7S435 but was not further resolved due to the refinement of the disease region. The total size estimate for the complete contig is approximately 6 Mb and for that spanning the refined RP9 disease region is approximately 3 Mb. The interval between D7S484 and D7S526 was originally estimated at 4 cM (equivalent to 4 Mb) by Weissenbach and colleagues, and at 1.6 cM (~1.6 Mb) by this laboratory. When comparing the ~6 Mb physical size estimate of the contig extending from D7S484 to D7S526, it suggests the estimated physical size is larger than the genetic size. The size determination of these clones were also later performed by NHGRI. When compared to the sizes estimated here, they were generally similar with only a few discrepancies that can be accounted for by the lack of resolution provided by standard PFGE settings for very large and very small clones. The physical extent of the RP9 critical region was established by estimating the total lengths of overlapping and non-overlapping clones. Although this gives some idea of the physical size of the region, the distance between individual STSs could be more precisely determined by restriction site mapping (section 1.3.3.2). By using rare-cutter restriction enzymes whose sites are often found in CpG island clusters, this method would also be a useful indicator for locations of gene sequences (Lindsay and Bird 1987).

A high density STS-content based physical map of chromosome 7 has recently been published (Bouffard et al. 1997a) and is the result of the large scale efforts of Dr. E. Green and colleagues. It is comprised of 2,150 STSs, which include genetic (Dib et al. 1996), random and EST markers (Bouffard et al. 1997b and Touchman et al. 1997), some of which have been FISH mapped. These have been mapped and ordered onto YACs from the chromosome 7 YAC resource and CEPH clones to build redundant sets of large contigs with an STS spacing of one every ~79 kb. This integrated physical map has circumvented the need, and eliminates the time that is taken for assembling contigs over large regions by individual groups and is especially a great benefit to any new positional cloning project on chromosome 7.

The contig described in this study was constructed prior to the recent high density physical map of chromosome 7 (Bouffard et al. 1997a). It has uniquely
ordered six gene transcripts relative to one another in a space of ~1.5 Mb between D7S1514 and D7S435. Some of these transcripts may be obvious candidates for certain disease phenotypes as information on their function has already been deciphered. GARS is one of 20 aminoacyl-tRNA synthetases which have been identified as targets of autoantibodies in the autoimmune disease polymyositis/dermatomyositis (Ge et al. 1994). GHRHR abnormalities have been implicated in severe growth hormone deficiency, which is analogous to the "little" (lit) mouse phenotype (Wajnrajch et al. 1996). Also AQPI has been proposed as a possible candidate for conditions that involve an imbalance in ocular fluid movement, due to its expression within the eye (Stamer et al. 1994). Although excluded as candidates for RP9, some of these genes are within the critical intervals for other positionally cloned diseases. Dominant cystoid macular dystrophy (CYMD, Kremer et al. 1994) has been mapped to the interval between D7S526 and D7S493, making AQPI, GARS and EST00601 positional candidates for this disease phenotype. Also the disease gene for dominant non-syndromic hearing loss (DFNA5) was located between D7S632 and D7S493 (Van Camp et al. 1995), making GARS and EST00601 positional candidates for this disorder.

In conclusion, the physical map presented here has provided a valuable resource for a variety of manipulations, for example in the identification of additional polymorphic markers (e.g. microsatellites) and random STSs to generate additional tools for mapping. Smaller clones can be produced from YACs by fragmentation analysis (Pavan et al. 1991). Also the recombination properties of the yeast host can be exploited to produce individual clones which contain large genes and their regulatory elements (e.g. as performed with the cystic fibrosis mutated gene CFTR, Green and Olson 1990a). At this stage, only YACs have been incorporated into the contig. This paves the way for a more higher resolution map constructed of smaller clones such as cosmids, BACs or PACs, which can be used for more refined STS ordering, or for use in gene identification procedures and ultimately, as a basis for sequencing.
Chapter 4. Regional Assignment of Chromosome 7 ESTs

4.1 Introduction.

One of the main objectives of the Human Genome Project is to identify and map all the genes residing in the human genome. This in turn should facilitate further studies aimed at providing more information about biological processes and elucidating their possible role in the molecular pathology of disease (Collins and Galas 1993). Expressed sequence tags (ESTs; section 1.3.5) are created by partially sequencing random cDNA clones from a source library. This rapid and simple process of generating these sequences has allowed them to be used as a resource to determine the range of genes that are expressed in certain tissue and cell types. One application for EST sequences is to generate cDNA specific STSs which serve as unique gene markers for physical mapping. Wilcox and co-workers (Wilcox et al. 1991) first described the concept of using the 3' untranslated sequence from cDNAs for this purpose. This region displays significant sequence variation, which is essential for discriminating between members of gene families and conserved genes from other species. In addition, it has been noticed that it is also less likely to contain introns, thus supplying an identical template sequence to that of genomic DNA, which is important when developing a diagnostic STS for use in PCR reactions. However, in earlier studies aimed at creating a large number of ESTs (Adams et al. 1991), sequences were synthesised from random areas of the cDNA sequence intentionally. This was done to increase the chance of generating sequence from the coding regions of cDNAs. These could then be used to search sequence databases in order to detect homologies to known genes and to discover new members of gene families, both in human sequence and in other species. This information in some cases may provide clues as to the protein coding function of the gene.

A number of investigators (Wilcox et al. 1991, Adams et al. 1991 and Okubu et al. 1992) predicted that the generation of ESTs would play an important role in accomplishing a near complete transcript map for the human genome. The systematic generation of ESTs has been undertaken by several groups over the past few years (Adams et al. 1991; Khan et al. 1992; Houlgatte et al. 1995; Hillier et al. 1996). The
value of ESTs is increased when they are assigned to a specific region in the genome. This can be performed by FISH (Korenburg et al. 1994), PCR analysis in a somatic cell hybrid panel (Durkin et al. 1992), or PCR in a radiation hybrid panel (Schuler et al. 1996). Initial localisations can be further refined to a higher resolution on YACs or cosmids (Berry et al. 1995). These gene-based tags can then function as anchor points in the genome for contig mapping and also as positional candidates for genetic disorders mapped to a defined region.

Soon after the commencement of large scale random cDNA clone sequencing by Venter and colleagues (Adams et al. 1991, 1992a, 1993a and 1993b), several groups took the initiative of mapping a proportion of the published ESTs to individual chromosomes. Researchers interested in particular chromosomal regions have to date mapped ESTs to distinct regions on chromosomes 5 (Feldblyum et al. 1996), 6 (Paapas et al. 1995), 11 (Slorach et al. 1995; Rosier et al. 1995), 13 (Hawthorn et al. 1996), 18 (Gerken et al. 1994), 21q (Chiang et al. 1995) and X (Mazzarella and Srividtava 1994). This information has benefited investigators who are trying to identify genetic disorders mapping to these chromosomes. In this manner, we proposed to add to the transcriptional map of chromosome 7 by assigning ESTs that had previously been allocated to chromosome 7, to further defined sub-regional locations. This would assist the search for positional candidates for chromosome 7 linked diseases such as cystoid macular dystrophy (CYMD), autosomal dominant and recessive deafness (DFNA5 and DFNB4), retinitis pigmentosa (RP9 and RP10), Charcot-Marie Tooth disease (CMT2D), osteogenesis imperfecta 4 (OI4), and craniosyntosis (CRS) (OMIM, 1997). To identify positional candidates for RP9, ESTs that were mapped to the 7p14-15 sub-region in this study, and known genes that have previously been mapped here, were analysed to test for their presence in YAC and cosmid clones that reside within the disease interval. The critical region for this disease had been defined by the distal and proximal flanking markers, D7S795 and D7S484 respectively (Keen et al. 1995a; section 1.6.1).

Chromosome 7 represents approximately 5% of the human genome, consisting of an estimated 170 Mb of DNA (Morton 1991). At the time of study, approximately 100 genes had been localised to chromosome 7, as well as two large gene clusters at the HOXA and TCRG loci (OMIM database, April 1995). This represented a small
proportion of the estimated 3,500-5,000 genes predicted to occupy chromosome 7 (Fields et al. 1994). The assignment of the ESTs to particular regions of this chromosome, would therefore further add to the gene map of chromosome 7.

4.1.1 Tools for assignment.

The use of a somatic cell hybrid panel in this study provided an efficient method for EST assignment, permitting “rapid binning” of ESTs to certain sub-regions of chromosome 7. This provides a resolution similar to that attained by low resolution FISH. The success of the technique relies on the PCR primers being human specific, so as not to give problems of distinguishing the target product from the rodent background. The PCR based approach to mapping is increasingly favoured in comparison to Southern blot hybridisation analysis of the EST probes onto the hybrids, as it requires less template DNA and is more rapid. The assignment of ESTs in a panel is determined by scoring the presence or absence of amplified products in each cell line, which deciphers a pattern of concordance by its localisation.

The somatic cell hybrid panel used in this mapping study is based on a mouse background and was a generous donation by Prof. K-H Grzeschik (Institut für Humangenetik der universität, Marberg, Germany). The cell lines consist of overlapping breakpoint fragments which together divide chromosome 7 into 12 sub-regions. These breakpoints are defined by terminal or interstitial deletions, or rearrangements of DNA obtained from Greig cephalopolysyndactyly syndrome patients (Vortkamp et al. 1991) as well as from other sources (Tsui and Farrall 1991). The human content in each cell line has been determined using chosen chromosome 7 specific genetic markers and genes that provide reference points for these sub-regions (Vortkamp et al. 1991; Tsui et al. 1994). The cell line 5387-3cl10, was used as a control as it contains an intact chromosome 7 as its only human material, while other somatic cell hybrids carry one or more other human chromosomes in addition to the chromosome 7 segment. To ensure that the primers were species specific, total human and total mouse DNA were also included as controls.
4.1.2 Aims

At the time of study, the number of identified genes on chromosome 7 was limited and many of those placed on chromosome 7 had not been sub-localised. Thirty ESTs that had previously been assigned to chromosome 7 by various groups were mapped in this investigation to specific sub-regions on chromosome 7 using a PCR based hybrid panel mapping strategy. The assignment of these ESTs were performed as a collaboration with several other groups working on chromosome 7 (Grzeschik et al. 1994), with the aim of further localising genes expressed on this chromosome and providing new positional candidates for genetically linked diseases. Database analysis was also performed on the EST sequences to identify and assess any nucleotide or protein similarities to known genes.

Several ESTs that were mapped in this study and known chromosome 7 genes were investigated further for their role as positional candidates for RP9. At the time, no gene transcripts had been identified to reside in the RP9 critical region. ESTs that were assigned to the 7p14-15 sub-region in this study were therefore tested for their presence in YAC clones from the contig which spans the RP9 disease region (Keen et al. 1995a; figure 3.16).

4.2 Results

4.2.1 Assignment of dinucleotide markers

To clarify the EST content of chromosome 7 in different cell lines, several dinucleotide genetic markers of known location were typed among the cell lines by PCR assay. Certain genetic markers have previously been examined to determine the breakpoints, and indicate which portions of chromosome 7 the individual cell lines contain (Tsui et al. 1994). STS primer pairs for the markers, D7S484, D7S519, D7S524, D7S486 and D7S550 were used for PCR amplification against the cell lines of the panel as an additional verification of cell line constitution. D7S484 and D7S486 gave a pattern of concordance that confirmed the localisations of these markers in this panel, which have previously been tested (Tsui et al. 1994). D7S519, gave a pattern that suggested its presence in a sub-chromosomal region defined by breakpoints that had not been formerly tested for genetic markers. D7S524 and D7S550 were localised
Regional Assignment of Chromosome 7 ESTs to sub-chromosomal regions, that were previously defined by other genetic markers. All assignments in this study are in agreement with the order of the Généthon genetic map (Dib et al. 1996).

4.2.2 Source of ESTs.

All of the ESTs in this study were previously assigned to chromosome 7 by the investigators, who selected PCR primers from their sequences and studied the segregation of amplified products from monochromosomal rodent-human somatic cell hybrid panel templates. The origins of the individual chromosome 7 assigned ESTs used in this study are listed in table 4.1. Of the thirty ESTs, twenty-five were first described by Adams and co-workers (Adams et al. 1991, 1992a and 1993a). These were derived from several different sources of brain libraries, whose cDNA sequences were obtained from random and poly d(T) priming simultaneously. Eighteen of these were placed on chromosome 7 previously by Polymeropoulos and colleagues (Polymeropoulos et al. 1993) or Durkin and co-workers (Durkin et al. 1992 and 1994). PCRs on the monochromosomal hybrid panels were performed either by employing one oligonucleotide primer pair per reaction (Polymeropoulos et al. 1993) or by multiplexing several primer pairs together, after confirming the STSs were species specific (Durkin et al. 1994). Seven other ESTs were localised to chromosome 7 by Maglott and colleagues (1994) at the American Type Culture Collection (ATCC, Maryland) in the same manner. The remaining five ESTs, were generated by Hayes and co-workers at the HGMP resource centre, UK. Three of these ESTs were derived from random clones from human fetal whole brain and two were from a human fetal liver library. Oligonucleotides and reaction conditions for the majority of these are deposited in dbSTS (at the National Centre for Biotechnology Information web site). These previously reported chromosome 7 EST assignments were sub-regionally refined in this study.

4.2.3 EST assignment.

To ascertain whether the EST primer pairs amplified human specific products,
### Table 4.1

<table>
<thead>
<tr>
<th>EST no./ Lab name</th>
<th>Locus Symbol / Accession no.</th>
<th>Parent Library</th>
<th>Source of EST:</th>
<th>Chr 7 assignment by:</th>
</tr>
</thead>
<tbody>
<tr>
<td>00022</td>
<td>D7S543E</td>
<td>FB</td>
<td>Adams et al. 1991</td>
<td>Polymeropoulos et al. 1993</td>
</tr>
<tr>
<td>00085</td>
<td>D7S534E</td>
<td>H</td>
<td>Adams et al. 1991</td>
<td>Polymeropoulos et al. 1993</td>
</tr>
<tr>
<td>00122</td>
<td>D7S536E</td>
<td>H</td>
<td>Adams et al. 1991</td>
<td>Polymeropoulos et al. 1993</td>
</tr>
<tr>
<td>00167</td>
<td>D7S532E</td>
<td>H</td>
<td>Adams et al. 1991</td>
<td>Polymeropoulos et al. 1993</td>
</tr>
<tr>
<td>00231</td>
<td>D7S524E</td>
<td>TC</td>
<td>Adams et al. 1991</td>
<td>Polymeropoulos et al. 1993</td>
</tr>
<tr>
<td>00439</td>
<td>D7S544E</td>
<td>FB</td>
<td>Adams et al. 1992a</td>
<td>Polymeropoulos et al. 1993</td>
</tr>
<tr>
<td>00446</td>
<td>D7S545E</td>
<td>FB</td>
<td>Adams et al. 1992a</td>
<td>Polymeropoulos et al. 1993</td>
</tr>
<tr>
<td>00548</td>
<td>D7S551E</td>
<td>FB</td>
<td>Adams et al. 1993a</td>
<td>Durkin et al. 1992</td>
</tr>
<tr>
<td>00601</td>
<td>D7S552E</td>
<td>FB</td>
<td>Adams et al. 1992a</td>
<td>Durkin et al. 1992</td>
</tr>
<tr>
<td>00631</td>
<td>D7S548E</td>
<td>FB</td>
<td>Adams et al. 1992a</td>
<td>Polymeropoulos et al. 1993</td>
</tr>
<tr>
<td>00654</td>
<td>D7S556E</td>
<td>FB</td>
<td>Adams et al. 1992a</td>
<td>Polymeropoulos et al. 1993</td>
</tr>
<tr>
<td>00817</td>
<td>D7S537E</td>
<td>H</td>
<td>Adams et al. 1992a</td>
<td>Polymeropoulos et al. 1993</td>
</tr>
<tr>
<td>00838</td>
<td>D7S553E</td>
<td>H</td>
<td>Adams et al. 1992a</td>
<td>Durkin et al. 1992</td>
</tr>
<tr>
<td>00969</td>
<td>D7S538E</td>
<td>H</td>
<td>Adams et al. 1992a</td>
<td>Polymeropoulos et al. 1993</td>
</tr>
<tr>
<td>00979</td>
<td>D7S539E</td>
<td>H</td>
<td>Adams et al. 1992a</td>
<td>Polymeropoulos et al. 1993</td>
</tr>
<tr>
<td>01644</td>
<td>D7S540E</td>
<td>SH</td>
<td>Adams et al. 1992a</td>
<td>Polymeropoulos et al. 1993</td>
</tr>
<tr>
<td>01888</td>
<td>D7S557E</td>
<td>FB</td>
<td>Adams et al. 1992a</td>
<td>Durkin et al. 1994</td>
</tr>
<tr>
<td>01995</td>
<td>D7S1877E</td>
<td>FB</td>
<td>Adams et al. 1992a</td>
<td>ATCC</td>
</tr>
<tr>
<td>02120</td>
<td>D7S552E</td>
<td>FB</td>
<td>Adams et al. 1992a</td>
<td>ATCC</td>
</tr>
<tr>
<td>02379</td>
<td>D7S1876E</td>
<td>FB</td>
<td>Adams et al. 1992a</td>
<td>ATCC</td>
</tr>
<tr>
<td>02419</td>
<td>D7S1878E</td>
<td>FB</td>
<td>Adams et al. 1992a</td>
<td>ATCC</td>
</tr>
<tr>
<td>06146</td>
<td>D7S11428E</td>
<td>IB</td>
<td>Adams et al. 1993a</td>
<td>ATCC</td>
</tr>
<tr>
<td>06238</td>
<td>T08346</td>
<td>IB</td>
<td>Adams et al. 1993a</td>
<td>ATCC</td>
</tr>
<tr>
<td>06275</td>
<td>D7S11559E</td>
<td>IB</td>
<td>Adams et al. 1993a</td>
<td>ATCC</td>
</tr>
<tr>
<td>AAAAKP</td>
<td>Z19607</td>
<td>HFB</td>
<td>Hayes et al. 1996</td>
<td>Hayes et al. 1996</td>
</tr>
<tr>
<td>AAABQZ</td>
<td>Z20232</td>
<td>HFB</td>
<td>Hayes et al. 1996</td>
<td>Hayes et al. 1996</td>
</tr>
<tr>
<td>AAADEER</td>
<td>Z20930</td>
<td>HFB</td>
<td>Hayes et al. 1996</td>
<td>Hayes et al. 1996</td>
</tr>
<tr>
<td>AAABWFR</td>
<td>—</td>
<td>HFL</td>
<td>Hayes et al. 1996</td>
<td>Hayes et al. 1996</td>
</tr>
<tr>
<td>AAAFENH</td>
<td>—</td>
<td>HFL</td>
<td>Hayes et al. 1996</td>
<td>Hayes et al. 1996</td>
</tr>
</tbody>
</table>

The table lists the EST name, locus symbol, source library of the corresponding clone, the group who first reported the EST and who assigned the EST to chromosome 7. FB: Stratagene human fetal brain library (cat. #936206); H: Stratagene human hippocampus library (cat. #936205); SH: Stratagene subtracted hippocampus library; IB: B. Soares infant brain library; HFB: HGMP fetal brain library; HFL: HGMP fetal liver library; and TC: Stratagene temporal cortex (cat. #935205).
oligonucleotides were initially tested by PCR amplification on 50 ng human and mouse DNA templates. Non-specific amplification products were observed in 20% of the ESTs in addition to the expected band, but were of distinctly different sizes and were easily distinguishable from the expected STS product size of the EST.

PCR conditions for most EST primer pairs were adjusted according to those specified for each individual primer pair in the dbSTS database. Some PCR conditions were optimised by assessing the ideal magnesium concentration or annealing temperature to reduce the level of non-specific amplified products, and to produce distinct bands. Once satisfactory conditions were obtained, the primers were used for amplification on 100 ng of individual cell line DNA from the somatic cell hybrid panel. Products were electrophoresed through a 2% agarose gel alongside a size standard (section 2.7.1 and 2.3.1). Sub-chromosomal assignment was determined by concordancy of the pattern of amplified products in each cell line. Figure 4.1 shows the assignment by PCR of ESTs 06238 and AAAFENH as examples. The results for each individual EST are listed in table 4.2.

In brief, four ESTs mapped to band 7p22, three to 7p13-21, one was localised to 7p12-13, one to the pericentric region, fourteen were assigned to cen-q22, one to q31.1-q22 and six to the q32-36 region. Individual assignment is illustrated in figure 4.2. Some of these localisations represent assignments equivalent in resolution to that achieved by FISH.

These mapped ESTs can now be considered positional candidate genes for diseases that have been associated with chromosome 7. As little is known about the expression pattern of the transcripts associated with these gene-based STSs, one can only speculate which diseases these possible genes could be candidates for based on the source of the original library. In this case, most of the ESTs were derived from brain expressed cDNA clones, so tissue specific genes could be considered as candidates genes for neurological diseases. However, unless the library that was used had been subtracted for housekeeping transcripts, a large proportion of these may be widely expressed. Until more is known about the expression patterns of each EST, the co-localisation of these to a defined region would make them acceptable as positional candidates. We can therefore speculate that ESTs 00601, 02120 and 06146 which map to the 7p21-13 region, could be candidates for diseases such as autosomal dominant
Figure 4.1 PCR amplification of a) EST06238 (89 bp) and b) AAAFENH (134 bp) on individual cell lines of the chromosome 7 somatic cell hybrid panel. This assigned EST06238 to 7q32-7qter and AAAFENH to 7p22. See table 4.2 for the human content in each cell line.
Regional Assignment of Chromosome 7 ESTs

Table 4.2 Table displaying results from PCR amplification of each EST through the human-mouse somatic cell hybrid panel. / indicates positive amplification scores. Blank spaces denote negative amplification. X indicates cell lines not tested. The above diagram illustrates the human content of individual cell lines in the panel. (Continued overleaf).
## Table 4.2

(continued from previous page).

<table>
<thead>
<tr>
<th>EST</th>
<th>1696 Reg 2-2</th>
<th>0044 Reg 1-15</th>
<th>0044 Reg 2-40.5</th>
<th>IT A9 2-139-25</th>
<th>IT A9 2-139-14</th>
<th>1355 Reg 2-21</th>
<th>Au Rag 8-18-17</th>
<th>Au Rag 6-19</th>
<th>Au Rag 6-22-2</th>
<th>1568 Reg 6-12-2</th>
<th>194 Reg 6-13-1</th>
<th>5387 Reg 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>00979</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>136</td>
</tr>
<tr>
<td>02429</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>136</td>
</tr>
<tr>
<td>06235</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>136</td>
</tr>
<tr>
<td>AAA</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>136</td>
</tr>
<tr>
<td>AAA</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>136</td>
</tr>
<tr>
<td>AAA</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>136</td>
</tr>
<tr>
<td>AAA</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>136</td>
</tr>
<tr>
<td>AAA</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>136</td>
</tr>
<tr>
<td>00922</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>136</td>
</tr>
<tr>
<td>00923</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>136</td>
</tr>
<tr>
<td>00946</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>136</td>
</tr>
<tr>
<td>00947</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>136</td>
</tr>
<tr>
<td>00995</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>136</td>
</tr>
<tr>
<td>00996</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>136</td>
</tr>
<tr>
<td>00997</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>136</td>
</tr>
<tr>
<td>00928</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>136</td>
</tr>
</tbody>
</table>

152
Figure 4.2  Schematic representation of the sub-regional assignment of the ESTs on chromosome 7. The vertical bars represent the human content of each cell line. EST localisation is represented on the left with the assignment of microsatellite markers on the right. The markers tested in this study are marked with an asterix.
non-syndromic hearing loss (DFNA5; Camp et al. 1995), Charcot-Marie Tooth axonal neuropathy (CMT2D; Ionasescu et al. 1996), cystoid macular dystrophy (DCMD; Kremer et al. 1994) and retinitis pigmentosa 9 (RP9; Inglehearn et al. 1993). Cerebral cavernous malformation type 1 (CCM1; Dubovsky et al. 1995) has been located to the 7q11-q22 region, with the disease critical region spanning the region between two breakpoints defined by the hybrid panel. EST 01995, situated in the 7p11.1 region and ESTs 00085, 00272, 00548, 00631, 00838, 00969, 00979, 02419, 06275, AAAFBWR, AAAAKP, AAADEER and AAABQRZ are assigned to 7qcen-q22. These could be considered candidate genes for this disorder. The ESTs that are mapped to this latter portion may be examined further as candidates for Zellweger syndrome (ZS), a condition mapped to this region by chromosomal rearrangements and deletions (Naritomi et al. 1989). EST 00167 at 7q22-32 may be a candidate for autosomal recessive deafness 4 (DFNB4; Baldwin et al. 1995) as well as retinitis pigmentosa (RP10; Jordan et al. 1993). Finally, ESTs 00231, 00654, 02379, 00446, 01888 and 06238 were assigned to 7q32-36. Diseases located here include Smith-Lemli-Opitz syndrome (SLOS; Alley et al. 1997). Once positional candidates have been localised at this level, the ESTs must be placed at a higher resolution to determine whether they reside in a defined critical region for the disease. This can be achieved by using the STSs to PCR the clones that map within that region. If any are placed within the disease region they would probably be first analysed to determine their pattern of expression, and their cDNA and genomic structure would be determined. Mutation screening would then be carried out to determine whether or not the genes associated with these ESTs play a role in the molecular pathogenesis of a disease.

To allow public sharing and accessibility of this information within the wider scientific community, a paper summarising these results have since been published (Patel et al. 1995; included at the end of this thesis).

4.2.4 Homologies of ESTs to sequences deposited in databases.

As the output of new sequences generated and deposited in databases is ever increasing, it is important to keep up to date with new comparisons against comprehensive nucleotide and peptide databases to detect similarities to recent updated sequences. As certain EST sequences mapped in this study are not solely
Regional Assignment of Chromosome 7 ESTs

derived from the 3' untranslated regions, and may derive from coding portions, they were recently reanalysed by similarity searches against sequences in nucleotide and protein databases to update information on these ESTs. Comparisons were accomplished using the “basic local alignment search tool” programs (Altschul et al. 1990), BLASTN, for nucleotide and BLASTX, for protein functions against sequences in several databases, operating through their servers at NCBI (http://www.ncbi.nlm.nih.gov/BLAST/; section 2.13). In this analysis, matches to database sequences were only considered significant if their p-value was $10^{-5}$ or less. Each EST sequence was compared against the non-redundant nucleotide database of Genbank, EMBL, DDBJ and PDB directories, and also against Alu repeats from the REPBASE database, using BLASTN. If no statistically significant matches were found, the sequences were then compared to the non-redundant SwissProt database after translation of the query sequence in all six reading frames via BLASTX.

ESTs 01644, 06146 and 06275 have previously been reported to share significant statistical sequence identity with nucleotide or amino acid sequences in database records (Adams et al. 1992a and 1993a). These matches were reanalysed with sequences from the updated versions of the databases. EST01644 is now found to be 97% identical to a 291 bp stretch of the cAMP-dependent protein kinase regulatory subunit RIβ gene (PRKAR1B; figure 4.3), and its localisation is consistent with the assignment of the gene to 7pter by genetic analysis (Solberg et al. 1992). EST06146 now matches the human amphiphysin gene with 98% identity over 334 bp (figure 4.4). Amphiphysin has been found to be expressed in the synaptic terminals of several neuronal cell types, the adrenal medulla, pituitary gland, pancreas and spermatocytes (Yamamoto et al. 1995a). Its localisation to the 7p13-14 region by FISH analysis also verifies the mapped position of this EST (Yamamoto et al. 1995b).

EST06275, representing sequence from the 3' end of the cDNA clone, was originally found to share 100 % identity corresponding to a 28 amino acid region within the 3' region of the rat metabotropic glutamate receptor 3 gene (mGluR3). At the time of study, the human homologue of this gene had not been localised, so the placing of this gene to 7q11-q22 represented a tentative new assignment. This localisation was confirmed later by Scherer and colleagues who assigned mGluR3 to 7q21.1-21.2 using a somatic cell hybrid panel and by FISH (Scherer et al. 1996).
Regional Assignment of Chromosome 7 ESTs

**Human cAMP-dependent protein kinase regulatory subunit RI-beta mRNA, 3' end.**

- gb|M65066|HUMRiB Human cAMP-dependent protein kinase regulatory subunit RI-beta mRNA, 3' end. Length = 2366
- gb|G18327|HUMSW1462 human chromosome 7 STS

**Plus Strand HSPs:**

<table>
<thead>
<tr>
<th>Score</th>
<th>Expect</th>
<th>Identities</th>
<th>Positives</th>
<th>Strand</th>
</tr>
</thead>
<tbody>
<tr>
<td>952</td>
<td>1.1e-103</td>
<td>194/199 (97%)</td>
<td>194/199 (97%)</td>
<td>Plus / Plus</td>
</tr>
</tbody>
</table>

**Figure 4.3** Sequence similarity alignment of EST01644 with the cAMP-dependant regulatory subunit RIβ gene using the BLASTN programme against the non-redundant nucleotide database of Genbank, EMBL, DDBJ and PDB.

**Human sviens mRNA for amphiphysin**

- emb|X81438|HSAmpA H.sapiens mRNA for amphiphysin Length = 3260

**Plus Strand HSPs:**

<table>
<thead>
<tr>
<th>Score</th>
<th>Expect</th>
<th>Identities</th>
<th>Positives</th>
<th>Strand</th>
</tr>
</thead>
<tbody>
<tr>
<td>1626</td>
<td>1.5e-127</td>
<td>328/334 (98%)</td>
<td>328/334 (98%)</td>
<td>Plus / Plus</td>
</tr>
</tbody>
</table>

**Figure 4.4** Sequence similarity alignment of EST06146 with human amphiphysin gene using the BLASTN programme against the non-redundant nucleotide database of Genbank, EMBL, DDBJ and PDB.
These investigators also integrated EST06275 into the genetic and physical map of chromosome 7 by mapping it to a CEPH YAC of known location. Current database analysis reveals that this EST has 97% identity over 380 bp with the human gene and 98% identity with 361 bp of sequence from the human BAC clone RG331C24, which has also been mapped to 7q21, yet again confirming its location (figure 4.5).

New EST matches with the updated sequence databases include EST00979, which possesses 99% nucleotide identity over 293 bp with sequence from the human BAC clone RG083M05. This clone had previously been localised to 7q21-7q22 by STS content analysis (Shizuya et al. 1992; Kim et al. 1996). Similarly, EST0631 shares 97% nucleotide identity over a 298 bp sequence from the human BAC clone RG072E11 assigned to 7q21. This BAC clone, as with the others mentioned above, is one of many clones chosen for the large scale sequencing project undertaken by NHGRI and the University of Washington Genome Centre (Shizuya et al. 1992; Kim et al. 1996). These findings agree with the position of these ESTs and provide refinement of their localisation.

Two ESTs were found to share homologies to gene sequences in recent database entries. EST00272 showed 99% sequence identity over 384 bp with the human Krit1 mRNA (figure 4.6), a novel gene sequence. It also shares significant homology to a human BAC clone RG161K23 that has been mapped to 7q21, yet again confirming the assignment of the EST. This BAC clone is part of an unanchored contig of overlapping BACs whose orientation has not yet been determined but contains STSs sWSS3201 and sWSS1708. It overlaps with a 200 bp region of BAC RG072E11 which contains EST00631, therefore maps these ESTs close together. Finally, EST00167 has 99% identity over 215 bp with the human putative endothelin receptor type B-like mRNA (fig 4.7), a novel gene originally isolated from a human hippocampus cDNA library (Zeng et al. 1997). The assignment of this gene has not previously been determined, so the mapping of this EST to 7q22-32 represents a tentative new gene assignment.

Many of the sequences for these ESTs have an "N" in positions where nucleotide sequences that were generated from single pass sequencing could not be deciphered. When observing the percentage homologies, these discrepancies probably account for the last few percent of sequences that are in fact almost certainly identical.
Regional Assignment of Chromosome 7 ESTs

A) em|X77748|BMGLUR3

H. sapiens mRNA for metabotropic glutamate receptor
type 3  em|A66194|A66194  Sequence 1 from Patent WO9522609
Length = 1410
Minus Strand HSPs:
Score = 1838 (507.9 bits), Expect = 3.4e-145, P = 3.4e-145
Identities = 372/380 (97%), Positives = 372/380 (97%), Strand = Minus / Plus

06275: 381 GACGCCAATCTCGCTCAACAGTATATCGCAGCTGAACTGGCAATGGCCAGGGCCG 322
Sbjct: 2800 GAACGCCAATCTCGCTCAACAGTATATCGCAGCTGAACTGGCAATGGCCAGGGCCG 2859
Score = 1838 (507.9 bits), Expect = 3.4e-145, P = 3.4e-145
Identities = 372/380 (97%), Positives = 372/380 (97%), Strand = Minus / Plus

B) gb|AC002081|BAC002081

Human BAC clone RG331C24 from 7q21; HTGS phase
3, complete sequence [Homo sapiens]
Length = 136,150
Minus Strand HSPs:
Score = 1761 (486.6 bits), Expect = 1.9e-138, Sum P(2) = 1.9e-138
Identities = 355/361 (98%), Positives = 355/361 (98%), Strand = Minus / Plus

06275: 141 TGTGATGTGCTAGAACCTCTAGGCTGAGTCTAGTGCCCCTATTATTAACACACCCCA 243
Sbjct: 47222 CCTCTGCAAGCACGTATGTGCCAACGGTGTGCAATGGGCGGGAAGTCCTCGACTCCACCA 303
Score = 1761 (486.6 bits), Expect = 1.9e-138, Sum P(2) = 1.9e-138
Identities = 355/361 (98%), Positives = 355/361 (98%), Strand = Minus / Plus

Figure 4.5 Sequence similarity alignment of EST06275 with A) the human metabotropic glutamate receptor 3 gene and B) BAC clone RG331C24 using the BLASTN programme against the non-redundant nucleotide database of Genbank, EMBL, DDBJ and PDB.
**Regional Assignment of Chromosome 7 ESTs**

---

**Human Krit1 mRNA, complete cds**

Length = 2004  
Minus Strand HSPs:  
Score = 1902 (525.6 bits), Expect = 5.2e-162, Sum P(2) = 5.2e-162  
Identities = 382/384 (99%), Positives = 382/384 (99%), Strand = Minus / Plus  
00272: 384 GAAAACGTGCAAGAGCAAGCAGGAGGGGCGAAGCGAAGAGGACAGGAGGTTGA 325  
Sbjct: 208 GAAAAATCGAGAGAACAGGCAAGCAGGAGGGGCGAAGCGAAGAGGACAGGAGGTTGA 267  
00272: 324 TGAATTACCTCTCCACGAGGAGGTGATTTTGTAGGTTATAGGTATTTACACGGCAAGC 265  
Sbjct: 268 TGAATTACCTCTCCACGAGGAGGTGATTTTGTAGGTTATAGGTATTTACACGGCAAGC 327  
00272: 264 CAGTGAAAGATTTCACTACCCAAGGATATGGAACCCATTACTTGCG 205  
Sbjct: 328 CAGTGAAAGATTTCACTACCCAAGGATATGGAACCCATTACTTGCG 387  
00272: 204 AGCTCTGTTAGAAGAGGTTGCTCTGCTTGGATAGTACACGATGTTGAGAAGCG 145  
Sbjct: 388 AGCTCTGTTAGAAGAGGTTGCTCTGCTTGGATAGTACACGATGTTGAGAAGCG 447  
00272: 144 AAAAAACGTTAAATGAACATGCTTTTCCTTTTCAGTTGTCAATTGCAT 85  
Sbjct: 448 AAAAAACGTTAAATGAACATGCTTTTCCTTTTCAGTTGTCAATTGCAT 567  
00272: 94 TUAATTCGATAAAATCCATGACACCCGACCAGATGACACATACAGAGGCAACA 25  
Sbjct: 508 TUAATTCGATAAAATCCATGACACCCGACCAGATGACACATACAGAGGCAACA 567  
00272: 24 AGGACAGTTCTCAATTTATCTG 1  
Sbjct: 568 AGGACAGTTCTCAATTTATCTG 591  

**Figure 4.6**  
Sequence similarity alignment of EST00272 with the human Krit1 gene using the BLASTN programme against the non-redundant nucleotide database of Genbank, EMBL, DDBJ and PDB.

---

**Human putative endothelin receptor type B-like protein mRNA, complete cds**

Length = 4156  
Minus Strand HSPs:  
Score = 1066 (294.6 bits), Expect = 6.8e-81, P = 6.8e-81  
Identities = 214/215 (99%), Positives = 214/215 (99%), Strand = Minus / Plus  
00167: 215 TTAAATAATCAATTATGTCAAATTTATTATAGCAATGGATGTAATTATATAT 156  
Sbjct: 3924 TTAAATAATCAATTATGTCAAATTTATTATAGCAATGGATGTAATTATATAT 3983  
00167: 155 TTGAGTTGATAAGTTGAGTGTTGGATGTGTGTGTATGATGATGGAGG 96  
Sbjct: 3984 TTGAGTTGATAAGTTGAGTGTTGGATGTGTGTGTATGATGATGGAGG 4043  
00167: 95 ATTCACAAAGTTACAAAAATTTTCTTCTTTATTATCTTCTTTACT 36  
Sbjct: 4044 ATTCACAAAGTTACAAAAATTTTCTTCTTTATTATCTTCTTTACT 4103  
00167: 35 TTAAACAATTTCTAATATAATTATTTTTTCTTCTTTATTATCTTCTTTACT 1  
Sbjct: 4104 TTAAACAATTTCTAATATAATTATTTTTTCTTCTTTATTATCTTCTTTACT 4138  

**Figure 4.7**  
Sequence similarity alignment of EST00167 with the human putative endothelin receptor type B-like mRNA using the BLASTN programme against the non-redundant nucleotide database of Genbank, EMBL, DDBJ and PDB.
Two ESTs showed homology to the Alu repeat element. EST01995 contains part of an Alu repeat. Sequence comparisons of this against those from select Alu repeats from the REPBASE database has identified strong similarity with 122 bp of the 5' ends of the -Sq and -Sp sub-family of Alu-S consensus sequences. EST01888 is almost entirely composed of Alu sequences, retaining similarity with the oldest family of Alu elements, the Alu-J family. Nevertheless, the EST primers used here define a unique STS that is present in a brain cDNA library. The location of these sequences in the cDNA for both these ESTs are not known as they were derived from a random primed cDNA library. The presence of these short interspersed repeats are not uncommon in cDNAs as it is estimated that approximately 5% of cDNAs derived from completely spliced mRNAs contain part of or all of an Alu sequence (Yulug et al. 1995), and are generally present at the 3' or 5' untranslated regions and usually absent from coding portions.

The remaining EST sequences share no significant homologies to sequences from Genbank, EMBL, DDJB and PDB database records. These may possibly represent novel gene sequences.

4.2.5 The search for positional candidates for RP9.

ESTs which were mapped to the 7p13-15 sub-region in this study (ESTs 00601, 02120 and 06146), and known genes which have previously been assigned to this region of the chromosome, were investigated further to assess whether any were present within the YAC contig spanning and extending out of the RP9 disease region (figure 3.16; Keen et al. 1995a). A set of pools containing selected YACs from the RP9 contig were combined and used as a source to test for the presence of these ESTs. If present, their individual YAC address and their position in the physical map of the RP region was determined. The pools contained several redundant YACs to assure a level of complete coverage was achieved and in case any of these clones possessed small undetected deletions. The RP9 disease interval had not been refined at the onset of this investigation, therefore YACs within and extending from the disease region were included. Pool A contained YACs isolated using markers from the distal portion of the region (D7S435, D7S632, D7S526 and D7S690). Pool B contained YACs
isolated using markers from the proximal portion of the region (D7S484, D7S497, D7S460 and D7S683). Some YACs in pool A overlap others in pool B.

4.2.5.1 Identification of a positional candidate EST for RP9.

EST06146 did not amplify in either of the two pools containing YACs from the RP9 contig, and was therefore excluded. EST00601 was found to be present in pool A, and when tested for each individual YAC, was found to reside on ICI YAC yWSS2458. STS content analysis on redundant clones in that area placed this EST distal to the defined critical region for the disease, which excluded it as a candidate for RP9. The localisation of EST00601 has been described in detail in section 3.2.6.

STS primers for EST02120 amplified the correct product size in both YAC pools A and B. It was found to be present in several overlapping YACs at the distal boundary of the RP9 critical region, but mapped within the disease region (see figure 4.8). At the time of this study, this was the first gene-based transcript that had been identified in the critical region. Its exact position, determined by STS content mapping, is between STSs A1 and 3586D. It was also found to be duplicated in a more proximal part of the contig, along with 3586D (figure 4.8 and Keen et al. 1995a), being present in two physically distinct locations within the critical region. This was discovered when using less stringent temperature PCR conditions, the STS for EST02120 also amplified the expected sized product in YACs yWSS3586, 908-F-12 and cosmids J0117 and F0249. The proximal version of EST02120 is designated EST02120P and the distal version is 02120D. Similarly the YAC end clone 3586RA has a duplicated equivalent in the distal part of the contig designated 3586D (Keen et al. 1995a). Regional duplications of relatively short segments outside tandem arrays has been observed as a feature of the large scale structure of the genome (Lupski et al. 1996). For example, the middle and long-wavelength photopigment genes, located within close proximity of each other on chromosome Xq, are thought have arisen from the duplication of a single gene locus (Nathans et al. 1986). In addition, a 1.5 Mb regional duplication on chromosome 17p11.2-p12 has been identified as the cause of disease for Charcot-Marie-Tooth (CMT) disease (Patel and Lupski 1994). The extent and characterisation of the duplication in the RP9 region is currently being investigated by colleagues.
Figure 4.8 The physical map location of EST02120. Only selected YACs and STSs are shown. Thin horizontal lines represent YACs, thick horizontal lines depict cosmids. Vertical dashed lines represent the location of STSs. The bold vertical line is the end clone for YAC yWSS3586. In this illustration EST02120 and STS3586 is represented twice in the region, due to the presence of a duplication. The position of J0117, F0249, M1040, yWSS2550 and yWSS798 in the contig was determined by employing oligonucleotides for 3586D and 3586RA which exploit nucleotide differences in sequences from the duplicated regions (T.J. Keen, personal communication).
The cDNA clone for which this EST had derived from, HFBCK70, was obtained from ATCC. This clone is from a commercial 17-18 week gestation fetal brain library which was both random and poly d(T)$_n$ primed. It is cloned within the pBluescript vector and is 908 bp in size. Sequencing of the cDNA clone was performed by a colleague using T3, T7 and an internal sequencing primer (dbEST ID: IOO_02, accession No: H52869 and NCBI ID: 354079). Initial analysis on this sequence revealed no significant open reading frames, no poly (dA)$_n$ tail and no consensus polyadenylation signal sequence. BLAST and FASTA similarity searches of nucleotide databases disclosed that it possessed high sequence homology to part of the MER37 medium reiteration frequency interspersed repeat. At the original time of analysis only 119 bp of the repeat element had been identified (Iris et al. 1993). Recent analysis has revealed the clone is actually part of the 2,418 bp defined consensus sequence of the human Tigger1 transposable element, which is a class of the MER group of repeats (Smit and Riggs 1996; figure 4.9). When compared to sequences from the cDNA databases and dbEST, nucleotide identity was only seen with those clones that possessed MER repeats. It is not yet possible to say whether this sequence it is actually part of a gene. It may actually represent a genomic DNA contamination artefact which was cloned when constructing the library.

4.2.5.2 Testing chromosome 7 genes as positional candidates for RP9.

The search for positional candidates continued by testing known genes that have been mapped to chromosome 7 for their presence in the RP9 disease region. Those that were excluded were subsequently assigned to chromosome 7 to confirm their sub-regional localisation, and to provide positional candidates for chromosome 7 linked diseases.

Adenylyl cyclases are integral membrane proteins which catalyse the formation of cAMP, and are second messengers for a variety of neurotransmitters, hormones and drugs. These enzymes are modulated by the formation of primary messengers via G proteins and in some cases via calmodulin (CaM). Type 1 adenylyl cyclase is CaM sensitive. Northern analysis indicated that it is expressed exclusively in the brain, retina and adrenal medulla, which suggests that the enzyme is neural
Figure 4.9 Part of the alignment generated in the BLAST analysis of the 908 bp cDNA sequence containing EST02120. This identified strong sequence identity with the human Tigger 1 transposable element consensus sequence over the entire length of the cDNA sequence.
Regional Assignment of Chromosome 7 ESTs

specific. The human form of this gene was originally mapped to 7p by in situ hybridisation by Villacres and colleagues (Villacres et al. 1993).

Oligonucleotides chosen from the 3' untranslated sequence of the cDNA (Genbank accession:35309; 173 bp product) gave no amplified product in the pool of YACs, therefore excluding Type 1 adenylyl cyclase (AC) from the RP9 disease region. At the onset of this study, the chromosome 7 hybrid panel was not available in the laboratory, therefore to further localise this gene on chromosome 7, a YAC 15C-C4 (1.1 Mb) was isolated from the ICI library by PCR screening (section 2.7.4) for the purpose of FISH mapping. However FISH analysis was not performed on the clone since the chromosome 7 hybrid panel had just then become available in the laboratory. PCR analysis of the AC STS with the chromosome 7 hybrid panel mapped this gene to 7p11.2-p13 (figure 4.2).

Acyloxyacyl hydrolase (AOAH) is a two subunit lipase present in leukocytes. Both the subunits are translated from the same ~2.2 kb mRNA molecule. It has been suggested that AOAH is involved in modulating the host inflammatory responses to gram-negative bacterial invasion, by hydrolysing the secondary fatty acyl chains of the lipopolysaccharide (endotoxins) found in the bacterial walls (Hagen et al. 1991). The human gene was sub-localised to 7p14-p12 by FISH using an AOAH containing cosmid (Whitmore et al. 1994). Although involvement of this gene in the pathology of RP9 is very unlikely, it was still useful to determine if this gene was present in the RP9 defined region and therefore to add to the transcription map. PCR using oligonucleotide primer pairs from the 3' untranslated sequence of the AOAH mRNA (Genbank acc: M62840; generates a 0.249 kb product) did not amplify in the YAC pools and proved the gene was not present in the RP9 defined region. Further localisation by amplifying the STS through the chromosome 7 hybrid panel mapped it to 7p13-p21 (figure 4.2), yet again confirming the localisation found by Whitmore and colleagues.

The aromatic hydrocarbon receptor protein (AHR) is a ubiquitous ligand-activated transcription factor that regulates the expression of a number of enzymes involved in many critical life processes by signal transduction mechanisms. It was found to be implicated in cell type-specific differentiation, cell division and apoptosis. It has also been seen to play a role in the metabolism of chemical carcinogens.
Le Beau and co-workers (1994) localised the AHR gene to chromosome 7 by PCR amplification using a gene specific STS primers on a monochromosomal somatic cell hybrid panel. It was then further localised to 7p21-15 by FISH. PCR with an STS primer pair flanking exon 10 of the AHR gene (Le Beau et al. 1994; sWSS2591) was used for amplification with the pool of YACs, but this excluded it from the RP9 disease region. It was then localised to the 7p13-p21 region by PCR through the chromosome 7 hybrid panel, confirming its localisation but not refining it.

Amphiphysin (AMPH; described earlier) was previously located to chromosome 7p13-p14 by FISH (Yamamoto et al. 1995a). Prior to mapping the related EST (EST06146), primers for a gene specific STS (Manfred Kilimann, personal communication; section 2.14.8) were used in a PCR reaction on the pools of YACs. This was not found to be present in the pools of YACs thereby excluding it from the RP9 disease region. It was then mapped through the chromosome 7 hybrid panel locating it to 7p13-p21 (figure 4.2) which confirms the FISH results but does not refine it. This localisation is also verified by the mapping of EST06146, which possesses high sequence homology to the amphiphysin gene, to the same region (sections 4.2.3 and 4.2.4). After finding these results in our study, the exclusion of amphiphysin from involvement with RP9 was also confirmed by Yamamoto and colleagues (1995b), who developed intragenic polymorphic RFLP markers for the gene that placed AMPH proximal to RP9 and the dominant cystoid macular (DCMD) disease region by genetic analysis.

In addition, both ADCYAP1R and GHRHR (see section 3.2.6), which are located on YACs in the contig that extends distal to the RP9 critical interval, were mapped through the chromosome 7 hybrid panel to confirm their localisation.

4.3 Discussion.

At the start of this study, the number of identified transcripts on chromosome 7 was low. In an effort to increase this collection, thirty chromosome 7 ESTs were mapped to twelve sub-regions, to benefit those searching for candidate genes for disorders mapping to chromosome 7. The positional candidate gene approach has been extremely successful in identification of many disease-causing genes (Collins et al. 1995). ESTs provide an important contribution towards this end, as recent
investigations have revealed that over 82% of genes mutated in human disorders which were identified by positional cloning efforts, are represented by significant matches with one or more ESTs in the dbEST division of Genbank (Genome Database, 1997).

Sub-regional assignment was accomplished by studying the segregation of PCR-amplified products in a well established chromosome 7 human-mouse somatic cell hybrid panel. EST sequences and STSs were obtained from several sources, which were mostly generated from The Institute for Genomic Research (TIGR, Maryland USA; Adams et al. 1991). In summary, eight mapped to the short arm, one to the pericentric region and twenty-one to the long arm. Although this number is insufficient to be representative, it portrays an impression of a relatively uniform distribution of transcripts along the length of chromosome 7.

Pairwise similarity searches against sequences in databases have revealed interesting information for a subset of these ESTs. Five of these are from known genes and in some cases verify their localisation. EST01644 was found to be part of the PRKAR1B gene. This encodes one of the regulatory isoform subunits of the holoenzyme which makes up the cAMP dependent protein kinase, and is ubiquitously expressed (Solberg et al. 1991). The function of PRKAR1B is clearly understood, being involved in the control of many biochemical events including ion transport and regulation of metabolism. EST06146 was associated with the human amphiphysin gene. As mentioned before, this protein is expressed widely in the nervous system and is believed to be involved in synaptic vesicle recycling. It has been found to be associated with the rare neurological autoimmune disease Stiff-Man syndrome, as it was discovered that a subset of patients with this disease accompanied by breast cancer seem to be positive for autoantibodies directed against amphiphysin (Yamamoto et al. 1995a) Its symptoms include progressive rigidity of body musculature with painful spasms. Amphiphysin is therefore a strong candidate for disorders associated with the nervous system. Recently, researchers cloned a novel isoform of amphiphysin, (amphiphysin II) by performing homology searches of amphiphysin against EST database sequences and found it expressed different splice variants (Tsutsui et al. 1997). This is just one example of exploiting EST sequence databases to find unidentified members of gene families. EST 06275 showed significant homology to the metabotropic glutamate receptor 3 gene (mGluR3).
MGlur3 is a member of a family of mGlurRs which are involved in glutamate neurotransmission and have been found expressed solely in neuron synapses among most brain cell types (Makoff et al. 1996). MGlurRs may be implicated in neurodegenerative disorders since abnormal glutamate transmission has been observed to cause neuronal death in patients with anoxia, hypoglycaemia and epilepsy (Makoff et al. 1996). The human Krit1 mRNA was found to be tagged by EST00272. Little information exists on the function of this gene, though it has been found to contain an ankyrin-repeat, a thirty three amino-acid motif found to be present in protein sequences across a wide range of phyla in numerous proteins of assorted functions, such as transcription factors and structural proteins (Michaely et al. 1993). EST00167 tags the human putative endothelial receptor type B-like mRNA. This gene has found to be abundantly expressed and distributed in various neuronal cells of the brain (Zeng et al. 1997). Mapping of 00167, provides a tentative new assignment for this gene. As with Krit1, its precise function is not known, but if these become disease gene positional candidates, associated information on their tissue distribution and expression patterns, will help to decide whether to prioritise these over others for further examination.

The majority of these ESTs represent sequences that show no homology to known gene sequences deposited in public databases. It remains to be determined whether these, and the ESTs that have homology to Alu, actually do represent true gene-based transcripts. As mentioned previously, Alu repeats have been found in approximately 5% of cDNA sequences (Yulug et al. 1995). However, there is always a chance that those sequences containing Alu could be genomic contaminants introduced when creating a cDNA library. Re-analysis of the EST sequences against those in public databases demonstrates the importance of repeating sequence alignments as the number of nucleotide sequence submissions increases at a rapid rate. Repeated examination in this case led to the detection in previously unidentified ESTs, of significant sequence homology to the human Krit1 gene, the human putative endothelial receptor type B-like mRNA, and several BAC clones.

The presence of the Mer 37 repeat sequence in the clone HFBCK70 of EST02120 suggests that this clone maybe a genomic contaminant of the library, a common problem in the production of many libraries. However, these repeats have
Regional Assignment of Chromosome 7 ESTs

been found in untranslated regions of cDNAs. Upstream and downstream sequence of this clone in genomic DNA is necessary to clarify the true origin of this clone. Although this EST is duplicated in the RP9 critical region, these primers define a unique STS which has been useful in physical mapping.

Soon after the publication of this map data information, Schuler and colleagues announced the results of a collaborative effort to sub-regionally localise the majority of existing ESTs in the Unigene database, in an effort to produce an integrated gene map of the human genome (Schuler et al. 1996; section 1.3.5). Of the 16,354 ESTs that were mapped onto either the Genebridge4 radiation hybrid panel, the higher resolution G3 radiation hybrid panel, or the integrated YAC-RH genetic map (Hudson et al. 1995; section 1.3.3.4), seven were from those localised in this study (ESTs 00022, 00085, 00122, 00167, 0439, 0446 and 0838). All these assignments confirm the localisations obtained in our study and provide a greater resolution of mapping information. Concurrent mapping is of great value when undergoing large operations. It is essential to check for discrepancies that may occur due to errors in various mapping applications since it has been estimated that 1% of assigned data is inaccurate (Schuler et al. 1996).

The transcript map of chromosome 7 has been greatly augmented by Eric Green and colleagues (Touchman et al. 1997) who generated 2,006 chromosome 7 ESTs by performing direct cDNA hybrid selection using chromosome 7 cosmids. Sequence comparisons have demonstrated that the majority of these represent new gene transcripts. Once these are mapped, those in the RP9 critical interval will be investigated as positional candidates for this disorder.

It is estimated that approximately 1,500 new EST sequences are submitted into dbEST every day (Boguski 1995). The level of redundant clones must therefore increase in proportion to this. Expression analysis would enhance the informative value of these ESTs. If a gene is identified as a positional candidate, similarity searches of the sequence to a redundant set of ESTs, may allow the assembly of an expression profile of the cDNA by observing the parent source library from which these are derived. For example if an EST sequence obtained from clones of a brain cDNA library was found to possess significant homology to EST sequences from liver library clones, it reveals some indication of where it is expressed. This is a simple and
rapid method for providing an idea of the tissue or cell type from which they were obtained. However, this will not replace more sensitive methods for analysing expression patterns. A full length sequence of a cDNA may be obtained by comparing it to redundant ESTs and forming overlapping contigs of sequences such as those in the Tentative Human Consensus (THCs) database (Adams et al. 1995) or UNIGENE (Schuler et al. 1996). Caution must be taken when further analysing ESTs as it has been revealed that a large proportion of 5' and 3' sequences in public databases that are assumed to derive from the same gene, are actually from separate transcripts (Wolfsberg and Landsman 1997). These errors are expected when dealing with a large output of sequences. The full sequence for many of the cDNA clones represented in dbEST are currently being generated. Studies have revealed that these sequences may also provide valuable information about mRNA splicing and modifications (Wolfsberg and Landsman 1997). In addition, ESTs may be used to detect previously uncharacterised gene alleles.

This approach has proved valuable in the route to reaching two of the goals of the human genome project: the complete transcript map of the human genome, and an STS map with a marker density of 1 every 100 kb. Most importantly, EST mapping efforts has greatly facilitated the positional candidate gene approach, which has now largely superseded other gene detection procedures employed in positional cloning.
Chapter 5. Sandwich Selection Hybridisation

5.1 Introduction

In a further attempt to identify transcripts within the RP9 disease region, a second approach was followed. cDNA hybrid selection methods have been successfully used to identify genes from specific genomic regions for disease gene identification and to add to the transcriptional map of the genome (section 1.3.4.4). A novel hybrid selection strategy, the sandwich selection hybridisation method (Yan and Swaroop 1994), was tested in this study to examine its capability as a new gene identification method, by using a control genomic source to “fish out” its respective cDNA. This technique was also applied to genomic clones from within the defined RP9 genomic interval in the hope of isolating cDNAs encoded by genes that reside within them. As this technique was relatively new and has not been widely discussed in the literature, the general principles, aims, strengths, limitations and refinements of this technique will be discussed.

5.1.1 The sandwich selection strategy.

The sandwich selection hybridisation method, developed by Yan and Swaroop (1994 & 1996), is a new technique for the selective isolation of cDNAs encoded by genomic DNA. It was developed to overcome some of the associated problems with the already established cDNA hybrid selection techniques (Lovett et al. 1991; Parimoo et al. 1991; section 1.3.4.4). This technique is thought to be advantageous over previously established hybrid selection methods, as it does not require any PCR steps, thereby eliminating any PCR bias in selected cDNAs. The cDNA source in this procedure is required in a single stranded plasmid form, therefore there is the possibility of selecting fuller length cDNA transcripts. After hybridisation, the selected single stranded plasmid is transformed directly into \( E. coli \) cells, removing the need for time consuming amplification and cloning of the selected material.

In summary, the selection procedure involves hybridisation in solution of (i) a cDNA library which is in a single stranded plasmid form, with (ii) the target genomic source, e.g. plasmids or cosmids which are subcloned into a specialised vector and are
in vitro transcribed to generate RNA probes that includes an SV40 sequence tag ("genomic RNA"), and (iii) biotinylated RNA capture probe ("capture RNA") which has an SV40 sequence tag that is complementary to that on the "genomic RNA". The "capture RNA" molecules and "genomic RNA" are hybridised together by the shared tag sequences. The "genomic RNA" acts as a "hook" to hybridise with strongly homologous cDNA sequences only. These cDNA- "genomic RNA"- biotinylated "capture RNA" hybrids are captured and retained on an avidin matrix. Non-specific cDNA products are washed off and the selected cDNAs are eluted from the matrix by digestion with ribonuclease A. This enzyme targets a single stranded RNA probe region in the sandwich hybrid complex to release the selected cDNAs. These are subsequently purified and transformed in *E. coli* cells for further analysis (see figures 5.1 and 5.2).

Genomic RNA is made by in vitro synthesis of template genomic DNA that is cloned in a specialised vector (pSV9Zf9). The genomic source of choice is fractionated to 0.2-1.0 kb size ranges by digestion using tetranucleotide recognition site restriction enzymes. These are then shot-gun cloned into the pSV9Zf9 vector (figure 5.2). This vector was designed specifically for this selection procedure (Yan and Swaroop 1994). It was constructed by cloning a 375 bp SV40 DNA fragment within the pGEM-9Zf (-) vector (Promega, Madison, WI), immediately after the SP6 promoter. Digested genomic DNA is cloned into an *Eco* RI site downstream of the SV40 sequence. After linearising the vector, SP6 polymerase is used for in vitro synthesis of RNA, transcribing both the SV40 tag and the genomic insert as one RNA transcript.

"Capture RNA" is produced by in vitro transcription of linearised pSV7Zf3 vector. This vector (figure 5.2), developed specifically for this strategy, was constructed by cloning a 395 bp SV40 fragment into the multiple cloning site of the pGEM-7Zf(+) vector (Promega), 225 bp of which is complementary to the SV40 tag in "genomic RNA". T7 polymerase is used for transcription, incorporating bio-11-UTP in the RNA product.
cDNA library in Charon BS phage vector (F episome E.coli host)  

Subclone genomic DNA into pSV9Zf9, and in vitro transcribe genomic template using SP6 RNA polymerase  

Single stranded circular cDNA  

"Genomic RNA" with SV40 tag  

Synthesise pSV7Zf3 capture RNA template using T7 RNA polymerase  

Excess Biotinylated "Capture RNA" with SV40 tag  

Preanneal with Cot-1  

Solution hybridisation  

(Biotinylated Capture RNA)  

(Genomic RNA)  

Capture on avidin matrix  

Capture on avidin matrix  

Wash off non-specific cDNAs  

Electroporation  

Direct analysis of clones  

Figure 5.1  
Summary of the sandwich hybridisation selection technique for the isolation of specific cDNAs (Yan and Swaroop 1994).
Figure 5.2
Diagram depicting pSV9Zf9 and pSV7Zf3 vectors (Yan and Swaroop 1994) and the hybridising structure of their RNA products.
Sandwich Selection Hybridisation

One main requirement of the Sandwich Selection technique is the necessity for the cDNAs to be available in a single stranded circular plasmid form. This is essential for the hybridisation step and also makes handling selected cDNAs easier due to the elimination of a cloning step after selection, and allows the direct analysis of clones after transformation. Charon BS (+/-) lambda vectors (Swaroop and Weissman 1988) combine the advantage of being in the form of a phage construct, thereby providing high efficiency cloning, ease in storage and screening, with the additional benefit of being able to convert into M13-based pBluescript (Stratagene cloning system, CA) phagemid form. This enables circular single stranded DNA to be generated using the helper phage R408. The Charon BS vector was constructed (Swaroop and Weissman, 1988) by cloning Not I digested and linearised pBluescript phagemid between Charon 15 lambda arms (Blattner et al. 1977, figure 5.3). Several directionally cloned, poly (A) primed cDNA libraries from various tissues and developmental stages have been created (Swaroop and Xu 1993) and are available from the American Type Culture Collection (ATCC, Rockville, MD).

5.1.2 Hybridisation and selection procedure.

Genomic RNA is first blocked using Cot-1 DNA to prevent the non-specific selection of intermediate and high copy number repeat containing cDNAs. Hybridisation is carried out in solution in the presence of 0.5 M sodium phosphate pH 7.2, 10 mM EDTA and 0.5% SDS in a final volume of 50 µl. A 2-10 fold molar excess of biotinylated “capture RNA” (approximately 20-50 µg) compared to “genomic RNA” is required. The “genomic RNA” in turn, is generally at a 10 to 100 fold molar excess (10-40 µg) compared to the single stranded cDNA (0.5-1 µg). The reaction is performed in a 200 µl eppendorf tube or in a sealed capillary, and incubated at 60°C for about 16-20 hrs to allow the formation of sandwich hybrids.

The resultant hybrids are captured on an avidin matrix via the biotinylated “capture RNA”. These are extensively washed to remove non-specific cDNAs. The selected clones are then eluted specifically using RNase A, which acts on the stretch of non-complementary SV40 single stranded RNA in the hybrid (see figure 5.2), to enzymatically release specific cDNAs. Two elutions are performed to ensure
Figure 5.3

Diagram depicting the Charon BS (-) vector and directionally cloned cDNA (Swaroop and Weissman 1988). Conversion to pBluescript allows amplification of the cDNA insert with M13 universal and reverse primers. Single stranded DNA can be rescued with the M13 fl ori, which recovers the antisense orientation of the LacZ gene.
all the selected material is gathered. Finally, the selected material is purified, precipitated and electroporated into *E. coli* host cells for direct analysis. A full description of hybridisation and washing conditions is given in sections 2.12.1 and 2.12.2.

### 5.1.3 Model sandwich selection experiments.

Model experiments performed so far by the developers of the method have primarily been conducted to investigate the success of the hybridisation method, and to a lesser extent to test its sensitivity. A series of model selection experiments (Yan and Swaroop, 1994) were set up using a genomic clone containing the human neural retina leucine zipper gene (NRL; Yang-Feng and Swaroop, 1992), in order to analyse the selection and isolation of the NRL cDNA clone from a complex mixture. A 5 kb genomic clone containing the NRL gene, was digested with *Rsa* I and *Mbo* I, and subcloned into pSV9Zf9. A human NRL cDNA in pBluescript M13 (-) was also constructed. The cDNA clone source used in the tests were either retinal cDNA library clones spiked with NRL cDNA, or human red opsin cDNA clones spiked with NRL cDNA. Both the library clones and the red opsin cDNA clones were in pBluescript M13 (-). The red opsin cDNA construct had been modified by the addition of a 1.6 kb *Not* I fragment containing the kanamycin resistance gene. Single stranded DNA was isolated from these. Selection steps and conditions were standard in each of the experiments, with the quantity of control cDNA and the source of complex cDNA mixtures being the main variables.

Initially an experiment was performed to demonstrate the selection potential under low complexity conditions. The target 5 kb genomic NRL containing clone was used to select its cDNA counterpart from a mixture of single stranded NRL cDNA (*amp* only) and red opsin cDNA (*kan*’ + *amp*’). Selected clones were plated onto kanamycin containing LB plates and ampicillin only plates in parallel. Clones that conferred both kanamycin and ampicillin resistance (red opsin containing cDNAs) grew on both sets of plates. NRL cDNA clones, which confer only ampicillin resistance, should not have grown on the Kanamycin containing plates. The differences in the number of clones that had grown on the respective plates were compared before and after selection. It was estimated by the authors that before
selection, the ratio of NRL to red opsin cDNA was approximately 1:40 (3,500 amp<sup>r</sup> clones: 137,500 kan<sup>r</sup> + amp<sup>r</sup> clones, see table 5.1). After selection, the first elution showed the NRL cDNA clones had been selected with an approximate ratio to the red opsin cDNA clones of 20:1 (1,000 amp<sup>r</sup> clones: 50 kan<sup>r</sup> + amp<sup>r</sup> clones). The authors suggest that it presents an increase of around 800 fold. In the second elution, no kan<sup>r</sup> + amp<sup>r</sup> colonies had grown (100 amp<sup>r</sup> clones only), which the authors argue represents total selection for NRL clones. As the genomic and cDNA source was not complex, this experiment demonstrates the upper limit of selection that can be achieved with this method.

Table 5.1 Increase of NRL cDNA after selection.

<table>
<thead>
<tr>
<th>Experimental stage</th>
<th>Kan&lt;sup&gt;r&lt;/sup&gt; + amp&lt;sup&gt;r&lt;/sup&gt; (Red clones only)</th>
<th>Amp&lt;sup&gt;r&lt;/sup&gt; (NRL)</th>
<th>Approximate ratio Red :NRL</th>
<th>Percentage NRL in total library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before selection</td>
<td>137,500</td>
<td>3,500</td>
<td>40:1</td>
<td>2.5%</td>
</tr>
<tr>
<td>Elution 1</td>
<td>50</td>
<td>1,000</td>
<td>1:20</td>
<td>95%</td>
</tr>
<tr>
<td>Elution 2</td>
<td>0</td>
<td>100</td>
<td>-</td>
<td>100%</td>
</tr>
</tbody>
</table>

When evaluating the success of the procedure in subsequent control experiments using more complex cDNA sources, only the quantity increase in the percentage of target cDNA in the starting, compared to the selected library, is being taken into consideration. The sandwich selection procedure does not increase the amount of selected material by amplification, but selects it.

A second experiment was performed to highlight the number of background clones that may filter through to the selected material. The target genomic source was more complex in this case, using DNA from a flow sorted X chromosome library, with NRL cDNA spiked into a retinal cDNA library. The human NRL gene maps to chromosome 14, and therefore should not be selected in this experiment. The proportion of NRL clones in the starting and selected library was determined by colony lifts of plated transformed bacterial colonies and hybridised with an NRL cDNA probe. Prior to selection, NRL clones comprised 26% of the cDNA mix (95 out of 364 clones tested). The proportion had decreased to 2.3% after selection (2 out of 86
total clones analysed in elution 2). This demonstrates that there is some level of NRL background clones passing through the selection procedure, but it is low.

**Table 5.2** Negative selection: reduction of NRL cDNA clones with unrelated “genomic RNA” (from Yan and Swaroop 1994).

<table>
<thead>
<tr>
<th>Experimental stage</th>
<th>Total clones analysed</th>
<th>NRL positive clones</th>
<th>Percentage NRL positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before selection</td>
<td>364</td>
<td>95</td>
<td>26%</td>
</tr>
<tr>
<td>Elution 1</td>
<td>43</td>
<td>2</td>
<td>4.6%</td>
</tr>
<tr>
<td>Elution 2</td>
<td>86</td>
<td>2</td>
<td>2.3%</td>
</tr>
</tbody>
</table>

Two later experiments were performed to assess the level of selection when the quantity of target cDNA is altered. Here a retinal cDNA library was spiked with NRL cDNA to two different levels. The genomic target source used was the 5 kb genomic clone. Quantitative analysis of selected and original material was performed with colony lifts and hybridised with NRL probe as before. In the first experiment, the amount of NRL in the starting library was 12% (6,000 out of 50,000 clones). After selection, approximately 99% (3,600 out of 3,650) of the clones recovered in the first elution were NRL. This gives an 8.25 times quantity increase of NRL from the starting to the selected material. In a second experiment, 1.7% (400 out of 23,600 clones) of the original library comprised of NRL clones, and 33% (500 out of 1,500) of the clones in the selected library consisted of NRL, demonstrating a 19.4 times quantity increase. The investigators have not mentioned results achieved from the second elution.
Table 5.3 Positive selection: selection of NRL cDNA from a spiked retinal cDNA library using NRL "genomic RNA" (adapted from Yan and Swaroop 1994).

A.

<table>
<thead>
<tr>
<th>Experimental Stage</th>
<th>Total clones analysed</th>
<th>Number of NRL positives</th>
<th>Ratio of NRL : library</th>
<th>Percentage NRL</th>
<th>Quantity increase of target clones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before selection</td>
<td>50,000</td>
<td>6,000</td>
<td>1:8.3</td>
<td>12%</td>
<td>-</td>
</tr>
<tr>
<td>Elution 1</td>
<td>3,650</td>
<td>3,600</td>
<td>1:1.01</td>
<td>99%</td>
<td>~8.25 x</td>
</tr>
</tbody>
</table>

B.

<table>
<thead>
<tr>
<th>Experimental Stage</th>
<th>Total clones analysed</th>
<th>Number of NRL positives</th>
<th>Ratio of NRL : library</th>
<th>Percentage NRL</th>
<th>Quantity increase of target clones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before selection</td>
<td>23,600</td>
<td>400</td>
<td>1:59</td>
<td>1.7%</td>
<td>-</td>
</tr>
<tr>
<td>Elution 1</td>
<td>1,500</td>
<td>500</td>
<td>1:3</td>
<td>33%</td>
<td>~19.4 x</td>
</tr>
</tbody>
</table>

The aim of these model experiments was to demonstrate the feasibility of the procedure, and to highlight its potential for isolating genes. The effect that the complexity of the genomic target material has on the specificity of selection has not yet been tested in a controlled manner. However, the authors have used this method to isolate a clone from a YAC spanning OTC, DXS140, within the RP3 disease region and will be discussed in section 5.2.11.3.

5.1.4 Adult Retina Library in Charon BS.

The adult retina library (AR2) in Charon BS vector was chosen for use in the selection procedure in this study. It was favoured in preference to a fetal eye library, due to its availability in the laboratory and after the comparison of insert sizes with each of the libraries (section 5.2.2). The AR2 library was also chosen because it has been enriched for tissue specific genes by subtractive hybridisation against a JY lymphoblastoid cell line (Swaroop et al. 1991).

At the time of construction (Swaroop and Xu, 1993), the investigators examined the quality of the library. The range of cDNA insert sizes in the library were determined by isolating and digesting DNA from at least eight random clones, and size separating on agarose electrophoresis gels. To ascertain the nature of the cDNA clones, they then screened a gridded library with class I HLA genes. Positive plaques were
isolated, converted into pBluescript KSM13(-) phagemid, and then sequenced from both ends. Most of the cDNA clones were full length. Approximately 80% of the clones contained three or more deoxyadenosines generated from the poly(A) site.

5.1.5 Choice of genomic clones for use in selection.

Cosmids that had previously been mapped within the genetically defined region of RP9 were pooled and used as a genomic source. A colleague had isolated these by screening a chromosome 7 gridded cosmid filter (Nizetic et al. 1994) with probes from the disease region. Cosmids were chosen over YAC clones due to the problems associated with the large molecular size of YACs. YAC clones have been observed to exhibit chimerism, rearrangements or deletions. Cosmid clones on the other hand are known to be more stable, and as their host is bacterial, the isolation of pure cosmid DNA is generally easier than that of YACs. Contamination of yeast chromosomal DNA from the YAC host has been known to be a major problem in previous selection experiments (Lovett 1991).

5.1.6 The blue cone pigment gene as a model for sandwich selection.

At the time of selection, no genes had been identified in any of the cosmids from the RP9 region. Therefore to assess if the selection process was effective, a control experiment was designed where a cosmid containing the blue cone photopigment (BCP) gene was used for the selection of its respective cDNA. The BCP gene encodes the blue cone opsin pigment, one of the three light sensitive photopigments that detect photons of light as the fundamental basis for the process of vision (Nathans et al. 1986). These are integral membrane proteins, maximally sensitive at different wavelengths, which together mediate human colour vision. The BCP gene has been mapped to chromosome 7q31.3-32 (Fitzgibbon et al. 1994). Blue cone photoreceptor cells are present at a density 150 times lower than rhodopsin photoreceptor cells in the retina, and the BCP cDNA represents approximately 1 in 30,000 clones in a human retinal cDNA library (Nathans et al. 1986), a frequency 150 times lower than rhodopsin. However, in the subtracted AR2 library, BCP represents approximately 1 in 1,125-1,250 clones (section 5.2.2).
The first control experiment was performed to test if selection with the BCP cosmid would be successful at isolating the BCP cDNA from the subtracted AR2 library. A second experiment involved spiking the AR2 library with BCP cDNA to a higher abundance level, to provide an indication that selection was indeed occurring, because this selection procedure has not been tested for its sensitivity when the cDNA species is represented at 1 in 1,125-1,250. A lorist B cosmid containing the human BCP gene, was donated by Dr. B. Appukuttan (Appukuttan 1997). The BCP cDNA was isolated by PCR amplification of a retinal cDNA library using gene specific primers (section 5.2.5).

5.1.7 Aims.

As this selection procedure is still in the developmental stage, the aim was primarily to analyse and determine the potential of this system using the model control sources, while also applying it to genomic clones from the RP9 region. The model experiments performed by the developers did not investigate the effect on selection when a complex genomic source was administered as the target material, as is the case when applying this technique to genomic clones from physically mapped disease regions. Control experiments were carefully followed through at each stage, and will be presented here, considering that the selection technique is novel and not established. The advantages, limitations and possible routes for improvement will also be discussed.
5.2 Results.

5.2.1 Amplification of AR2 library.

The AR2 bacteriophage library stock was amplified to prepare large quantities of DNA for conversion to the pBluescript phagemid form (section 5.2.3). This was later used to create single stranded cDNAs to administer in the selection procedure. Growth of recombinant phage in solution may cause a bias in the representation of certain cDNAs, as some clones may be eliminated due to competition from faster growing ones. Amplification was therefore performed on a solid media (section 2.10.3). An initial titration of the amplified laboratory stock of AR2, indicated there were $12 \times 10^9$ pfus/ml (section 2.10.2). A total of approximately $3 \times 10^6$ pfus were plated onto several 22 cm x 22 cm Nunc plates to ensure representation of the majority of cDNA clones. The resulting recombinant phage were eluted off and combined to provide a general phage library stock (section 2.10.3). Titration was repeated after amplification, demonstrating that the amplified library contained approximately $1.6 \times 10^{11}$ pfus/ml.

5.2.2 Library characterisation.

At each stage of phage library manipulation, the library was tested to determine whether the complexity of insert sizes was sustained, as there are a number of steps involving amplification. Swaroop and co-workers had determined the size range of inserts in the library at the time of construction, by preparing DNA from eight random clones, digesting out the insert and fractionating on an agarose gel (Swaroop et al. 1993). The AR2 library had been estimated to have on average >1 kb inserts. We investigated the size range of inserts of the amplified library by PCR-amplification using primers flanking the cloning site of the Charon BS(-) vector. This method is less time consuming, involves less steps than digesting random clones of prepared phage DNA, and also permits a larger number of clones to be analysed simultaneously. Approximately 100 individual random plaques were picked, after plating an aliquot of the original library at a density that permitted clear separation of individual plaques (section 2.10.2). Single plaques were isolated by removing a plug of agar containing the plaque and transferring to 20 µl of SM buffer. Phage particles were allowed to
diffuse out into the media at room temperature for several hours. An aliquot of this was used for PCR using the pBluescript vector specific primers SK+ and T3 (sections 2.7.1, 2.7.2 and 2.14.8). As the Charon BS phage is constructed with linearised pBluescript that is cloned between phage arms, only the vector specific primers SK+ and T3 will amplify the inserts (see diagram 5.3).

Approximately 65% of those clones that did amplify, possessed inserts. Such a low percentage of recombinants has been observed by the creators of the library, who argue that it is a result of the simultaneous enrichment of non-insert containing molecules during the subtraction procedure (Swaroop et al. 1991). Non-recombinant clones produced a standard band of 118 bp, which is the distance between the two primer pairs flanking the multiple cloning site (figure 5.4). The size of inserts varied.

A fetal brain (FB2) library (Swaroop and Xu 1993) was also tested this way. It was found to have an even lower proportion of recombinant clones (~45%), which were also smaller in size range (data not shown). On this basis it seemed preferable to use the AR2 library in the selection process.

As the BCP cDNA was used in control experiments (section 5.2.5), it was necessary to determine the number of BCP cDNA transcripts that were present in the AR2 library. Nathans and co-workers (Nathans et al. 1986) observed that BCP cDNA was present in approximately 1 in 30,000 clones in a retinal library. $10^5$ pfus from AR2 were plated onto XL-1 Blue *E. coli* host cells on each 22 cm x 22 cm Nunc LB agar plate, supplemented with tetracycline (section 2.10.1 and 2.10.2). Two replica plaque lifts were prepared for each plate. The BCP probe was obtained by PCR amplification of a cDNA clone (see section 5.2.5) with primers BCP1+ and BCP5- (sections 2.14.8 and 2.7.1). The probe was purified by passing through a sephacryl S-200 HR spin column (section 2.1.2.3) and was labelled by random priming (section 2.5.1). Hybridisation was carried out at 65°C, as described in sections 2.6.2 and 2.6.3. The probe was first tested on a Southern blot of fractionated *Hind* III digested genomic DNA to ensure it was specific. It was then hybridised to the plaque filters. From this, BCP was estimated to be present at a density of approximately 1 in 1,125-1,250 (~0.088%) in the phage form of the AR2 library (figure 5.5). This is about 30 times higher than that observed by Nathans and colleagues (Nathans et al.
Figure 5.4 Determination of the percentage of recombinants by PCR amplification of inserts from random clones of the AR2 retinal library. The inserts of individual plaques were PCR amplified with pBluescript SK+ and T3 primers. Non-recombinant clones produced a standard product of 118 bp. Roughly 100 clones were tested, and an estimated 65% possessed inserts.
Figure 5.5  A representative section of the AR2 phage library screen by hybridisation with a BCP specific probe. The arrows point to positive plaques that are represented in both replicas. A total of 4 X 10 pfus were screened and the abundance of BCP cDNA in the library was estimated as 1 in 1,125-1,250 (0.088%).
Sandwich Selection Hybridisation

1986), but this can be accounted for as AR2 is a subtracted library.

As an additional check to ensure the library was representative for retinal expressed sequences, PCR amplification (sections 2.7.1 and 2.7.2) was performed with the amplified library as a template, using primers for several known retinal genes that are expressed at different levels in the retina. Oligonucleotides chosen from the cDNA sequence of rhodopsin, peripherin, and the blue cone opsin gene gave the expected sized product from all stages of AR2 library manipulation (data not shown).

5.2.3 Conversion of AR2 library to pBluescript.

The pBluescript phagemid contains the fI phage intergenic region required for the synthesis of single stranded DNA. The phagemid can be easily recovered from the Charon BS vector by digesting with Not I restriction enzyme for phagemid release and then self-ligated to circularise it (Swaroop and Weissman 1988). Several 50 ml amplified library stocks were required to obtain a high yield of bacteriophage phage DNA (section 2.10.4). Quantity and purity was estimated by OD measurements at $A_{260}$ and $A_{280}$. At least four 20 µg samples of the library DNA were digested with 20 units of Not I to release the phagemid (section 2.2). Sample aliquots of the digested products were separated onto 0.8% agarose gels (section 2.3.1). The Charon left and right arms are clearly visible as two discrete bands of sizes 21.6 kb and 18.25 kb respectively, which migrate at close proximity to each other (figure 5.6a). This is observed when the clones have been excised, thereby confirming complete digestion. Due to the high molecular weight of the phage arms, visibility of these two fragments was possible on a gel stained with ethidium bromide. However, as the released phagemid library comprising different sized inserts, are of a lower molecular weight, they cannot be visualised in this way.

To be certain that the digestion was complete and that the phagemid was released as expected, a simple check was performed before proceeding onto ligation. Aliquots of the digested samples were fractionated by gel electrophoresis alongside undigested AR2 phage DNA and appropriate size markers, and were transferred to a nitrocellulose filter by Southern blotting (sections 2.3.1 and 2.4.1). This filter was hybridised with [$\alpha^{-32}$P]dCTP labelled digested pBluescript vector probe...
Figure 5.6 *Not* I digested DNA samples of the AR2 library.
Lane 1 and 2  5 µg of digested samples from DNA preparation 1
Lane 3 and 4  2.5 µg of the digested samples from DNA preparation 2
Lane 5 and 6  1 µg of the digested samples from DNA preparation 2
Lane 7  1 µg undigested library DNA.

a) The high molecular weight left and right Charon arms are visible on the 0.8% agarose gel. The excised library in pBluescript cannot be seen.

b) A Southern blot hybridised with a vector probe allows visualisation of the excised library in pBluescript. Non-recombinant pBluescript clones are seen as a band at 2.961 kb. The smear directing up from this represents recombinant cDNA library clones (variable sized cDNA inserts in pBluescript). The DNA of preparation 1 (Lanes 1 and 2) appeared to be degraded and were not used in further experiments.
Sandwich Selection Hybridisation (sections 2.5.1 and 2.6). The results are seen in figure 5.6b. A band of 2,961 bp is clearly detected on the autoradiograph, whereas it is not visible on the agarose gel. This represents the proportion of non-recombinant pBluescript phagemid vector. A smear directed upwards indicates the range of inserts. The smear is denser at the 4 kb range (cDNA insert sizes of 1 kb in a 2.9 kb plasmid), which may confirm the 1 kb average insert size observed by the library constructors (Swaroop et al. 1993). An absent downward smear, indicative of degradation, substantiates the quality of preparation 2.

The digested material was subsequently ligated under low concentrations to recircularise the phagemid (methods 2.10.5). All of the ligation samples were transformed by electroporation into XL1-Blue MRF' cells for single stranded rescue (section 2.8.4). 100 µl aliquots of the transformed culture were plated out onto selective LB agar plates supplemented with ampicillin for insert analysis (section 2.8.5 and next section). The remaining culture was stored in glycerol stocks or was immediately used for the production of single stranded DNA (sections 2.10.6 and 5.2.4).

5.2.3.1 Insert analysis and confirmation of circularisation

A PCR screen on transformed AR2 pBluescript clones was conducted to verify that conversion was successful. M13 universal and reverse primers are located on the pBluescript vector, situated at positions that are flanking the cloning site. When the vector is linearly cloned into the Charon BS vector, they are positioned facing away from each other. Only when the pBluescript vector has been released and circularised, may amplification of the cloning site and prospective inserts be possible with these primers (figure 5.3). Random clones were picked and tested by PCR assay for the size and presence of inserts. When the transformants were plated onto media containing X-gal and IPTG (section 2.8.5), only ~60-65% of the clones were white in colour, confirming the observation that the library was poorly represented with recombinants. Individual white colonies were randomly picked with a sterile toothpick and placed in 10 µl of enzyme incubation solution for the disruption of cell walls (section 2.7.2). 2 µl of this was used in a standard PCR reaction with M13 forward and reverse primers (section 2.7.1 and 2.14.8).
No considerable change was observed in the range of insert size estimated at this stage compared with recombinants before circularisation. A 227 bp product indicative of the cloning site when no insert was present, was seen in approximately 40% of cases (figure 5.7). The consistency of results indicated no apparent bias in selection was taking place, which is common when conducting several ongoing manipulations with clones. This assay was repeated several times when further steps with the libraries were carried out to ensure that clone bias was not introduced.

5.2.4 Single stranded DNA rescue.

pBluescript KS (-) phagemid permits the secretion of single stranded DNA when clones are co-infected with helper phage (section 2.10.6). The AR2 cDNA library is directionally cloned into Charon BS (-) (Swaroop 1993). Secretion of the (-) strand of pBluescript will result in recovery of the phagemid DNA in the antisense orientation of the LacZ gene, hence the cDNA insert equivalent of the mRNA strand in the directionally cloned library. These represent the cDNA source in the selection procedure.

Single stranded cDNA was secreted and prepared from the supernatant of the culture of the AR2 library, and non-recombinant pBluescript as a control (section 2.10.6). Double stranded phagemid DNA was prepared from the pellet of cells in parallel (section 2.1.1). Aliquots of double stranded DNA alongside single stranded DNA, were then electrophoresed through a 1% agarose gel to confirm that single stranded DNA had been rescued (section 2.3.1). As seen in figure 5.8, single stranded DNA molecules run at a different mobility than double stranded DNA. They adopt a different conformation to their double stranded counterparts due to the formation of secondary structures and tend to run faster. A clear shift in migration is visible between the double stranded and single stranded preps. The 2.961 kb single stranded pBluescript vector DNA migrates in a non-denaturing agarose gel at approximately 1.1 kb. The 6.4 kb R408 helper phage migrates at the range of a 3.0 kb double stranded size marker, and is clearly visible in the lanes containing the single stranded DNA. In preparations of the cDNA library, a smear is representative of cDNA inserts in the vector. In addition, single stranded BCP DNA (section 5.2.5) was sequenced without denaturation, to confirm the isolation of single stranded DNA.
Figure 5.7 Amplification of random clones from the pBluescript form of the AR2 retinal library. PCR with M13 universal and reverse primers would only amplify the inserts if the library was successfully converted from the phage to the pBluescript form. Non-recombinant clones produced a standard product of 227 bp. Roughly 100 clones were tested in total, which confirmed that insert size distribution and the percentage of recombinants was maintained.
Figure 5.8  Electrophoretic comparison of single and double stranded DNA preparations on a non-denaturing 1% agarose gel. Single stranded DNA migrates faster than double stranded DNA. R408 helper phage is present in the single stranded preparations.

Lane 1 and 2  Double stranded undigested AR2 phage library DNA.
Lane 3  Single stranded AR2 phage library DNA.
Lane 4  1 kb ladder (Gibco).
Lane 6  Double stranded undigested pBluescript vector DNA.
Lane 7  Single stranded pBluescript DNA.
Lane 8  R408 helper phage.
5.2.5 Construction of model sandwich selection experiments.

A lorist B cosmid clone, containing the full length BCP gene was obtained from Dr. Appukuttan. To verify that the cosmid did contain the full length BCP gene, primers amplifying exon 1 to exon 2 (BCP 1+ and 2-) and exons 4 to 5 (BCP 4+ and 5-) (see section 2.14.8), were used for PCR amplification from cosmid DNA. Electrophoretic separation on a 1% agarose gel (section 2.3.1) gave an expected band of 804 bp for exon 1 to 2. This consists of exon 1, intron 1 and exon 2 genomic BCP sequence. Also, an expected size band of 1,342 bp was produced for exons 4 to 5 (data not shown). This consists of exon 4, intron 4 and exon 5. Like the other cosmids used in selection it was shotgun cloned into pSV9Zf9, then in vitro synthesised to “genomic RNA” (sections 2.11.2 and 5.2.8). There are eight Rsa I and four Sau 3AI sites in the BCP gene, leaving all except exon 4 intact.

A human retinal cDNA library (Nathans et al., 1986) was used as a template for the amplification of the BCP cDNA. The BCP primers (section 2.14.8) were selected from positions directly flanking the start codon (oligonucleotide BCP1+) and the polyadenylation recognition site (oligonucleotide BCP5-) of the gene. PCR was performed with Vent<sup>®</sup> polymerase (New England Biolabs) in the presence of 4mM MgSO<sub>4</sub> (sections 2.7.1) to generate templates with blunt termini. An expected size product of 1,043 bp was recovered. This blunt ended PCR template was ligated into Eco RV digested and dephosphorylated pBluescript II KS(-) vector (section 2.8.2). After electroporating into XL1-Blue MRF' cells (section 2.8.4) the resulting libraries were plated out onto standard selective LB agar plates with ampicillin and tetracycline for selection of recombinants (section 2.8.5). 25 white clones were randomly picked and phagemid DNA was prepared by mini-preparations (section 2.1.1). These were checked to confirm they contained the correct size insert by PCR amplification with the cDNA specific primers, BCP 1+ and BCP 5-. 24 clones contained the correct size insert. As an additional verification that the desired insert was obtained, two clones were partially sequenced from both ends using the M13 universal and reverse primers. This sequence was compared and aligned to the published sequence (Nathans et al., 1986). Roughly 160 bp from the 5' end and 140 bp from the 3' end of the gene was obtained for each clone, all of which correlated exactly to the published sequence, therefore confirming the integrity of these clones. The cDNA orientation of both these
clones was such that the 3' end of the gene neighbours the T3 promoter end, and the 5' end is near the T7 promoter region. Single stranded cDNA was produced from one of these clones, as described in 2.10.6.

5.2.6 Choice of genomic clones from the RP9 region.

At the time, several cosmids had been mapped within the RP9 critical interval. These did not form a contig, but were binned in the region. These were chosen as genomic templates for sandwich hybridisation selection in parallel to the control experiments. All cosmids are in the lawrist 4 vector. J0117 and F0249 were isolated with the STS 3586RA (see figure 3.16 for STS order); L0651 was isolated with 3029LA; 1211RA is in E1924 and F1049, the latter also contains D7S460. M1040 contains 3586D and EST02120. Overlaps between cosmids, were determined by restriction digest analysis.

5.2.6.1 Vector assay for cross-contamination.

As the LacZ (-) strand of the pBluescript vector is present in the selection reaction as part of the single stranded cDNAs, it could introduce a problem of hybridising with similar species of sequences which may be present as part of the cosmid vector in the "genomic RNA". For the genomic clones in both Lawrist 4 and Lorist B vectors, the genomic inserts could not be isolated away from the cloning vectors easily as there were no single rare cutter restriction sites flanking the cloning site. Therefore the complete genomic clone, including the vector DNA, was digested and subcloned. At the time of study, it was uncertain whether vector DNA of cosmids would hybridise with pBluescript vector of the cDNA library. This could cause products to be falsely selected. To investigate this, cosmid DNA with genomic inserts, uncloned cosmid vector DNA, and pBluescript DNA as a positive control, were digested and separated onto a 1 % agarose gel sections (sections 2.2 and 2.3.1). This was Southern blotted (2.4.1) and the filter was hybridised to random labelled digested pBluescript probe at 60°C (sections 2.5.1 and 2.6). The results indicated that pBluescript did not hybridise to the cosmid vector when genomic DNA is cloned. However there was homology with the uncloned cosmid vector, though only to portions of the cosmid vector which are removed during construction of the clone (not
shown). Therefore using cosmids as the genomic source should not cause any disruption in specificity during selection.

5.2.7 Cloning genomic DNA into pSV9Zf9 vector.

4 µg of pSV9Zf9 vector was linearised by Eco RI digestion (section 2.2). Staggered termini were filled in by Klenow polymerase to create blunt ends, and were then dephosphorylated (section 2.8.1). The enzyme was inactivated by the addition of EGTA to a final concentration of 0.05 M and incubation at 65°C for 45 min.

2 µg of each cosmid DNA was digested to completion with four base pair recognition cutters, Rsa I and Sau 3AI in independent reactions (section 2.2). These frequent cutter enzymes were chosen as they produce fragment sizes of a manageable length for the production of *in vitro* synthesised probes. The cosmid digests produced fragments ranging from 0.2 kb to 2.5 kb. Two separate sub-libraries were created with different enzymes to provide full representation of cosmid inserts. Rsa I enzyme results in blunt ended cleavage, therefore digested DNA was used directly in a ligation reaction. However Sau 3AI digested products had to be pre-treated with Klenow polymerase to generate blunt termini prior to ligation (section 2.8.1). Both the vector and insert DNA were subsequently purified by phenol:chloroform extraction, and then precipitated (section 2.1.2). To confirm that the digestion was complete and to estimate the quantity of DNA, samples were separated on 1% agarose gels (section 2.3.1) alongside known molecular weight standards. The quantity of DNA was also determined by O.D._260_ measurements. Ligation was performed such that the ratio of molar ends of the vector to insert was 3:1 (vector: insert) (section 2.8.1). Blunt ended, dephosphorylated pSV9Zf9 vector was included in a self ligation reaction as a negative control, and uncut vector was used as a positive control to monitor the efficiency of transformation.

These ligations were used to transform XL1-Blue *E. coli* cells by electroporation (section 2.8.4). After recovery of transformed cells in 1 ml SOC, 10 µl aliquots were taken for the analysis of recombinants (section 5.2.7.1). 4 ml LB with ampicillin was added to the recovered transformed culture, and the samples were left to grow overnight at 37°C to amplify the library. Plasmid clone DNA from the
amplified culture was isolated (section 2.1.1) to provide templates for \textit{in vitro} synthesis of RNA (section 5.2.8). The negative control produced few transformed clones, as expected when plated onto LB agar plates supplemented with ampicillin.

\textbf{5.2.7.1 Genomic sub-library evaluation.}

Recombinant clones of pSV9Zf9 cannot be identified on the basis of blue white colour selection. This is due to the construction of the vector. The SV40 fragment was cloned into the \textit{Hind} III site in the multiple cloning region of the pGEM9Zf(-) vector (Promega, USA). This causes insertional inactivation of the LacZ gene, hence eliminating the application of the blue white selection (methods section 2.8.5).

Since the number of recombinants could not be determined by eye, random clones were picked from each sublibrary and tested by PCR assay as described previously (section 5.2.3.1). This test was done to verify that shotgun cloning was successful and developed no bias for selection of insert size. 10 µl aliquots of the transformed sublibrary grown before and after growth in LB were plated onto standard selective LB agar plates supplemented with ampicillin (section 2.8.5). The PCR screen was performed using M13 universal and reverse vector primers (section 2.7.1 and 2.14.8) and at least 25-50 clones from each cosmid sub-library were tested. The sizes of amplified inserts were determined after separating on 1% agarose gels (section 2.3.1 and figure 5.9). PCR products of non-recombinants, gave a fragment size of 0.575 kb which includes the vector and SV40 sequence. As estimated, the sizes of inserts ranged from 0.2 kb to 1 kb, taking into account the 0.575 kb additional sequence. Approximately 90 % of the clones tested were recombinant. No single band dominated the random clones that were analysed indicating that no bias for size selection had occurred during ligation. This assay was done for each cosmid sub-library at each stage in its manipulation to ensure there was no introduction of preference for certain sized clones.

\textbf{5.2.8 In vitro synthesis of “genomic RNA”}.

Pooled products from the cosmid sub-libraries were linearised by digestion to prepare the templates for run off transcripts during \textit{in vitro} synthesis of “genomic RNA”. It has been reported that \textit{in vitro} transcription using templates which have been linearised using restriction enzymes that produce 3' overhangs may cause undesired
Figure 5.9  PCR amplification of random clones from the genomic shotgun cloned libraries in pSV9Zf9, using M13 universal and reverse primers. Non-recombinant clones generated a product of 575 bp. Approximately 90% of clones possessed inserts.
synthesis of the complementary sequence of the insert and succeeding vector DNA (Schenborn and Mierendorf 1985 and Promega Riboprobe *in vitro* transcription systems, USA). To avoid this occurring, *Sal I* restriction enzyme was therefore chosen for linearisation of the vector, producing 5' overhangs. The recognition site is 8 bp away from the *Eco RI* cloning site, thereby producing run off transcripts immediately after the insert. This is important in the selection procedure as synthesis of these unwanted transcripts may cause superficial hybridisation to the pBluescript DNA of the single stranded library, thereby selecting non-specific cDNAs. Unwanted transcripts also exhaust the rNTP source, therefore reducing the quantity of target product. Aliquots of the digested products were run on a 1% agarose gel to ensure complete digestion had been achieved.

Transcriptions of genomic DNA were performed on pooled sub-libraries, for convenience and for a reduction in the number of samples to synthesise. Cosmids J0117, F0249 and L0651 were pooled into group A and cosmids F1049, E1924, M1040 into group B. After incubation with proteinase K and subsequent purification with phenol chloroform, samples were resuspended in nuclease free dH2O (section 2.11.2). Quantitative measurements and purity of samples were determined by O.D.\textsubscript{260/280} ratios. Approximately 5 \( \mu \)g of the cosmid sublibrary template was used to synthesise "genomic RNA" by SP6 polymerase (section 2.11.2). To purify the synthesised RNA, the template DNA was first degraded by incubation with DNase I, then unincorporated nucleotides and enzymes were removed by phenol:chloroform extraction and ethanol precipitation (sections 2.11.2.) The quantity of RNA produced was determined by spectrophotometric analysis at O.D\textsubscript{260}. 5-10 \( \mu \)g of the samples was electrophoresed on a denaturing agarose gel (section 2.11.3) to confirm transcript size and to evaluate the level of degradation. As seen in figure 5.10 transcription products from the BCP cosmid sub-library are represented as a smear of products from 0.417 kb upwards, indicative of genomic inserts. This confirms transcription has occurred. The 0.417 kb product represents transcription products from the SP6 transcription site to the *Sal I* site in pSV9Zf9, inclusive of the SV40 fragment, which are from non-recombinant clones. An insignificant amount of RNA degradation was observed, confirming the quality of the transcriptions.

\[198\]
Figure 5.10 *In vitro* transcribed RNA products separated on a denaturing formamide gel.

Lane 1  Genomic RNA from the BCP genomic sublibrary synthesised using SP6 polymerase. The smear directing upwards consists of transcription products from the library inserts.

Lane 2  RNA ladder (Gibco BRL)

Lane 3  Biotinylated RNA products generated from digested pSV7Zf3 vector template using T7 polymerase. The target band is 0.486 kb. Additional non-specific products of higher molecular weight may derive from partially digested template DNA.

Lane 4  The 0.535 kb biotinylated RNA product generated from a PCR template of the target capture sequence.
5.2.9 *In vitro* synthesis of biotinylated "capture RNA".

Biotinylated RNA was synthesised from the pSV7Zf3 plasmid using T7 RNA polymerase. Plasmid DNA was prepared by the miniprep method (section 2.1.1). Several 10 μg aliquots were digested to completion with *Pvu* II and *Sac* I (section 2.2), as recommended by the authors, to produce a template for synthesis. Complete digestion was confirmed by fractionating aliquots through a 1% agarose gel (section 2.3.1). *Sac* I is not an ideal enzyme for linearisation of the template as described in section 5.2.8. Taking this into consideration, the products were filled in with Klenow polymerase and dNTPs to create blunt ends. This is an additional step for prevention of extraneous transcripts. Digested DNA products were purified with phenol:chloroform after treatment with proteinase K solution as described previously. Transcription was then carried out with T7 polymerase with UTP substituted for biotinylated-11-UTP (section 2.11.2). Several transcriptions were performed to produce large amounts of target RNA. The samples were purified by spinning through G-50 Sephadex Quick Spin columns. Quantification was determined by spectrophotometry as described previously. Confirmation of biotinylation was determined by running 5-10 μg aliquots onto a denaturing formamide agarose gel (section 2.11.3). The presence of products would indicate biotinylation had occurred as RNA products would not have been transcribed without the incorporation of Bio-11-UTP. As shown in figure 5.10, a band of the expected size (0.486 kb) was synthesised. In addition several bands of higher molecular weight were produced and are likely to be spurious products derived from partially digested templates.

These unwanted products should not interfere with the selection process but deplete the rNTP source, thereby reducing the quantity of target transcribed SV40 RNA. Therefore an alternative method was also employed, *in vitro* transcribing RNA from a PCR template which contained only the sequence of interest. M13 forward and SP6 primers were used to amplify a product, using Vent™ polymerase, from the pSV7Zf3 vector that included the T7 promoter and recognition site and the SV40 target DNA. The origin of the product was confirmed by size confirmation on an agarose gel and by sequencing. The product was purified and transcription was performed as described (section 2.11.2). Using this procedure, the dominant product was the expected band of 0.535 kb (figure 5.10).
5.2.10 Selection procedure.

For all selection reactions in this study cosmids instead of smaller plasmid clones, were used as the genomic source. As explained in section 5.1.3, the selection procedure has only been tested to a limited extent by its developers, principally to demonstrate its feasibility through certain model experiments. The lowest quantity of control cDNA to capture in the starting library that had been tested was to an abundance of 1 NRL clone in 59 retinal library clones (1.7%). This is extremely high and does not represent moderate or low transcript levels of cDNAs in common libraries. The genomic source in this experiment was from a 5 kb plasmid clone. This too does not represent situations where larger more complex clones are used as a genomic source. Although the tests performed by the developers demonstrate the success of the selection process, they do not address the extent of its sensitivity, especially when applying the technique to YACs or cosmids, clones which are more commonly used in physical mapping efforts. The control cDNA selection experiments in this study were devised to verify that selection was indeed taking place when it was applied to cosmids.

The BCP gene is usually present at 1 in 30,000 cDNA clones in a retinal library (Nathans et al. 1986), but represents approximately 1 in 1,125-1,250 clones in the normalised AR2 library (section 5.2.2). The first selection experiment involved using a cosmid containing the BCP gene, to select its cDNA counterpart from the AR2 library (section 5.2.5). The objective of this was to determine if complex genomic clones can select cDNAs that are represented at a relatively high level in the library. The “genomic RNA” of this cosmid was mixed with “genomic RNA” from cosmid Ml 040 to increase the complexity of the “genomic RNA” source with the aim of reflecting the conditions of the selection with pooled cosmids from the RP9 region.

As the selection procedure had not been tested for this level of sensitivity, it was necessary to have a control experiment which assures that selection was occurring at a level which is detectable. Therefore a second control selection experiment was performed, where the BCP containing cosmid was used to select its cDNA counterpart, from the AR2 library which was spiked with the BCP control cDNA (section 5.2.5). Again, the “genomic RNA” of this was mixed with “genomic RNA” of cosmid M1040. A calculated amount of the BCP single stranded cDNA was added to the
Sandwich Selection Hybridisation

single stranded cDNA from the AR2 library. The abundance of control cDNAs in the starting library was 0.375% (15:4,000 clones analysed, see fig. 5.11). This was determined by transforming an aliquot of the spiked library prior to selection and plating out onto large nylon filters overlaid on LB agar plates (section 2.8.4 and 2.8.5). Replica colony lifts were made from this and were hybridised with the BCP cDNA probe (2.5.1 and 2.6).

Finally, two more selection experiments were performed using the pooled cosmids from the RP9 region to “fish out” specific cDNAs from the AR2 library. Cosmids J0117, F0249 and L0651 (pool A) were used in one selection experiment and cosmids F1049, E1924 and M1040 (pool B) were used in a second experiment.

Selection was carried out as described by the developers of the method (Yan et al. 1996; see section 2.12.1 and 2.12.2 for full details). The concentration of RNA and single stranded DNA was accurately measured at O.D.\_\text{260} (1 OD = 40 µg/ml) for RNA, and (1 OD = 33 µg/ml) for single-stranded DNA. For each selection experiment, there was a 2-10 fold molar excess of “capture RNA” than “genomic RNA”, which itself is of a 10-100 fold molar excess compared to single stranded cDNA. Generally, 40 µg “capture RNA”, 20 µg “genomic RNA” and 1-1.5 µg of single stranded DNA was used for selection. The “genomic RNA” was first blocked for repeat sequences with Cot-1 DNA for two hours prior to selection. Single stranded cDNAs and biotinylated RNA were then added to the pre-blocked samples, and selection was performed at 60°C for at least 16-20 hours in 0.5 M sodium phosphate pH 7.2, 10 mM EDTA and 0.5% SDS.

5.2.11 Analysis of selected clones.

After the selection reaction, the mixture was transferred to an avidin matrix in binding buffer for capturing the selected clones. The matrix was washed several times to remove non-specific cDNAs and the captured material was released by incubation with RNase A. The selected material was finally purified, precipitated and resuspended in 10 µl. Aliquots of these were electroporated for direct analysis.
Figure 5.11 A representative sample of the Southern blot hybridisation screen using a BCP cDNA specific probe onto duplicate filters of the starting spiked AR2 library. Out of a total of ~4,000 clones screened, 15 were identified as BCP cDNAs by hybridisation and confirmed by PCR with BCP specific primers. The abundance of BCP cDNAs to AR2 clones was therefore estimated as ~1 in 266 (0.375%).
5.2.11.1 BCP controls.

To assess the success of the selection procedure quantitatively, for both BCP control experiments, several 2 µl samples of the final selected material were electroporated and together spread over a large nylon filter (Hybond-N+), overlaid on an LB agar plate. These were left to grow overnight. Two replica colony lifts on Hybond-N+ were prepared for each plate (section 2.4.2) and the number of BCP positive clones was determined by hybridisation with a BCP cDNA probe that was PCR amplified using oligonucleotides BCP1+ and BCP5- (section 2.14.8 and 2.7.1). The probe was prepared as described in sections 2.1.2.3 and 2.5.1. To determine if the selection was a success, the percentage of BCP clones in the starting library (section 5.2.10) was compared with those in the selected library.

In the first control selection experiment, using the BCP containing cosmid mixed with additional cosmid DNA for selecting BCP cDNA from the AR2 library which had not been spiked, approximately 2,300 clones were obtained from a representation of the selected material from elution 1, and approximately 800 from elution 2. Unfortunately, the hybridisation revealed no positive signals for BCP. A control, denatured BCP cDNA probe dotted onto a small filter, was included in the hybridisation to ensure labelling and hybridisation was effective. This illustrates that the selection was unsuccessful in these circumstances. One possibility for this could be because the proportion of BCP cDNA in the starting library may have decreased to a lower level than the estimated amount, during the steps involved in conversion of the phage library to single stranded phagemid form. This could be a result of competition from other clones, which reduced it to an amount to which this technique is not sensitive enough to detect. The abundance of BCP in the AR2 library in the phagemid form or the single stranded form was not determined and may have been worth checking to see if clone bias had actually occurred. A control experiment using a complex genomic source to isolate a target cDNA that was represented at this abundance had not previously been tested. The lowest level of control target cDNA in a library that the investigators had tested for was with an abundance of 1 NRL in 59 clones (1.7% NRL) in the starting library, using a 5 kb genomic target source. This selection technique may just not be as sensitive for the lower quantity tested in this study. A second control experiment was therefore set up in this study to test the
sensitivity of this procedure.

To test the selection of BCP cDNA in a spiked retinal library, using the BCP containing cosmid mixed with DNA of another cosmid clone from a different chromosomal region as the genomic source, the selected material was screened for the presence of BCP cDNAs as before. As shown in table 5.4, out of a total of 4,200 clones analysed in elution 1, 35 BCP positives were detected (figure 5.12a). This indicates BCP clones comprise 0.833% of clones in the selected library in elution 1 (an abundance of 1 BCP in 120 clones). This is a 2.2 fold percentage increase from the starting spiked library. Elution 2 gave 25 positive clones out of a total of 2,200 analysed (figure 5.12b). Therefore this indicates BCP clones comprise 1.13% of the selected library in elution 2 (1 BCP in 88 retinal library clones), a percentage increase of 3.01 fold from the starting library. This clearly indicates that selection is occurring, although the level is low. However, when compared to the selection results from Yan and Swaroop’s controls, this decrease in sensitivity can be accounted for by the lower amount of BCP cDNA in the starting library, and the increased complexity of the genomic source used in this experiment. This clearly suggests that both the complexity of genomic source and the abundance of target cDNA in the starting material are critical. However, this selection has not been tested with spiked cDNA at the same level as Yan and Swaroop’s controls.

Table 5.4 Selection of BCP cDNA from a spiked retinal cDNA library using BCP containing “genomic RNA”.

<table>
<thead>
<tr>
<th>Experimental stage</th>
<th>Total clones analysed</th>
<th>Number of BCP Positives</th>
<th>Ratio of BCP to library clones</th>
<th>Percentage BCP</th>
<th>Quantity increase of target clones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before selection</td>
<td>4,000</td>
<td>15</td>
<td>1:266</td>
<td>0.375%</td>
<td>-</td>
</tr>
<tr>
<td>Elution 1</td>
<td>4,200</td>
<td>35</td>
<td>1:120</td>
<td>0.833%</td>
<td>2.2 X</td>
</tr>
<tr>
<td>Elution 2</td>
<td>2,200</td>
<td>25</td>
<td>1:88</td>
<td>1.13%</td>
<td>3.01 X</td>
</tr>
</tbody>
</table>

All positive clones were picked from the plates and confirmed as BCP clones by PCR with 1+ and 5- primers. Several additional clones were sequenced to ensure they were true BCP positives. When the selected material was plated out, the level of
Figure 5.12a Southern blot hybridisation of a BCP cDNA specific probe onto duplicate filters of the selected material from Elution 1 of the BCP spiked control. Out of a total of 4,200 clones screened, 35 were identified as BCP cDNAs by hybridisation and were confirmed by PCR with BCP specific primers. The abundance of BCP cDNAs to AR2 clones in the selected library was therefore estimated as ~1 in 120 (0.833%), a 2.2 fold increase from the starting library.
Figure 5.12b Southern blot hybridisation of a BCP cDNA specific probe onto duplicate filters of the selected material from Elution 2 of the BCP spiked control. Out of a total of 2,200 clones screened, 25 were identified as BCP cDNAs by hybridisation and were confirmed by PCR with BCP specific primers. The abundance of BCP cDNAs to AR2 clones in the selected library was therefore estimated as ~1 in 88 (1.13%), a 3.01 fold increase from the starting library.
blue colonies, indicative of non-recombinants had decreased to approximately 10%. This is still a large amount considering the technique should not have selected these clones. The possible homology between lawrist 4 and lorist B cosmid vectors and the pBluescript vector of the single stranded DNA was tested (section 5.2.6.1) and showed no detectable cross-hybridisation. A possible explanation could be the contamination of the cosmid DNA preparations with pBluescript or related sequence, which could have been subcloned into the pSV9Zf9 vector. Alternatively, the washing was not satisfactorily completed.

5.2.11.2 Sequence Analysis of random clones from the RP9 selected library.

5.2.11.2.1 E2-1.

This clone (elution 2, clone 1) was randomly chosen from the selected material in the experiment with pool A cosmid clones used as the genomic target. Double stranded template DNA was prepared, and the complete sequence (1,051 bp) was obtained by automated sequencing using vector primers and additional internal sequencing oligonucleotides (E2-1A and E2-1B; section 2.14.8 and figure 5.13). A poly (A)n tract begins at nucleotide position 1,035 and two putative consensus polyadenylation signals (AATAAA) are present 18 bp and 68 bp upstream from this. A search for coding sequences revealed that it has several small open reading frames (ORF), but none that were significant. As the orientation of the cDNA is known, it is likely that this sequence may be part of a large 3' untranslated region and does not represent a full length transcript. It is possible that one potential ORF, from nucleotide position 1 to 174 in frame 1 on the sense strand, may extend further 5' upstream and actually be part of a coding region. This sequence was searched for amino acid similarity using BLASTX in the SwissProt database. However, no significant homology was detected.

To determine if this clone was truly selected from sequences in the cosmids that were used in the procedure, an STS was chosen from the E2-1 sequence for use in PCR analysis (E2-1C and E2-1D; section 2.14.8 and figure 5.13). This STS and the STS from primers E2-1A and E2-1B, did not amplify in any of the cosmids used as
Figure 5.13 Nucleotide sequence of clone E2-1. M13R and M13F refer to the vector primers used for initial sequencing. Boxed sequences represent the putative polyadenylation recognition sites. Primers E2-1A and E2-1B were used for generating the full length sequence. E2-1C and E2-1D were used as an STS for chromosomal assignment.
templates in this selection experiment. E2-1 may therefore represent a non-specific clone passing through to the selected material, or may show significant homology, not detectable by STS mapping, to a sequence in the cosmids, such as a member of a gene family. To determine its true localisation, the STS was used for PCR on the HGMP monochromosomal hybrid panel (Kelsall et al. 1995). No amplification was observed in the rodent DNA only control, confirming its specificity. The actual map location for E2-1 was found to be on chromosome 4. Even though this does not map to the RP9 region, it represents a cDNA from a brain library and can become a candidate for diseases mapping to this chromosome if its localisation is further refined.

Clone E2-1 was analysed for homology to other sequences (after removing the poly (A)ₙ tract) using the FASTA algorithm against the Genbank, EMBL, cDNA, SwissProt and repeat databases. This revealed that E2-1 possessed considerable nucleotide identity (>80%) with several human EST sequences in the dbEST database. In addition, strong identity was observed with two mouse ESTs (discussed below).

EST826227/zp62a10.r1 exhibits the top scoring alignment to E2-1. This represents the 5' sequence from a clone (IMAGE: 624762) that was derived from an umbilical vein endothelial cell line cDNA library. The complete sequence for this EST (526 bp) has 91.7% identity with the 5' end of the E2-1 sequence (E2-1 nucleotide position 18-544, with the introduction of gaps by FASTA; figure 5.14). This is one of the many ESTs generated by the Washington University/Merck EST project who single pass sequenced the 5' and 3' ends of directionally cloned cDNAs. This group, like others, submits their ESTs to Genbank without manual editing. They predict these retain a typical sequence accuracy of 98%, which may therefore affect the alignment score. The 3' end of clone 624762 (EST821057/zp62a10.c1), has 87% identity over its complete 340 bp sequence, with the 3' end of E2-1 (figure 5.14). It is possible that E2-1 and clone 624762 are cDNAs from different members of a gene family, as the nucleotide sequence homology was very significant, but not precise enough, to suggest that E2-1 represents the same cDNA clone as 624762.

Two mouse ESTs, vd94g09.r1 and vd88b04.r1, showed striking nucleotide identity with internal E2-1 sequence (81.5% over 557 bp overlap and 86.4% over 434 bp overlap respectively, with the introduction of gaps by FASTA; see figure 5.15). Both were generated from the 5' end of directionally cloned mouse heart cDNAs, and
Figure 5.14 Sequence similarity alignment of E2-1 with human ESTs 826227/zp62a10.r1 and 821057/zp62a10.sl, using the FASTA programme against the Genbank database. These ESTs were generated from the 5' and 3' ends of the human cDNA clone 624762 respectively, by the Merck/Washington University EST project. The alignment of zp62a10.sl is with the antisense strand of E2-1.
Figure 5.15 Sequence similarity alignment of E2-1 with mouse EST vd94g09.r1 using the FASTA programme against the Genbank database.
their exact sequence identity suggest they are derived from the same mRNA, although vD94g09.r1 is a longer sequence. The 3' respective sequences of these clones had not been submitted to Genbank at the time of analysis. This cDNA could possibly represent the murine homolog for the E2-1 cDNA.

Other human ESTs that were homologous to E2-1 include clones that were derived from brain, breast, lung and cochlea libraries. All these ESTs were generated from directionally cloned sequences, and most show 83-87% nucleotide identity with the extreme 3' end of the E2-1 sequence. The corresponding 5' ESTs, where available, possessed the same percentage identity to E2-1 but did not extend further upstream than E2-1 or EST826227. None of these exhibit amino acid similarity to known genes or proteins. Clone E2-1 sequence was also analysed in the TIGR database for Tentative Human Consensus sequences (THCs; Adams et al. 1995). This presents aligned groups of strongly homologous redundant human ESTs, including all those generated by TIGR and most of those represented in dbEST. These have been aligned to form Tentative Consensus Human sequences (THCs). Query sequences can be searched for homology to these THCs. Those with significant identity can be compared for sequence accuracy and may even extend its length further. Strongly homologous THCs can also provide expression information from the source library of the EST clones and abundance of the EST transcripts, and can also identify alternative splice forms. The 3' sequence of E2-1 displayed 83% nucleotide identity and 85.9% amino acid similarity to THC175367. All the redundant ESTs that form the THC had already been identified by FASTA and did not extend further (see figure 5.16). The highest scoring THC for the 5' end of E2-1 showed only 69% identity and 78% similarity. A number of ESTs present in THC175367 were also represented in Unigene (Schuler et al. 1996; section 1.3.5), but this did not provide any additional information. The top scoring alignment, EST826227 was not represented in either the Unigene or the THC database. No significant homology of E2-1 was detected with searches against the repeat database or with BLASTX against the SwissProt database.

As E2-1 has been excluded as a candidate for RP9, further analysis on this clone was not performed. Nevertheless, this clone represents a cDNA expressed in a retinal library. With fine mapping, E2-1 can be considered a positional candidate for
Figure 5.16 The position of the top scoring sequence alignments for E2-1. vd94g09.r1 and vd88b04.r1 are mouse ESTs generated from the 5' end of directionally cloned cDNAs. The highest sequence homology was found with zp62al0.rl, which represents the 5' sequence of a directionally cloned human cDNA. The EST representing the 3' end of the same cDNA is zp62al0.sl (represented in bold in the Tentative Human Consensus (THC) alignment box). E2-1 has 74% nucleotide identity with the consensus sequence of THC 175367 (not shown). Sequences that are part of the THC individually possess 83-87% nucleotide identity with E2-1.
chromosome 4 linked diseases and would contribute to the transcription map. The isolation of the full length cDNA sequence should result in the identification of a statistically significant open reading frame, and provide coding sequence that may assign a putative role through amino acid similarity searches.

5.2.11.2.2 E2-19.

This clone (elution 2, clone 19), was randomly chosen from the selected material in the experiment where pool B cosmids were used as the genomic target. PCR using vector primers indicated the insert was ~535 bp. Double stranded DNA was prepared, and 296 bp of nucleotide sequence was obtained by single pass sequencing using the M13 reverse vector primer (figure 5.17). Four deoxyadenosines, possibly part of a poly (A)\textsubscript{n} tail, are present at the start of this sequence. A putative polyadenylation signal is 19 bp upstream from this.

Sequence similarity searches of this sequence using the FASTA algorithm against the dbEST and cDNA databases revealed it possessed striking homology (98-100% nucleotide identity) with a number of ESTs. These include ESTs from clones that were derived from retina, brain, pineal gland, testis, and B cell mRNA. The top scoring alignment was with EST nz28e07.sl. This is from a clone derived from B cell mRNA and has 99.324% identity with the complete E2-19 sequence, although others represented 100% identity over shorter regions. The top scoring THC alignment with E2-19 is with THC212875 (99% nucleotide identity and 99.7% amino acid similarity). This THC includes all the top scoring ESTs that had already been identified to possess homology with E2-19 in Genbank and dbEST (figure 5.18). Most of these ESTs align with E2-19 at the 3' end, but possess longer poly (A)\textsubscript{n} tails. This clearly indicates that E2-19 is derived from the same cDNA as these ESTs and most likely represents its 3' untranslated end. However, one EST (80876) extends 23 bp further downstream from the poly d(A)\textsubscript{n} site. This was derived from the 5' end of a directionally cloned cDNA. It is possible that this transcript is alternatively spliced. No poly(A)\textsubscript{n} tail or consensus AATAAA site was present in this extra sequence. However, the possibility that this extended sequence may actually represent a chimeric portion of its parent clone cannot be discounted.

Most sequences in this THC are also represented in the Unigene cluster
Figure 5.17  Partial nucleotide sequence of clone E2-19. The complement of the sequence is shown. M13R refers to the vector primer that was used in single pass sequencing. The boxed sequence represents the putative polyadenylation recognition sites. A001W23F and A001W23R are the primers used for mapping E2-19.
Figure 5.18  a) Nucleotide identity of E2-19 with THC212875, and b) THC 212875 structure and content. Most of the redundant ESTs that together form the THC begin at nucleotide position 24 (after removal of poly d(T) tracts). The three deoxyadenosines of E2-19 were not removed for the similarity search. The highlighted red sequence indicates the position of the E2-19 partial sequence in the THC.
Sandwich Selection Hybridisation

Hs.7535. An STS specific to sequences in this cluster (STSG4476) has been mapped through the Genebridge 4 radiation hybrid panel. This has assigned the Unigene cluster to chromosome 9q21.12, and is an estimated 76.4 Mb from the 9p telomere. As Hs.7535 contains most of the THC sequences that are exactly identical to E2-19 sequence, its placement can suggest that E2-19 is excluded from the RP9 region. An STS (A001W23 GB accession: G19715) specific to the THC 212875 sequence has also been generated by TIGR, but has not yet been mapped. The primer sequences matched the E2-19 sequence exactly (figure 5.17). To confirm that this clone is truly excluded from the RP9 region, the STS was used for PCR assay on the cosmids which were used in the selection experiment, in case they possessed any undetected chimerism. This STS, did not amplify in any of these cosmids, thereby confirming the exclusion.

Like E2-1, this clone seems to have passed through the selection experiment and is not a true selected cDNA. However, it is a cDNA that is shown to be expressed in a wide variety of tissues, based on the source library of the homologous EST clones. Its location to 9q21.12 makes it a positional candidate for diseases mapping here. No strong protein matches have been identified with sequences that make up the THC, however it is noted in Unigene, that Hs.7535 sequences possess weak similarity to the S. cerevisiae hypothetical 48.1 kD protein. The full sequence and further analysis of this clone was not performed due to its exclusion.

5.2.11.2.3 Additional selected material.

As selected clones were sequenced, it became apparent that a large number of inserts were not from true cDNAs but actual contaminants containing pieces of vector DNA inserts. Seven clones which were partially sequenced, appeared to contain the same sequence that derived from the ColE1 origin of replication, present in plasmid vectors. The start of the insert sequences had identity to nucleotide position 1,070 in the pBluescript vector. The insert in the single stranded form of these clones would be recovered in the M13 (-) orientation, the same as the pBluescript vector. The origin of these clones is unknown, but they are most likely to have been derived from contaminants which were present and subsequently cloned into Charon BS during the
stage of library construction. The reason why these had passed through to the selected material is unknown. The generation of the full sequence and further analysis of these clones were not pursued. When choosing selected clones, the blue white colour assay would not have identified these. Contamination of cDNA libraries from a number of sources is a common occurrence and is recognised as a major problem in cDNA selection procedures (Lovett 1994).

A number of possibilities could have caused the passing through of non-selected cDNAs into the final elutions. For example, non-stringent washing conditions may have resulted in the elution of cDNAs which are not captured by the hybrids. In addition, the Vectrex matrix may not have been quenched sufficiently with salmon sperm DNA before capture and washing, and could have resulted in non-specific retention of some cDNAs in the matrix. However, the washing conditions were followed thoroughly and the specificity of the enzymatic elution under the relatively high salt and low temperature conditions should not have released any cDNAs that are not captured specifically by the hybrids. Also the repeated washing and the large volume of wash solution (10 ml) is sufficient to remove most non-hybridising cDNAs. Nevertheless, all the established techniques have had considerable amounts of non-specific selected material passing through (Lovett et al. 1991), and the analysis of many more clones are needed to identify truly selected clones that map back. Due to a limitation in time, only a handful of clones could be analysed in this study, and further analysis on these selected clones are being continued by a colleague.

5.2.11.3 Sequence analysis of clone E1-10.

To date, Swaroop and co-workers have isolated one novel clone which they claim maps back to the original genomic target source that was used in this procedure (personal communication, 1997). A 400 kb YAC clone (55B), from the RP3 region on Xp21.1, was used in the selection procedure after being isolated and separated away from the yeast chromosomal host DNA by PFGE. Rsa I and Mbo I digested YAC DNA was subcloned into the pSV9Zf9 vector. Selection was performed with a retinal cDNA library, after blocking the "genomic RNA" with Cot-1 DNA. Clone E1-10 (elution 1, clone 10), was randomly picked from the selected library and its insert size was estimated as 2.6 kb by restriction digestion. The authors stated that they had
mapped the cDNA clone back to the target YAC (55B), after hybridisation of the E1-10 insert onto Southern blot filters containing digested products of this YAC clone and several cosmids from the RP3 critical region (personal communication).

Partial sequence analysis of this clone was subsequently conducted in this project as part of a collaboration. Initial sequencing was performed using T7 and T3 primers. Additional primers were chosen from these to generate further sequences (primers 1 and 2 from the T7 and T3 ends respectively; 2.14.8). The E1-10 insert contains 2 Eco RI sites and yields insert fragments of ~1,400 bp, ~900 bp, and ~350bp (figure 5.19). The ~1,400 bp and ~900 bp fragments are released from the E1-10 clone when digesting with Eco RI only. These were subcloned into pBluescript, then partially sequenced to generate extra internal sequencing primers (primers 3, 4 and 5; 2.14.8). The sequences generated from these were used to construct a sequence contig of 1,791 bp, joining the two subclones sequences together and linking these with the T3 end sequence of E1-10 (figure 5.21). However this sequence was not long enough to link up with that from the T7 sequence (574 bp; figure 5.20). The generation of the full sequence was not completed upon request from the collaborators.

Upon sequencing it became apparent that the insert was not directionally cloned. It appeared that the vector had been cut with just Hind III during cloning, but not also with Eco RI. Both the T7 and T3 insert ends have at least two deoxyadenosines at the first nucleotides. When the linkers which should be used for directing the orientation of the insert in respect to poly (A)_n tails, are ligated to these, it generates Hind III sites at both ends. Therefore the clone can be ligated in any orientation. Only three deoxyadenosines are present at the start of the sequence generated from the T3 end. No polyadenylation consensus site was present upstream from this. Two deoxyadenosines were at the start of the T7 sequence. No consensus polyadenylation sequence (AATAAAA) was found, but a variant of the signal (AAGAAAA) at nucleotide position 60-55 was present, which is thought to function as a less efficient signal (Wickens et al. 1990).

The partial sequence was analysed for homology to other sequences using the FASTA and BLAST algorithms against Genbank, repeat and cDNA databases.
Figure 5.19  The primer positions and orientation of sequences in the E1-10 insert. The horizontal arrows represent sequencing primers. The insert is not directionally cloned, and is flanked by two Hind III sites. The numbers in italics show the sequence orientation referred to in figures 5.20 and 5.21.

Figure 5.20  Partial nucleotide sequence from the T7 end of clone E1-10. This sequence is part of the ~1,400 bp Eco RI fragment. The sequence in red represents sequencing primer 1.
Figure 5.21 Partial nucleotide sequence from the T3 end of clone E1-10. The 906 bp and ~1,400 bp subclones were partially sequenced to generate additional sequencing primers (represented above). These formed a sequence contig that linked the subclones and the T3 end of E1-10, to form the above sequence.
The highest homologies scored were to sequences that contained L1 repetitive elements (long interspersed repeat elements). These account for about 5% of the human genome and there are approximately 50,000 to 100,000 copies (Moyzis et al. 1989). Full length L1 elements of mammalian genomes are between 6 kb and 7 kb long. However, 95% are truncated at the 5' end. Other internal substitutions, deletions and rearrangements are common and there are estimated to be only ~3,500 full length copies in the genome. These elements are widely dispersed in human chromosomes and frequently occur in introns and untranslated regions of mRNA. Similarity searches against the repeat databases, revealed that all the sequence generated from E1-10 possessed significant homology to a number of L1 repeat sequences. However, there was not identity over its complete sequence to just one element, as certain rearrangements and deletions, common with L1 repeats were present. Figure 5.22 presents the alignment to part of E1-10, which shows 74% identity in a 557 nucleotide overlap with the L1ma4a repeat. There was no long stretch of sequence in E1-10 which did not show homology to L1. From these alignments it is evident that the E1-10 homology, is near the 3' end of the L1 repeat.

L1 repeats retain several sequences which are thought to encode proteins. No significant open reading frames were detected in the E1-10 sequences. It is uncertain whether or not this clone is truly derived from a cDNA. L1 repeats are associated with the untranslated regions of some mRNAs. Sequencing further upstream and downstream from this should provide additional clues. No obvious poly (A)_n tail was present which would have indicated the orientation of the clone. Also there are questions arising from the true map location of this clone. The repeat element in this clone would have caused problems in obtaining specific hybridisation. It is also possible that this clone derived from genomic contamination during library construction. Further analysis and completion of the clone sequence is being undertaken by the collaborators.
Figure 5.22 An example of the sequence alignment of the T3 end sequence of clone El-10 by FASTA against sequences in the repeat database. The entire length of the El-10 sequence displayed homology to L1 repeat sequences.
5.3 Discussion.

The method tested in this study, demonstrated a detectable level of selection in the control experiment where the starting library was spiked with BCP cDNA to an abundance of 0.375% (1 in 266). Although the percentage increase in number of target clones is far lower than that observed in Swaroop and co-workers controls (section 5.1.3), it can be argued that this is an effect of the increased complexity in genomic target template and the lower quantity of target cDNA in the starting library. In Yan and Swaroop’s control experiments, the genomic target was a 5 kb NRL containing plasmid, whereas in this experiment we used pooled cosmids. This plasmid clone is at least 10 times smaller in length, providing more genome equivalents of the NRL gene than any other gene, if present, in the cosmids. These investigators argue that the 5 kb clone used in their controls represents a “complex genomic mixture” as sequence data has shown it contains a stretch of repetitive elements and an Alu repeat. However, database analysis has revealed it represents low homology to these repeats and only constitutes a small proportion of the clone. In addition, there are no other genes in the 5 kb subclone that may have decreased the percentage of NRL to non-NRL clones in the selected library (if their cDNAs were represented in the starting library). Their control experiments show the maximum levels of selection that the procedure can achieve. The main aim of the sandwich selection technique is to select specific transcripts rather than enrich for them by PCR amplification, therefore the levels of enrichment are not as high as the established hybrid selection methods (section 1.3.4.4). When applied to more complex genomic sources such as the cosmids tested here, the level of selection in the eluted material was not as high as that observed in the developers model experiments (section 5.1.3). This does not benefit the isolation of rare transcripts, and these therefore may not be selected.

In the control experiment performed in this study where the library had not been spiked, the abundance of BCP in the starting library was ~0.088% (1 in ~1,125-1,250). This is almost 4.5 times lower than the level in the spiked control. At this level, no selection for BCP cDNA was observed, however the estimate was obtained by screening the starting library prior to amplification and conversion to single stranded cDNA. The large number of manipulative steps involved with conversion could have
caused a skewed reduction in the level of BCP clones due to competition from other clones. The lowest quantity of control cDNA in the starting library that had been tested in Yan and Swaroop’s controls were 1.7% NRL clones in the library (1 NRL in 59 retinal cDNA library clones). This is almost 20 times higher than the level of BCP in the non-spiked starting library, and 4.5 times more than the level of BCP in the spiked library in the experiments performed in this study. At the abundance of 1.7% NRL clones in the library, using a 5 kb genomic clone, Swaroop and co-workers achieved a 19.4 times increase. This suggests that both the complexity of the genomic template and the level of cDNA in the library affects the success of selection.

PCR-based cDNA hybrid selection techniques (Lovett et al. 1991, Parimoo et al. 1991 and Tagle et al. 1993) use vector specific oligonucleotides to amplify cDNA library inserts as a source for the selection, thereby eliminating many time consuming steps. This has the advantage that any library of choice may be used, as long as the vector sequence flanking the cloning site is known. Alternatively if there is not a library of choice available, it is possible to add linker-adapter oligonucleotides to freshly synthesised cDNAs (Morgan et al. 1992). This would also reduce the number of steps in amplification. The cDNAs and the genomic source are double stranded and are denatured before mixing together. The genomic template and cDNA sources in the Sandwich Selection procedure are single stranded, thereby eliminating competition for hybridisation from their antisense strands. The use of single stranded cDNAs also permits their direct analysis after selection as there is no need to clone them. It also supplies fuller length clones as there are no PCR steps that can result in smaller sized inserts due to PCR bias. Unfortunately one is limited in the choice of cDNA libraries available as they must be capable of rescuing single stranded circular cDNAs. A variety of libraries have been constructed in Charon BS (Swaroop and Xu 1993) which are available from the American Type Culture Collection (Rockville, MD). In addition a large number have been constructed in λZap vectors (Stratagene Cloning Systems) which can also secrete single stranded DNA’s. However, if there is a direct requirement for a tissue or developmental stage, which is not represented in these libraries, one must be constructed in the appropriate vector. Another concern, is the problem of time consuming conversion of the phage library to pBluescript, followed by the subsequent recovery of single stranded DNA. These steps can introduce the risk
of clone bias, loss of complexity, and most likely, the loss of rare transcripts. λZap libraries do not require conversion to double stranded pBluescript prior to single stranded secretion, therefore may reduce any clone bias occurring. It is necessary that the cDNA vector is in the (-) orientation for fl to avoid sequence similarity with the genomic and capture vectors. The pGEM derived pSV9Zf9 and pSV7Zf3 vectors are M13 based but transcription products are also in the (-) orientation so they should not hybridise with each other.

The resulting sandwich tripartite hybrid in the selection technique is believed to provide a higher efficiency for retention onto a solid matrix. RNA-DNA hybrids are supposedly more stable than DNA-DNA hybrids. The maintenance of a ribonuclease free environment is required for the selection and subsequent washing steps. This is easily achievable in the laboratory, however, any ribonuclease contamination entering into these stages can cause disturbance in the selection process, which is difficult to monitor. The procedure utilises biotin-avidin interaction for capturing the cDNA clones. Although the materials required for biotinylation and in vitro synthesis of RNA are expensive, especially when considering the large amounts of biotinylated RNA probe required, the biotin-avidin system provides efficient capture of the tripartite hybrids. The use of streptavidin coated magnetic beads may prove to be more efficient by providing additional physical separation from non-selected products.

The selection is believed to be specific as it employs enzymatic methods for elution. RNase A will digest the single stranded RNA in the sandwich tripartite so that cDNAs hybridised to the genomic DNA will be specifically released. Non-specific clones are often retained on the avidin-matrix and should not be eluted under the low stringency conditions (high salt and low temperature). Swaroop et al. believed that the number of non-selected clones that could come through to the eluted material is minimal. Their assumption was tested by control experiments (described in section 5.1.3) where the genomic source is X-specific flow sorted DNA, to analyse the amount of NRL cDNA (which maps to chromosome 14) that comes through with any selected material. NRL comprised 26% of the starting material, and reduced to 4.6% in elution 1 and 2.3% in elution 2 of the selected clones. This clearly demonstrates that NRL is not being selected for, although the percentage that is coming through is accounted for by the high level in the starting material. However, they have not determined what the
other clones were, or whether they map back to the X chromosome. These could include cDNAs containing repeat, vector or a number of contaminants. Even though NRL is not being selected, it is not known what percentage of the other clones are background (i.e. non selected) products. This has not been tested further in the control experiments performed in this study due to a limitation in time. Future investigations should address this topic, as it would clearly demonstrate the specificity of this selection.

Even by comparing the limited selected clones analysed here, it was evident that there was contamination from clones that contained vector DNA only. This has been observed in other selection procedures, and is possibly due to problems in the construction of the cDNA library. It is necessary to suppress these and other contaminants before selection. In addition, when choosing clones from the selected library for further analysis, it would be better to screen Southern blots of PCR amplified cDNA inserts with vector DNA. This would immediately indicate clones that are possible contaminants.

The selection procedure requires pre-blocking the genomic source with Cot-1 DNA. This is highly enriched for intermediate repeats. It is has been used instead of total human genomic DNA as a blocking agent, because it should not drastically increase the overall sequence complexity of the mixture in the hybridisation by providing single copy templates that may compete with the “genomic RNA” for specific cDNAs. However, it does not quench low copy repeats, a common contaminant in selected libraries. The investigators have not recommended using any other material for pre-blocking. In recent experiments using YACs from chromosome Xp as the genomic source, Swaroop and co-workers only used Cot-1 DNA as a quenching reagent (personal communication). When purifying YAC DNA from PFGE there will always be some contamination from yeast chromosomes and previous selection experiments have repeatedly reported this problem (Lovett et al. 1994). The amount of ribosomal cDNAs comprised up to 70% in the selected library in one selection experiment (Lovett et al. 1991). Ribosomal cDNAs were selected due to hybridisation with ribosomal loci on yeast chromosomes. Sufficient suppression of repeat, yeast, vector and ribosomal DNA’s in a library has been one of the most noted refinements in previous selection procedures. Clone E1-10 was the only clone isolated
from Swaroop’s experiment using YACs, which apparently mapped back to the YAC clone (personal communication). Sequence analysis in this study showed that this may not represent a cDNA and may in fact be a genomic contaminant from the library.

The main problem encountered with cDNA selection techniques is the availability of a high quality and representative cDNA library. Even commercial libraries are not devoid of contaminating sequences. In addition, it is not known if a transcript encoded by the genomic DNA template is represented in the library that is used. In the Sandwich Selection experiment performed here, an adult retinal library was used, as the objective was to isolate candidate genes for RP9, a disease associated with retinal degeneration. However, the mutated gene may not be expressed in the neural retina, even though most predisposing genes for RP are expressed in this tissue (section 1.5.4). In further experiments it may be an improvement to include cDNAs from a pool of libraries from different tissues or developmental stages to increase the chances of selecting any genes from the cosmids. However, it is not known if the disease causing gene or any gene for that matter is coded by DNA in these cosmids, as the RP9 critical region is in a gene poor region. In addition, using cosmids singularly instead of pooled cosmids would be an advantage. Many more clones from the selected library using RP9 cosmids will have to be analysed to determine if they truly were selected. Unfortunately due to a limitation in time, this could not be completed in this study, but is being performed by colleagues working on the RP9 disease project.

Even though true success with this technique is yet to be demonstrated, this procedure has certain potential, and there is a lot of room for improvements. After testing this selection procedure here, certain changes to the protocol can be suggested. Firstly, the selected cDNAs are eluted and transformed straight into bacterial cells. Single stranded cDNA has a lower transformation efficiency than double stranded DNA. This may result in the loss of some selected material, and as the procedure does not amplify but simply selects, rare transcripts may be lost. It can be proposed that the single stranded templates be converted to double stranded templates prior to transformation, using DNA polymerase to synthesise its complementary strand after priming with a vector specific oligonucleotide.

Another modification of the selection technique involves using inter-Alu PCR products from the insert DNA of YACs as the genomic target sequence. This
Sandwich Selection Hybridisation
circumvents the problem of isolating YACs away from the host background and could
cut down the background level which affects selection, however, suppression of these
repeats is essential as primers specific to these repeats will partially amplify the repeat
sequence. Not all inter-\textit{Alu} products may amplify, especially if they are larger than
amenable for PCR, and some regions of the genome are deficient in these repeats
\cite{Moyzis1989} thus resulting in a non-representable genomic source.

PCR-based selection methods are well established, being employed in many
positional cloning efforts. They have also been tested and optimised to various extents,
and the limits and successes of the techniques have been well documented. The
Sandwich Hybridisation Selection technique is novel and no new genes have been
identified to date which map back to the genomic template either in this study or by
the developers. However the test experiments clearly demonstrate that it has the
potential to be a successful method, but it requires further examination and
optimisation which is now being performed by Yan and colleagues (Kellogg Eye
Centre, Ann Arbor, Michigan).
Chapter 6. General Discussion

6.1 Overview of work presented.

The purpose of this study was to explore physical mapping and gene identification procedures to identify candidate genes that lie within a critical region to which a disease had been linked. These were applied to the RP9 positional cloning project. At the onset of this work, the RP9 disease locus had been mapped to chromosome 7p14-15 in only one large family. The critical region had been refined to a 1.6-4 cM interval between the flanking markers D7S526 and D7S484. The absence of suitable candidate genes and disease-associated cytogenetic abnormalities necessitated the construction of a contig of overlapping YACs across the defined region.

The work in this study contributed to a high resolution YAC contig spanning approximately 5.9 Mb of chromosome 7p within which the adRP9 gene must lie (Keen et al. 1995a). Microsatellite markers were used to identify seed YACs in this region, which were subsequently placed in a contig of overlapping clones by the isolation of other YACs and the inclusion of additional STSs. The high clone redundancy in the complete contig has enabled the physical linkage and ordering of STSs, genes and ESTs to a density of ~1 every 175 kb. Seven gene fragments have been placed on the contig thus contributing to the transcriptional map of 7p. Only one was found to lie in the RP9 critical region, which requires further analysis to determine if it is a true transcript. The apparent gene cluster at the distal end suggests that the contig lies in the 7p14-7p15 boundary. 7p14 is a giesma positive dark staining band which is associated with as being gene poor. In contrast, 7p15 is a giesma negative light staining region and previous studies have shown that these tend to be gene rich. To further characterise the clones in this contig and to provide a more accurate estimation of its physical size and the distance between STSs, long-range restriction mapping should be applied, preferably utilising enzymes that identify CpG island sites. This physical and transcriptional map will aid the search for the underlying mutations causing RP9 and other diseases mapping to this area, by providing clones from the region that can be manipulated and which facilitate further transcript mapping. The next stage that may
follow initial physical mapping experiments is the construction of a contig of BACs, PACs or cosmids. This will enhance the map resolution and these clones are the preferred substrates for gene identification methods and sequencing procedures.

The search for positional candidates that lie in the RP9 critical interval was an ongoing process. At the start, the collection of genes that had been identified and localised sub-regionally to chromosome 7 was small. In an effort to increase the number of mapped transcripts to this chromosome, thirty ESTs and several chromosome 7 genes were mapped through a chromosome 7 somatic cell hybrid panel and placed in “binned” sub-chromosomal regions (Patel et al. 1995). Sequence similarity searches were performed against sequences in databases and revealed that five of these ESTs tagged known genes. One of these, EST06275, is the human homologue of its mouse counterpart. Most of these genes and ESTs were excluded from the RP9 disease region, but two ESTs were placed on the YAC contig which spans and extends from the RP9 critical region.

After exhausting all known genes and ESTs available at that time, a second approach to gene identification was employed, while also continually searching for positional candidates via genome databases. The sandwich selection technique is a novel cDNA hybrid selection procedure for identifying gene sequences within cloned DNA. This approach was developed to overcome some of the problems of established cDNA hybrid selection methods, and was tested here to determine its relative merits in comparison with the many other established gene isolation techniques. As the method was novel, at every step checks had been made to ensure no bias was occurring and also to assess that selection was set up correctly. Before testing the procedure fully, Swaroop and co-workers used this method in the hope of isolating cDNAs from YACs. Apart from clone E1-10, which is still yet to be proven as a true cDNA, no other clones were selected that mapped back to the target interval. The selection experiment performed here clearly demonstrates that this method is sensitive to both the abundance of cDNA in the starting library and the complexity of the genomic target, as is the case with other selection procedures. This strongly suggests improvements must be made before performing selection on larger genomic clones. The 3.01 fold quantity increase of blue cone pigment (BCP) control cDNA provides a detectable level of selection, but is lower than initial selection experiments performed
by the developers of the method. This can be accounted for by the lower abundance of target cDNA in the starting library and the use of cosmids rather than 5 kb plasmid clones. This procedure does not have the advantage of enrichment. PCR-based techniques provide high levels of enrichment due to amplification, but can also suffer from PCR bias, sequence alterations and non-specific amplification. The Sandwich Selection procedure was designed to select, not to enrich. Therefore in theory, specific clones will be favoured and collected together but not amplified. This poses problems when the level of target cDNA in the starting library is rare, since the quantity of the target cDNA in the selected material will then also be small and therefore more difficult to detect. In theory however, there should be very little non-specific products in the selected material, as the specificity of selection should reduce the number of these passing through. Nevertheless, in this and previous selection experiments (section 5.1.3), a large number of non-specific clones were detected. Further evaluation is necessarily required to demonstrate the specificity of the procedure and this was not carried out here due to the limitation of time. From the results of this selection procedure performed in this study, it can be concluded that it does possess some of the problems associated with other cDNA hybrid selection techniques (section 1.3.4.4) and requires modifications and further testing before its efficiency can be substantiated.

The most common problem encountered by groups employing gene searches using EST sequencing, cDNA selection and direct screening, is contamination and misrepresented clones from the cDNA library. Analysis of sequences deposited in dbEST has demonstrated that in normalised directionally cloned libraries ~5% of inserts are cloned in a reverse orientation and ~2.5% of clones are internally primed (Gerhold and Caskey 1996). This can cause confusion regarding identity of the EST. Reports from the large scale EST generation groups have also identified poly (A)$_n$ only inserts (Adams et al. 1991), chimeric clones (Adams et al. 1992b) and contamination of libraries with sequences from other organisms (Dean and Allikmets 1995). Also, from previous cDNA hybrid selection experiments, contamination of the cDNA library with non-transcribed DNA is a frequent problem. This is the main setback for these procedures and many investigators have stressed the importance of developing high quality cDNA libraries.
The Sandwich Selection procedure for gene identification has the potential to be applied as a cDNA hybrid selection technique, but needs improvement. But as positional cloning methods have progressed in the past years, perhaps one needs to consider if there really is a need for a new technique when so many have been already established? In future, concerning the RP9 positional cloning project, if a cDNA hybrid selection method were to be employed, the well established selection techniques would be favoured over the sandwich selection procedure, simply because these have already been tried and tested with reported successes. More importantly, all efforts will be concentrated on the identification of positional candidates in the critical region, especially from ESTs that are rapidly accumulating in public databases. Recently, several ESTs have been found to map in the RP9 region. Analysis of these will be the initial priority and only if these are excluded as candidate genes for RP9 and no other positional candidates are identified, will sequencing, exon trapping, direct screening or cDNA selection be considered.

Meanwhile, the RP9 positional cloning project will be greatly aided by the large scale sequencing project being undertaken by a collaboration between the National Human Genome Research Institute (NHGRI) at Washington University and Washington University Genome Sequencing centres. Together these groups are generating mass sequence information from chosen clones mapped to high resolution, which are being assembled into overlapping sequence tiling paths. A shotgun sequencing strategy project has commenced by targeting clones from the 7p14 and 7q31.3 regions. The sequencing substrates are Washington University hybrid cell line YAC clones (Green et al. 1995), chosen because of their low level of reported chimerism (15%), which were subcloned into cosmids. 1997 has seen significant advances in the analysis of human chromosome 7. 2,006 chromosome 7 ESTs have been generated from cDNA libraries prepared by a cDNA hybrid selection procedure (Parimoo et al. 1991) using binned chromosome 7 cosmids as genomic targets. A fraction of these have been mapped sub-regionally (Touchman et al. 1997). Also a large number of STSs have been created and localised, producing a high density STS based radiation hybrid and YAC physical map covering 97% of the chromosome (Bouffard et al. 1997a and b). All of this will undoubtedly contribute significantly to chromosome 7 positional cloning projects.
6.2 The future for positional cloning.

For the majority of monogenic disorders where the molecular defect has not yet been discovered, the aetiology of the disease is not known. Positional cloning has become invaluable in these instances to identify the causative gene. Genetic analysis has been and remains a powerful means of mapping loci and disease traits. With the current status of the genetic map, a disease region can potentially be delimited to a 1-2 cM interval, depending on the pedigree size and structure. Isolation of novel polymorphic markers in key intervals can make it possible to refine the critical region even further. More importantly this task can be performed in modest sized labs within a relatively quick time, as long as there are sufficient numbers of informative meiotic events in families that exhibit a classical Mendelian inheritance pattern.

Positional cloning projects have demonstrated the successes of most gene identification and isolation procedures, and most groups employ more than one strategy. Both direct screening and exon amplification were used to successfully identify the gene that is implicated in the pathogenesis of Treacher Collins syndrome (The Treacher Collins Syndrome Collaborative Group, 1996). In the search for the BRCA1 disease gene one group employed three procedures (genomic sequencing, direct screening and cDNA hybrid selection) using BACs, PACs and P1 clones. cDNA hybrid selection was considered to be most successful and was recommended for use in positional cloning by the authors, as it identified most of the total transcripts in the critical region. But direct screening was only performed using probes from a limited portion of the BRCA1 critical region, and these identified all the cDNA clones in this area that had been isolated by cDNA hybrid selection. Genomic sequencing and coding prediction detection methods identified a proportion of these transcripts but proved most valuable in building full length gene sequences. The BRCA1 transcript was identified by all three procedures (Harshman et al. 1995).

Even with the reported successes of cDNA hybrid selection, the efficiency is hampered by the amount of work required to sift through the selected material to identify true selected clones. Previous cases have demonstrated that up to 30% of the selected material comprises of non-selected background single copy clones (Lovett et al. 1991). Post selection screening of the library will not identify these. In addition, many contaminants in the selected material include genomic, vector and repeat
sequences. In exon trapping projects, the recovery of products derived from cryptic splicing and vector inserts is a frequent problem, even though vectors have been designed to overcome these occurrences. Direct screening suffers from a large number of false positives and a very poor signal to noise ratio. Thus all these strategies have their limitations and advantages, but the choice of gene identification method employed in a laboratory will inevitably depend on the resources and personal preference.

In any event, these procedures may be used less in the future. Francis Collins predicted that positional candidates will take over from these "traditional" gene isolation techniques (Collins 1995). With the increasing output of ESTs contributing to the transcript map of the human genome, this prediction is being borne out. In September 1994 only 38% of the positionally cloned genes mutated in human disease were represented by ESTs in dbEST, whereas in June 1996 this had increased to 82% (NCBI). This demonstrates the power of using ESTs as positional candidates and supports the notion that disease searching by database scanning will overtake and eliminate traditional rate limiting gene searches. A large number of ESTs are now being mapped and the next stage may be to perform expression studies. With further characterisation of ESTs, candidates can be preferentially favoured for analysis by examining their expression patterns. However, this information is beneficial yet not always required. As the aetiology of many diseases is not known, it is crucial that no gene should be overlooked.

Currently, ~46,000 sequences are represented in the Unigene database, which may account for 50% of the human genes. Those that are undiscovered so far, are likely to be rare tissue specific transcripts. The completion of a transcript map still requires a lot of work, but the large number of ESTs available has placed less focus on the more traditional gene identification approaches. Even though single pass sequencing of clones from different libraries is likely to find many of the undiscovered genes, some will nevertheless remain undetected until the human genome is completely sequenced.

With the focus of the major laboratories now turning to large scale sequencing, improved exon prediction programs will provide rapid means of identifying a large proportion of genes. Gene hunting in this way may be favoured over the other methods as it is independent of tissue and developmental time expression. Also the resources
and time needed for employing the traditional methods do not yield a quick result, which is further hampered by the large measure of non-specific products. Rapid returns may be provided from efforts put into analysing DNA sequences in critical regions, especially as sequencing becomes faster and cheaper. Sequencing also reaches straight to the source. The power of this approach is becoming increasingly evident. While searching for the disease causing gene for RP3, candidate genes in the critical region were isolated by direct screening and exon trapping but these were excluded by mutation screening. Disease causing mutations were finally discovered in a novel transcript (RPGR) which had been identified by analysing sequences from cosmids in the critical region (Meindl et al. 1996). Also, as mentioned previously, when positionally cloning the Pendred syndrome disease gene, ESTs and known genes in the region were primarily chosen for mutation screening. These were subsequently excluded and mutations were then discovered in a novel transcript pendrin, identified by analysis of sequences from BAC clones in the critical region (Everett et al. 1997). In addition, the causative gene for autosomal dominant non-syndromic deafness (DFNA1; Lynch et al. 1997) was discovered by sequencing BACs from the critical interval on chromosome 5q31. Comparisons of the sequences with ESTs identified a human homologue of the Drosophila gene diaphanous and pathological mutations were later confirmed.

With this approach, the rate limiting step in positional cloning will then be the characterisation of full length gene structure and mutation analysis. This can be accelerated by forming contigs of overlapping ESTs, if they are represented in EST databases, and comparing the cDNA and genomic sequence. Sequencing of ESTs and full length cDNAs also complements genomic sequencing by identifying splice variants, expression patterns and gene structure. In addition, identification of promoter sequences and regulatory elements is necessary. There are many well established procedures for mutation analysis developed for rapid and large scale use, but all have varying efficiencies. It has been suggested that the most efficient method for mutation screening for small samples is simply to perform sequence comparisons between affected and unaffected patient DNA. This defines the nature of the change precisely and it is the necessary final step of mutation detection.

Positional cloning has been immensely accelerated by the Human Genome
Project initiative. The availability of the high resolution genetic, physical and transcript maps can now be applied to the more perplexing genetic frontier, the dissection of complex polygenic diseases and quantitative traits.

6.3 The future in understanding the disease mechanisms involved in retinal degenerations.

At least 68 genes causing retinal degenerations, syndromically and non-syndromically, have been mapped to date to specific chromosomal sites, mostly by linkage analysis. Only 32 have been cloned and characterised (RetNet: http://utsph.sph.uth.tmc.edu/www/utsph/RetNet/disease.htm). When disease causing mutations have been identified it is then important to determine how the mutation affects protein function and ultimately the phenotype. It is only by understanding the precise effects of a mutation in the aetiology of disease that complete and effective treatments can be developed.

In retinitis pigmentosa, allelic and non-allelic heterogeneity, as well as phenotypic variability between and within families, makes genotype-phenotype correlation difficult. In addition, different diseases are caused by mutations in the same gene. For example, peripherin/RDS mutations can cause both RP and macular degeneration (Wells et al 1993). Different rhodopsin mutations can cause autosomal dominant RP, recessive RP and congenital stationary night blindness (Gal et al. 1997). Studies of the effects of certain mutations in rhodopsin have shown that some mutants produce very abnormal molecules that are non-functional and fail to be transported from the endoplasmic reticulum to the plasma membrane. However, other mutations lead to rhodopsin molecules with few detectable differences. It is not known why different mutations in the same gene are highly variable in their phenotypic expression, but several studies have tried to define a correlation between mutations and their biological and clinical implications (e.g. Daiger et al. 1995; Keen et al. 1996).

Genes so far known to be implicated in RP, are largely those expressed in the RPE or specifically in the photoreceptors. These include genes encoding proteins involved in the phototransduction pathway, structural proteins and RPE metabolic proteins. How do these different biochemical defects result in the same clinical
phenotype of retinal degeneration? In RP and some macular diseases, degeneration of rod photoreceptor cells leads to degeneration of cone cells and vice versa. Why do cones die as a result of genetic abnormalities in rod proteins? Rods and cones are closely interdigitated in a dense cellular matrix and share many metabolites and cofactors, but each have their own specific proteins involved in the phototransduction cascade. Although the molecular pathology and associated biochemical abnormalities are diverse, a final common metabolic pathway may be involved. Several hypotheses have been proposed to explain how retinal degeneration occurs, but the most favoured view in recent years is that photoreceptor cell death is the result of apoptosis, a form of programmed cell death.

Unlike simple necrotic processes, where chemical/structural alterations lead to the arrest of metabolic activity, then cell lysis and phagocytic activity, apoptosis is a highly regulated cellular activity and acts naturally as a normal process in embryonic development. It differs from necrosis as it affects isolated cells rather than patches of tissue, is not usually accompanied by inflammation or scarring, and apoptotic cells are usually phagocytosed by adjacent cells with minimal effects to the microenvironment. A hallmark of apoptosis is the initial signs of chromosomal condensation and nuclear fragmentation, which is detectable by the appearance of a characteristic DNA ladder on agarose gel electrophoresis.

These hallmarks of apoptosis have been observed in a number of animal models of retinal dystrophy. This includes the rds mouse, caused by a peripherin mutation, the rd mouse, caused by a defect in cGMP phosphodiesterase and transgenic mice with different rhodopsin mutations (Portera-Cailliau et al. 1994; Chang et al. 1993). Even though these animal models have different mutations, their subsequent cell death was remarkably similar. During development, as many as 50% of neurons die from apoptosis by failure to interact with their target cells. In mouse models, the appearance of apoptotic cells in affected retinas began and extended beyond the stage at which developmental neuronal cell death reaches a peak in normal animals and was restricted to the photoreceptor layer. This suggests that apoptosis may be used as a response to aberrant stimuli as well as in development (Alder et al. 1996).

It is not yet known how each mutation triggers photoreceptor apoptosis but recent studies have shed light on genes whose products are involved in commencing or
blocking apoptosis. For example, the interleukin-1β-converting enzyme is essential for apoptotic cell death to occur. In contrast, over-expression of the \textit{bcl-2} gene in transgenic mice carrying mutations in three different genes that produce an RP phenotype, results in increased photoreceptor survival (Chen \textit{et al.} 1996). This gene is believed to negatively regulate apoptosis. Also, intraocular injections of the fibroblast growth factor in the RCS rat, where the primary genetic defect is in the RPE, has been proved to reduce the progression of photoreceptor degeneration (reviewed in Alder \textit{et al.} 1996). More insight into the mechanisms and factors involved in this process is required, but as it has been associated with retinal degenerations caused by diverse genetic mutations, the intervention in this process is being investigated as a single form of therapy for all RP cases (Gregory and Bird 1995; Travis 1998).

When studying single mutations, the creation of a transgenic animal model by gene targeting which harbours the specific mutation, should prove most informative for studying its effects histopathologically. By gene replacement or gene knockout, the effect on the organism of absent gene function or the presence of a dysfunctional product can be evaluated (Capecchi 1994). Although the physiological abnormalities in the mouse are frequently not identical to those in human patients, they are extremely valuable for testing gene therapeutic strategies. These models would also allow disease progression to be followed from development to advanced stages in more detail than has been possible in humans. These may also provide information on possible environmental factors that may have influenced the phenotypic variability between and within families. Considerable effort is still required to identify the entire set of genes that are implicated in RP, but once discovered they will provide an important insight into the complex yet fascinating process of vision. The development of diagnostic and therapeutic tools for tackling RP will be the next major challenge.
Publication list


References


Baldwin C. T., Weiss S., Farrer L., De Stefano A., Adair R., Franklyn B., Kidd K. K.,
deafness (DFNB4) to chromosome 7q31 and evidence for genetic heterogeneity in the

Banfi S., Borsani G., Rossi E., Bernard L., Guffanti A., Rubboli F., Marchitiello A.,
and mapping of human cDNAs homologous to Drosophila mutant genes through EST

pigmentosa locus on 17q (RP17): fine localisation to 17q22 and exclusion of the

1895-1902.

Bates G. P., McDonald M. E., Baxendale S., Youngman S., Lin C., Whaley W. L.,

Bellanne-Chantelot C., Lacroix B., Ougen P., Billault P., Beaufils S., Bertrand S.,
Georges I., Glibert F., Gros I., Lucotte G., Susini L., Codani J. J., Gesnouin P., Pook
S., Vaysseix G., Lu-Kuo J., Ried T., Ward D., Chumakov I., Paslier D. E., Barillot E.


Genet.* 3: 342-347.
References


References


References


References


References


Regional Assignment of 30 Expressed Sequence Tags on Human Chromosome 7 Using a Somatic Cell Hybrid Panel

RESHMA J. PATEL,* T. JEFFREY KEEN,* KARL-HEINZ GRZESCHIK, † WILLIAM C. NIERMAN, ‡ PHILLIP HAYES, § SHOMI S. BHATTACHARYA,* AND CHRIS F. INGLEHEARNE*1

*Department of Molecular Genetics, Institute of Ophthalmology, London, EC1V 9EL, United Kingdom; †Institut für Humangenetik der Universität, Bahnhofstraße 7, D-35037 Marburg, Germany; ‡Department of Molecular Biology, American Type Culture Collection, Rockville, Maryland; and §HGMP Resource Centre, Hinxton Hall, Hinxton, Cambridgeshire, United Kingdom

Received June 2, 1995; accepted August 21, 1995

The regional assignments of 30 expressed sequence tags (ESTs) on human chromosome 7 were determined by studying the segregation of their PCR-amplified products in a panel of mouse somatic cell hybrids. ESTs are important molecular landmarks for physical mapping and can be considered as tags to candidate genes for genetically linked human inherited diseases. These results contribute further potential gene sequences to the transcriptional map of chromosome 7.

As a first step toward the identification of all human genes, several laboratories have adopted a strategy of randomly selecting clones from cDNA libraries and analyzing them by single-pass sequencing to produce expressed sequence tags (ESTs) (1, 2, 7). In many cases PCR primers have been selected from the sequences generated to produce sequence tagged sites (STSs). This approach allows researchers to test for the presence of an EST in genomic or cloned DNA. ESTs can then be assigned to a chromosome or subchromosomal region using somatic cell hybrid panels (9), YAC localization (3), or FISH analysis of the original cDNA clone (8). Researchers interested in particular subchromosomal regions have to date mapped ESTs on chromosomes X, 6, and 18 (6, 10, 12). Once localized they provide landmarks for physical and expression mapping and could be new candidate genes for genetic diseases. We adopted the hybrid panel strategy in this study to determine the regional assignment of 30 ESTs on chromosome 7.

Of the 30 ESTs that have been localized in this study, 25 are brain ESTs first described by Adams et al. (1, 2). Eight ESTs map to the p arm, 1 to the pericentric region, and 21 to the q arm. These ESTs can now be considered positional candidates for chromosome 7 genetic diseases and traits in which the causative gene has not been found. There are 16 such defects known at this time, including cystoid macular dystrophy (CYMD), retinitis pigmentosa (RP9 and RP10), congenital chloride diarrhea (CCD), craniosynostosis (CRS), and osteogenesis imperfecta 4 (O14) (11). ESTs from human brain cDNAs could be candidates for genes associated with neurological diseases. On this basis it is possible to speculate that ESTs 02120, 00601, and

been published elsewhere; 7 have been localized to chromosome 7 by Maglott and colleagues (9), and their primer sequences are given in Table 1. The remaining 5 ESTs, 3 brain and 2 liver, have been mapped to chromosome 7 using a monochromosomal hybrid panel and are described here for the first time. Primer sequences for these are also included in Table 1. Regional assignments for all ESTs were obtained by PCR analysis on a well-established panel of human–mouse hybrid cell lines (16). The panel contains 13 cell lines that subdivide chromosome 7 into 13 regions. Total mouse and human genomic DNA were also included for assay control.

The EST content of individual hybrid cell DNAs was tested by PCR amplification, followed by analysis on agarose gels. The regional assignments of the ESTs were deduced from the patterns of amplification, and these are given in Fig. 1. In addition, to check the content of chromosome 7 in the different cell lines, several markers and genes with known locations were tested in the same way. In each case the pattern of amplification was consistent with published data. Grzeschik and colleagues (15) describe the localization of a large number of chromosome 7 genetic markers in the same hybrid panel, and by comparison with these data it is also possible to deduce approximate genetic locations for the ESTs.

Approximately 100 genes have been localized to chromosome 7 to date, as well as two large gene clusters at the HOXA and TCRG loci (11). This study localizes a further 30 expressed sequences on chromosome 7. Eight ESTs map to the p arm, 1 to the pericentric region, and 21 to the q arm. These ESTs can now be considered positional candidates for chromosome 7 genetic diseases and traits in which the causative gene has not been found. There are 16 such defects known at this time, including cystoid macular dystrophy (CYMD), retinitis pigmentosa (RP9 and RP10), congenital chloride diarrhea (CCD), craniosynostosis (CRS), and osteogenesis imperfecta 4 (O14) (11). ESTs from human brain cDNAs could be candidates for genes associated with neurological diseases. On this basis it is possible to speculate that ESTs 02120, 00601, and
FIG. 1. A schematic representation of the hybrid cell lines used in this study, with EST localizations shown to the left. The vertical bars represent the human content of each cell line. ESTs are identified by their EST numbers or laboratory names and by their locus symbols (italic) where these exist. Primer sets for several chromosome 7 genes and microsatellite markers with known locations were also tested in this study to confirm the human DNA content of the cell lines. The localizations found, which are consistent with published data, are included in this figure. The genes tested were amphiphysin (AMPH), the AH receptor gene (AHR), acylcoyxy acylhydrolase (AOAH), the human growth hormone-releasing hormone receptor gene (GHHR), and adenylyl cyclase type 1 (ACD1). The markers tested in this study are those marked with an asterisk to the right of the figure. Other markers shown to the right define the upper and lower limits of the intervals (15).

06146 might be considered candidates for the RP9 and CYMD loci, while EST00167 may be a candidate for the RP10 locus. Our own work on physical and genetic mapping around the RP9 locus has established that EST02120 is in the RP9 interval, while EST00601 lies immediately distal to it and is therefore a positional candidate for CYMD (unpublished data).

Three of the ESTs regionally assigned in Fig. 1 have previously been reported to have substantial sequence identity with nucleotide or amino acid sequences in database records (1, 2). We have reanalyzed these in current DNA and protein sequence databases. EST01644 is 99% identical to a 382-bp stretch of the cAMP-dependent PK regulatory subunit R1-β gene (PRKAR1B). The localization of this EST in this study is consistent with the previous localization of PRKAR1B by genetic analysis to 7pter (14). EST06146 matches the human amphiphysin gene with 99% identity over 334 bp. The localization of this EST is consistent with the known location of the amphiphysin gene (17). EST06275 was found to have 100% identity in a 28-amino-acid region with the rat metabotropic glutamate receptor 3 gene. The human homolog of this gene had not previously been localized so the placing of this EST at 7q11–q22 represents a tentative new gene assignment.

The remaining ESTs were also analyzed by BLAST searches of DNA and protein databases. No substantial new gene homologies were revealed. Two ESTs did show homology to the Alu repeat. EST01995 contains part of an Alu repeat and has been assigned to loci on chromosomes 7, 9, 12, and 20. The multiple sites for this EST may reflect a repeated component in the primers, or the cDNA may be a member of a gene family.
TABLE 1
Primers and Product Sizes for ESTs Not Previously Reported in the Literature

<table>
<thead>
<tr>
<th>EST identifier</th>
<th>Primers</th>
<th>Product size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EST01895</td>
<td>AGTATCCCCTTCCAGTTCC</td>
<td>124</td>
</tr>
<tr>
<td>EST02120</td>
<td>GCTAAGAGCCAGAGAGGC</td>
<td>251</td>
</tr>
<tr>
<td>EST02379</td>
<td>AGTCCTAAGAACACGCTGCA</td>
<td>89</td>
</tr>
<tr>
<td>EST06146</td>
<td>AGACCTTGCACCTCAAAAG</td>
<td>88</td>
</tr>
<tr>
<td>EST06238</td>
<td>CGGATAAACGCTTGGT</td>
<td>88</td>
</tr>
<tr>
<td>EST06275</td>
<td>TAATTGGGCACTAGAAGCTACG</td>
<td>114</td>
</tr>
<tr>
<td>AABQ2RZ</td>
<td>AACATCTACTTGGGGATGTAGCCTCC</td>
<td>124</td>
</tr>
<tr>
<td>AAAAEP</td>
<td>GCTACTGAAAACATGCTGCA</td>
<td>101</td>
</tr>
<tr>
<td>AADDEER</td>
<td>TTAGGTTTGCAAGGAATCTCAG</td>
<td>96</td>
</tr>
<tr>
<td>AAABFBWR</td>
<td>TTTGTAATCTCAGGAGCAGCTCACA</td>
<td>124</td>
</tr>
<tr>
<td>AAFENH</td>
<td>TGGTTGGAGACTGCTAATCTGGA</td>
<td>134</td>
</tr>
</tbody>
</table>

Note. ESTs AABFBWR and AAFENH derive from a liver cDNA library, while the rest are from brain cDNAs.

EST01888 is almost entirely composed of Alu sequences. Nevertheless, the primers used define a unique STS that is present in a brain cDNA library.

In summary, we have sublocalized 30 ESTs on chromosome 7, including three with substantial sequence identity to known genes. As well as providing potential candidates for chromosome 7 genetic diseases, these ESTs are useful tags for STS content mapping and contribute new sequences to the expression map of chromosome 7.

ACKNOWLEDGMENTS

R.J.P. is supported by an MRC/HGMP studentship. We gratefully acknowledge the Welcome Trust (Grant 035535/Z/92/Z) for funding this research. We also thank the HGMP Resource Centre for on-line bio-informatics facilities.

REFERENCES

A YAC Contig Spanning the Dominant Retinitis Pigmentosa Locus (RP9) on Chromosome 7p

T. Jeffrey Keen,* Chris F. Inglehearn,=* Eric D. Green,† Aimee F. Cunningham,† Reshma J. Patel,* Rachel E. Peacock,* Steve Gerken,‡ Ray White,‡ Jean Weissenbach,§ and Shomi S. Bhattacharya*

*Department of Molecular Genetics, Institute of Ophthalmology, Bath Street, London, EC1V 9EL, United Kingdom; †National Center for Human Genome Research, National Institutes of Health, 9 Memorial Drive, Bethesda, Maryland 20892; ‡Department of Human Genetics, University of Utah, Salt Lake City, Utah 84112; and §CNRs URA 1922, Génethon, Evry, France

Received April 12, 1995; accepted June 5, 1995

The dominant retinitis pigmentosa locus RP9 has previously been localized to 7p13–p15, in the interval D7S526–D7S484. We now report refinement of the locus to the interval D7S575–D7S484 and a YAC contig of approximately 4.8 Mb spanning this region and extending both distally and proximally from it. The contig was constructed by STS content mapping and physically orders 29 STSs in 28 YAC clones. The order of polymorphic markers in the contig is consistent with a genetic map that has been assembled using haplotype data from the CEPH pedigrees. This contig will provide a primary resource for the construction of a transcriptional map of this region and for the identification of the defective gene causing this form of adRP.

MATERIALS AND METHODS

Pedigree analysis in the RP9 family. The RP9 pedigree used in this study and the phenotype associated with it have been described elsewhere (Jay et al., 1992; Moore et al., 1993). Microsatellites were typed by PCR amplification (30 cycles of 94, 55, and 72°C, with 30 s at each step) with incorporation of [α-^32P]dCTP during the reaction. Alleles were resolved by size fractionation on a 6% polyacrylamide denaturing gel and then visualized by autoradiography.

Genetic map construction in the CEPH pedigrees. A two-step algorithm was used for construction of a genetic map flanking the RP9 region of chromosome 7 (Gerken et al., 1995). First, the method of location scores (Lathrop et al., 1985) was used to identify published PCR-based genetic markers that showed high likelihoods for localization to the D7S526–D7S484 interval. Genotypic data for this study were obtained from the CHLC database, version 2.5, or from the Utah database (Utah Group, 1995). Next, the order of markers was determined by sorting segregation patterns and by minimizing double-recombination events using a meiotic breakpoint mapping panel for the D7S526–D7S484 interval and the computer programs RBUild and CSORT (Eisen et al., 1996). Genotypic data for this analysis were retrieved from the CEPH (version 7) and CHLC (version 2.5) databases. Any multiple recombinant chromosomes whose crossovers

INTRODUCTION

Retinitis pigmentosa (RP) is a heterogeneous inherited disorder of vision, characterized by night blindness, visual field loss, and pigmentation in the peripheral retina (Bird, 1988). To date there are seven published loci for the dominant form of RP. These are the rhodopsin and RDS-peripherin genes and five other loci defined by linkage analysis on chromosomes 7p, 7q, 8q, 17p, and 19q (Rosenfeld et al., 1994; Greenberg et al., 1994). The locus on chromosome 7p, which has been assigned the locus symbol RP9, was first reported by this group in a single adRP pedigree and has been refined to an interval of approximately 4 cM between markers D7S526 (distal) and D7S484 (proximal) (Inglehearn et al., 1994).

Since no obvious candidate RP genes have been mapped in the RP9 interval, we proceeded with further genetic and physical analysis to identify the causative gene. Additional markers were obtained from all of the available high-resolution genetic maps (Gyapay et al., 1994; Cooperative Human Linkage Center, 1994; Matise et al., 1994; Buetow et al., 1994; Utah Marker Development Group, 1995; Tsui and Grzeschik, 1995). YACs used in this research were either derived from a chromosome 7 YAC resource (Green et al., 1994, 1995), isolated from the ICI YAC library (Anand et al., 1990), or were CEPH mega-YACs placed in the region on the CEPH/Génethon physical map (Cohen et al., 1993). By meiotic breakpoint analysis with the CEPH genotypic data, we have developed a high-resolution genetic map with 22 loci flanking the RP9 locus. Haplotype analysis in affected RP pedigrees with the new microsatellites placed RP9 on all of the published genetic maps and further refined the locus to the interval D7S795 (distal) to D7S484 (proximal). Subsequently, using microsatellite markers, 11 new STSs, and an EST, we have assembled an STS content contig of 28 YACs spanning the RP9 region and extending over approximately 4.8 Mb of chromosome 7p.

1To whom correspondence should be addressed. Telephone: 071 608 6920. Fax: 071 608 6863.
could not be verified by data from independent laboratories were
excluded from the mapping panel. Finally, the support for the order
of markers on the genetic map was tested using LINKAGE (version
5.1) by determining the odds against inversion of adjacent loci. Geno-
typic data from the full CEPH families were used for this analysis
and to estimate the recombination fraction. Data for the most infor-
mative locus was used when two or more loci were clustered (not
separated by a recombination event).

**YAC DNA isolation and analysis.** For analysis of STS content,
YAC DNA was extracted in solution, while intact yeast chromosomes
were prepared in agarose plugs for sizing of YAC inserts on pulsed-
field gels. YACs were sized on a 1% agarose gel run on a Bio-Rad
CHEF pulsed-field apparatus with 16- to 22-h run times, 5 V/cm,
and a pulse time ramped from 60 to 90 s (Birren and Lai, 1993).
YAC sizes were estimated by comparison of their mobility relative
to yeast chromosome size standards. STS content of YACs was deter-
mined by PCR amplification and resolution on 2% Nusieve agarose
gels, with detection by ethidium bromide staining.

**Characterization of YAC insert terminal STSs and Alu PCR prod-
uct STSs.** The conventional method for isolating YAC insert termi-
nal sequences involves the creation of a library of YAC restriction
fragments to which "vectorette" units have been ligated, followed by
PCR amplification using a PTAC vector arm primer against the vec-
torette promoter (Riley et al., 1990). We have developed a simpler
method that relies on the linear amplification of DNA adjacent to
the vector arms with a vector arm primer, and the ability of the same
primer to anneal nonspecifically (at 40°C) in the linearly amplified
single-stranded product. Five hundred nanograms of YAC DNA in
solution was subjected to linear amplification in the presence of only
a single vector primer, either for the right arm (ata tag gcc cca gcc
acc cca ctc ggt gc) or the left arm (acc ccc ttc tcg gac cac tgt ccc acc
gc). This amplification was achieved by 100 cycles of 10 s at 95°C,
15 s at 65°C, and 30 s at 72°C in a Perkin–Elmer 9600 thermal
cycler. At this point fresh Top polymerase was added to the reaction,
and a further 5 cycles of 10 s at 95°C, 15 s at 40°C, and 45 s at 72°C
was carried out. The initial cycling profile was then resumed for a
further 35 cycles. The resulting products were sequenced using [γ-
32P]ATP end-labeled primers internal to those used in the amplifica-
tion (for the left arm, gtt gtt tta agg cgc sag, and for the right arm,
gtc gaa cgc cgc atc gta ag). Direct sequencing was performed with a
Pharmacia T7 sequencing kit using the manufacturer's protocol.
Primers were then selected from the resultant sequence to produce
STs with products as large as possible.

Three additional STs were isolated by Alu PCR of YACs in the
contig. Fragments amplified with primers Ale 1 (gcc tct cca aag
gcc att aca g) or Ale 3 (ccg tac ccc acc cgc cgg) were cloned and
then sequenced.

**RESULTS**

To facilitate the localization of the RP9 locus and to
initiate the development of a YAC contig, a genetic map
was constructed using genotypic data available for the
CEPH families. The genetic map consisted of 22 loci,
of which 18 were linearly ordered into 12 clusters
spaced an average of 3 cM between clusters (Fig. 1).
Two clusters, which contained 3 and 5 loci, showed no
recombination between markers in the clusters. The
locations of four loci that could not be uniquely ordered
are shown. Support for the order of loci was significant,
with odds greater than 200:1 against inversion of adja-
cent loci. Using the new microsatellites and this
comprehensive map in the RP9 family, we have obtained
a further refinement of the locus as follows: telomere—
D7S526-D7S795-RP9-D7S484-centromere. This is
illustrated by the haplotype analysis in Fig. 2. The four
polymorphic microsatellites within the RP9 interval,
D7S683, D7S656, D7S460, and D7S497, are uninforma-
tive for both the proximal and the distal crossovers.

To create a YAC contig across this interval, 21 YACs
from the chromosome 7 YAC resource (Green et al.,
1994, 1995) were selected on the basis that they con-
tained D7S690, D7S683, D7S656, D7S460, D7S497,
or D7S484. These YAC clones formed four apparently un-
connected contigs around D7S484, D7S497, D7S460/
D7S656/D7S683, and D7S690. To join and extend these
contigs, seven YAC insert terminal sequences were de-
termined and STs generated from them (Table 1).
Three additional STs were generated from Alu PCR
products of selected YACs. One of these STs (MS0003)
contained a (CA)n tract that was shown to be polymor-
phic but was uninformative in the RP9 family. From
the initial STS content of these YAC clones a single
contig was assembled. To increase clone redundancy at

![FIG. 1. High-resolution genetic map for region around chromosome
7p13–p15. The map is scaled in centimorgans according to the
Haldane mapping function. Loci that show no recombination with
others are bracketed. The odds against the inversion of adjacent
loci were determined from the most informative locus for each cluster.
The locations of markers that did not have unique orders are shown
to the side of the map. While errors in the genotypic data could lead
to an incorrect order, only a single possible solution was obtained,
indicating that the data did not contain any errors that resulted
in a close double-recombination event. Also, although only single
recombination events separated some of the clusters, the odds were
>1000:1 against inversion of adjacent loci in the region of the RP9
locus. Haplotypes in the linked family resolved a series of blocks of
linked markers and confirmed the order shown in this analysis. In
one case the adRP family data improve on the CEPH information
shown by placing D7S252 distal to both D7S817 and D7S795 (data
not shown). Markers contained within the YAC contig are indicated
by asterisks.](image-url)
YAC contig spanning the RP9 locus

Note. The letters LA and RA refer to YAC terminal STSs and denote left and right arms of the YAC insert. The number before these letters identifies the parent YAC clone. The STSs sWSS3586, MS0002, and MS0003 are derived from cloned Alu PCR products of YACs from the contig. AFM067x9, AFM190yh2p, and AFMb073zc9 are Généthon microsatellites that failed polymorphic criteria.
FIG. 3. A YAC contig encompassing the RP9 region and adjacent markers. The heavy horizontal line represents the chromosome, with the RP9 locus interval denoted by the hatched area below this line. Above it are STSs in the order in which they occur in the contig. In all but five cases the physical order could be determined uniquely. Where two STSs proved inseparable with the data available, they are assigned to a single position. The order of D7S795 and D7S690 is based on the STS content of YACs not included in this study (E.D.G., unpublished data). The lighter horizontal lines represent YACs, with sizes given where known (kb). A filled rectangle on the end of a YAC indicates a terminal sequence STS. In the case of yWSS2609 an end is not indicated because it was not possible to determine which end had been cloned. Vertical lines show the position of STSs in YAC clones. Since this contig is based purely on STS content it is not drawn to scale, and intervals may vary widely in size. For instance, the STSs D7S460 and 1211RA have now been identified in a single cosmid, so the interval between them is less than 40 kb. YAC clones with the prefix yWSS are derived from the chromosome 7 YAC resource (Green et al., 1984, 1995). This consists of YACs from several genomic libraries together with clones from a chromosome 7 hybrid cell line library. Those from the hybrid cell line library are identified principally by their yWSS number. The other YACs from the chromosome 7 YAC resource also have plate numbers from their library of origin. YACs yWSS298 and 443 correspond to Washington University clones D89G7 and D56C8. YACs yWSS2597, 3011, 3043, 3255, 2535, 2609, 3029, 3215, and 3263 correspond to CEPH clones 757_h_11, 808_c_11, 812_f_05, 860_f_06, 739_b_12, 762_g_09, 810_f_08, 849_a_02, and 867_a_06, respectively. However, as in some cases these clones have come from mixed wells and are now unique clones, their yWSS identity is preferred. YACs 16BC5, 3EG9, 26FH4, and 30GD3 were isolated from the ICI YAC library. The remaining three YAC clones came directly from the CEPH YAC library, with plate numbers shown in italics.

**DISCUSSION**

This manuscript describes the completion of a YAC contig encompassing the dominant retinitis pigmentosa locus RP9 on chromosome 7p13–p15. The RP9 interval is estimated to be approximately 2.75 Mb, while the contig extends both distal and proximal to it and is estimated to be 4.8 Mb in length. The contig includes 28 YACs from three different sources, ordered according to STS content using 29 STSs mapping to the region. The inclusion of the genetic markers locates the RP9 locus on all of the major genetic maps, while YAC terminal STSs provide useful reference points to determine which of a set of overlapping YACs extends the furthest.

The construction of this contig was facilitated by the high density of PCR-based genetic markers now available. These together with the 11 new STSs allowed assembly of the contig without the need for extensive YAC walking or restriction mapping to establish overlaps. The use of marker genotyped in the CEPH families in a meiotic breakpoint ordering strategy has provided additional evidence to confirm the order of the STSs. Furthermore, this order is consistent with that obtained from high-resolution mapping in a radiation hybrid panel (E. D. Green, in preparation).
The lack of a long-range genomic restriction map means that chimerism cannot be excluded in the YACs used. However, there are several large YACs in the contig that contain 5 or more contiguous STSs and whose sizes appear to be internally consistent with each other, making chimerism in these YACs unlikely. In particular, yWSS3029 contains over 1 Mb of DNA from the RP9 interval, including five microsatellites that show no recombination with the disease phenotype. In addition, STSs exist for both ends of this YAC, making it highly unlikely to be chimeric and therefore a good starting point in the search for candidate genes for RP.

The presence in the RP9 region of a segment of duplicated sequence presented substantial impediments to the correct assembly of clones in the contig. There have been several recent reports of local sequence duplications in the literature, suggesting that this kind of event may be a common problem in the assembly of physical maps. In this case those difficulties were resolved by sequence analysis to display the unique attributes of individual map elements. This facilitated the design of an assay that differentiated the discrete loci. The extent of the duplication around the end of YAC yWSS3586 and in a DNA segment between D7S795 and D7S603E remains to be determined.

The inclusion of ESTs derived from single-pass sequencing of cDNA libraries in YAC contigs will probably provide the most rapid route to a transcriptional map of the genome. By testing known chromosome 7 genes and ESTs, we have been able to show the presence of one EST, D7S603E, in this region. Translations of the parent sequence do not show significant homology with any known proteins. No other genes have been placed in the RP9 interval to date. In contrast, the genes AQP1, GHRHR, and ADCYAPIR have been physically localized in YACs containing the markers D7S526 or D7S632, immediately distal to the contig (Keen et al., 1995; Wajnrajch et al., 1994; Stoffel et al., 1994). In situ localization data for markers in this region have in the past been contradictory, but recent reports place D7S526 at band 7p15, while a YAC containing marker D7S484 (yWSS3586), which marks the proximal end of the RP9 interval, localizes to the 7p14–p15 boundary (Green et al., 1994). 7p14 is a G-positive or "dark staining" band, while 7p15 is a G-negative band when stained with Giemsa dye. It has been suggested that G-positive bands contain few genes, while the majority of genes lie in the G-negative bands (Bickmore and Sumner, 1989). This may therefore imply that the RP9 locus lies in band 7p14, in a region of relatively few genes.

Using the YAC clones described here, work is now underway to create a transcriptional map of this region. This should facilitate the identification of the defective gene causing this form of RP and provide candidates for other genetic diseases that may map to this region of chromosome 7.

ACKNOWLEDGMENTS

We thank the Wellcome Trust (Grant 035535/Z/92/Z) and the Foundation Fighting Blindness of the U.S.A. for funding. We also acknowledge the contribution made to this research by the UK Human Genome Mapping Project Resource Centre, the Reference Library Database, and Fondation Jean Dausset (CEPH).

REFERENCES


Localization of the Aquaporin 1 (AQPl) Gene within a YAC Contig Containing the Polymorphic Markers D7S632 and D7SS26

T. Jeffrey Keen,*† Chris F. Inglehearn,* Reshma J. Patel,* Eric D. Green,† Dale C. Peluso,‡ and Shomi S. Bhattacharya*

*Department of Molecular Genetics, Institute of Ophthalmology, Bath Street, London, ECIV 9EL, United Kingdom; and †National Center for Human Genome Research, National Institutes of Health, 9000 Rockville Pike, Bethesda, Maryland 20892

Received July 27, 1994; revised November 1, 1994

The aquaporin protein acts as a water selective, transmembrane channel (5). It is expressed in a wide range of tissues and organs and is especially abundant in the anterior segment of the eye (6, 7). Studies have shown that there is only a single AQPl gene locus, which has been localized by in situ hybridization to chromosome 7p14 (1, 5). A study by Deen et al. (1) failed to identify any RFLP at the AQPl locus. A poly(CA) sequence 400 bp upstream of the transcription start site also proved to be nonpolymorphic (1). The same authors identified six other sequence-tagged sites from a single cosmid containing the entire AQPl gene. These all contained repeat motifs, but none was investigated for possible size variation at that time.

In an attempt to obtain a more accurate genetic localization, we designed primers to amplify the longest and least interrupted of these repeat motifs (EMBL Accession No. Z21985). A forward primer GGCACGAGAGGTGTGGC and a reverse primer ATGCTGGCAAACACATGCAC gave a product size of 86 bp containing the repeat. PCR analysis in 20 individuals revealed no size variation. However, amplification of this STS from DNA of YAC clones known to localize to the 7p14–p15 region (2) revealed that it was located on a contig of three YACs also containing the microsatellite markers D7S632 and D7SS26 (Fig. 1). It was possible to deduce from the contig that AQPl is proximal to D7S632, but its position relative to D7SS26 is unknown. The published genetic distance between D7S632 and D7SS26 is only 1 cM (2). Therefore, despite AQPl not having a polymorphism directly associated with it, these two highly informative microsatellites are within 370 (D7S526) and 400 (D7SS632) kb of the genetic distance between D7S632 and D7SS26 (Fig. 1). It was possible to deduce from three YAC clones also containing the microsatellite markers D7S632 and D7SS26 (Fig. 1), proximal to D7S632 and the disease region proximal to D7SS26 (data not shown). Thus, aquaporin is excluded as a candidate gene in this family. Another group has mapped dominant cystoid macular dystrophy (DCMD) to the interval D7S526–D7SS493 (4), which overlaps our physical placement of AQPl. DCMD is characterized by an early-onset cystoid macular oedema, and it has been suggested that the disease involves a dysfunction of the retinal pigment epithelium. On the basis of the localization described here and its role in fluid movement, the aquaporin gene should therefore be considered a candidate for DCMD.

ACKNOWLEDGMENTS

We thank the Wellcome Trust (Grant 035535/Z/92/Z) for funding C.F.I. and T.J.K. E.D.G. is a Lucille P. Markey Scholar. This work was also in part supported by grants from the Lucille P. Markey Charitable Trust and National Institutes of Health (Grant F50-H009201) to E.D.G.

REFERENCES


600 BRIEF REPORTS


**503**

**494**

**7717**

**494**

**8393**

**litt@ohsu.edu**

**Health Sciences University, L333, Portland, OR 97201-3098.**

**Received August 19, 1994; revised November 21, 1994.**

**503**

**494**

**7543/95 $6.00**

**Copyright © 1995 by Academic Press, Inc.**

**All rights of reproduction in any form reserved.**

**503**

**494**

**7543/95 $6.00**

**Copyright © 1995 by Academic Press, Inc.**

**All rights of reproduction in any form reserved.**

---

**Dinucleotide Repeat Polymorphism at the Human Chromosome 11p Telomere (D11S2071)**

David L. Browne,* Eric A. Smith,* Jeanne Dietz-Band,+ Harold C. Riethmann,+ Thas Phromchotikul,* and Michael Litt*".

*Departments of Biochemistry and Medical Genetics, Oregon Health Sciences University, Portland, Oregon; +Oncor, Inc., Gaithersburg, Maryland; and *Wistar Institute, Philadelphia, Pennsylvania

---

The recent publication of detailed PCR-based linkage maps for all human chromosomes (3, 5, 9) has greatly facilitated linkage mapping of disease genes. However, most of the markers on these maps have been isolated using approaches based on random screening of genomic libraries. As a consequence, the completeness of coverage of the map in the telomeric chromosomal regions is uncertain. To remove this uncertainty and to aid in positional cloning of disease genes in telomeric regions, it is necessary to develop polymorphic markers that are physically linked to the telomeres of all chromosomes. The need for telomeric markers is highlighted by the discovery that the highest gene concentrations in the human genome are in subtelomeric regions (11).

The successful cloning of human chromosomal telomeres in modified YAC vectors (10) has paved the way for the development of such markers. Thus, polymorphic markers have been developed from YAC clones containing the telomeres of human chromosomes 7q (6), 2p, 10p, 12p, 13q, and 14q (12). Here, we report the isolation and characterization of a highly polymorphic (CA), microsatellite marker from a YAC containing the telomere of human chromosome 11p.

**Table 1**

<table>
<thead>
<tr>
<th>Allele</th>
<th>Size (nt)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>202</td>
<td>0.013</td>
</tr>
<tr>
<td>A2</td>
<td>200</td>
<td>0.021</td>
</tr>
<tr>
<td>A3</td>
<td>198</td>
<td>0.006</td>
</tr>
<tr>
<td>A4</td>
<td>196</td>
<td>0.19</td>
</tr>
<tr>
<td>A5</td>
<td>194</td>
<td>0.07</td>
</tr>
<tr>
<td>A6</td>
<td>192</td>
<td>0.17</td>
</tr>
<tr>
<td>A7</td>
<td>190</td>
<td>0.05</td>
</tr>
<tr>
<td>A8</td>
<td>188</td>
<td>0.00*</td>
</tr>
<tr>
<td>A9</td>
<td>186</td>
<td>0.007</td>
</tr>
<tr>
<td>A10</td>
<td>184</td>
<td>0.00*</td>
</tr>
<tr>
<td>A11</td>
<td>182</td>
<td>0.013</td>
</tr>
<tr>
<td>A12</td>
<td>180</td>
<td>0.007</td>
</tr>
<tr>
<td>A13</td>
<td>178</td>
<td>0.00*</td>
</tr>
<tr>
<td>A14</td>
<td>176</td>
<td>0.05</td>
</tr>
<tr>
<td>A15</td>
<td>174</td>
<td>0.007</td>
</tr>
<tr>
<td>A16</td>
<td>172</td>
<td>0.13</td>
</tr>
<tr>
<td>A17</td>
<td>170</td>
<td>0.007</td>
</tr>
<tr>
<td>A18</td>
<td>168</td>
<td>0.007</td>
</tr>
</tbody>
</table>

*Alleles A9, A10, and A13 were not observed in the population studied, but may be observed as the sample size increases.