

**Gene dosage and the molecular mechanisms
of Pelizaeus-Merzbacher disease**

Maria Cundall

A thesis submitted to the University of London
for the degree of Doctor of Philosophy

Institute of Child Health
University College London

August 2004

UMI Number: U602657

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI U602657

Published by ProQuest LLC 2014. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

Pelizaeus-Merzbacher disease (PMD) is an X-linked neurological disorder characterised by dysmyelination of the central nervous system. The major molecular defect involved in PMD is a variably sized duplication of Xq22 including the *PLP1* gene.

Determining the copy number of *PLP1* is the first line of enquiry when trying to establish a molecular diagnosis of PMD. Multiplex amplifiable probe hybridisation is a quantitative technique that has been explored as a method for diagnosing *PLP1* duplications in PMD patients and female carriers as part of this project.

The rearrangements in a number of PMD families were investigated using FISH, PCR and sequencing of chromosome breakpoints. One tandem duplication and two more complex rearrangements were characterised during this study. A bioinformatic investigation of sequences present at the breakpoints in these families, and the genomic sequence throughout Xq22, has provided some insights into possible mechanisms causing duplications of this region.

Increased dosage of Xq26-27 has been associated with pituitary hypoplasia and a mutation involving the *SOX3* gene, located at Xq27.1, has been reported in a family with growth hormone deficiency. Characterisation of a duplication involving *SOX3* in a family with X-linked hypopituitarism has been carried out using a similar strategy to the *PLP1* duplications. This *SOX3* duplication is the smallest yet reported and only contains two other genes. This provides strong evidence for increased dosage of *SOX3* being involved in the aetiology of X-linked hypopituitarism.

Gene dosage is increasingly being recognised as a cause for human genetic disorders, and it is important for diagnosis that changes in gene dosage can be reliably detected. As more cases of human genetic disease involving gene dosage become apparent, it is clear that duplications are frequent occurrences in the human genome, which may often be undetected as a cause of human genetic variation and disease.

ACKNOWLEDGEMENTS

I would especially like to thank my supervisor Dr Karen Woodward for her constant encouragement, guidance, understanding and patience throughout this project. Many thanks also to Professor Sue Malcolm, Dr Mehul Dattani and Professor Pete Scambler, who have contributed to the supervision and development of this project over the past few years and whose input has been invaluable.

Many thanks to all the past and present members of the Molecular Genetics laboratory for all their friendship and help throughout my time here, and especially Kerra for running all my sequencing. I would also like to thank the many people that have collaborated on and contributed to this project, including Dr Nigel Carter and Dr Susan Gribble, for their help with flow-sorting and fibres, Dr Gareth Howell and Ian Barrett for their knowledge of Xq22, Professor John Armour and Dr Jess Tyson for advice on MAPH, Dr Grace Hobson and Karen Sperle for their breakpoint sequences and Rodger Palmer and Paula Stubbs for all their help with the pituitary samples.

Finally, I would like to thank my family, especially my parents for their love and support, and Simon, for everything.

TABLE OF CONTENTS

ABSTRACT.....	2
ACKNOWLEDGEMENTS.....	3
TABLE OF CONTENTS.....	4
LIST OF FIGURES	15
LIST OF TABLES.....	18
LIST OF ABBREVIATIONS.....	23
1.0. INTRODUCTION	27
1.1. Pelizaeus-Merzbacher Disease.....	27
1.2. Myelin.....	28
1.3. Proteolipid protein and DM20	30
1.3.1. Evolutionary conservation of PLP1/DM20	33
1.3.2. PLP1 mutations.....	33
1.3.3. Function of PLP1 and DM20	34
1.3.4. Animal models.....	35
1.3.5. Gene duplication	36
1.3.6. Atypical duplications	38
1.3.7. Point mutations and PLP1 deletions	40
1.3.8. Disease mechanisms	41
1.3.8.1. Duplications.....	41
1.3.8.2. Point mutations	42
1.3.8.3. Deletions/null mutations	43
1.3.8.4. Affected females	43
1.4. Gene dosage.....	44
1.4.1. Genomic disorders	47
1.4.1.1. Direct repeats	48
1.4.1.2. Inverted repeats.....	51
1.4.1.3. Gross chromosomal rearrangements.....	54

1.4.1.4. Partial gene deletions and duplications.....	55
1.4.2. Gene dosage in cancer	56
1.4.3. PMD – is it a genomic disorder?.....	56
1.5. Detection of gene dosage	58
1.5.1. Real-time PCR	58
1.5.2. MLPA	59
1.5.3. Comparative genomic hybridisation.....	59
1.5.4. MAPH.....	60
1.5.4.1. MAPH and PLP1 gene dosage.....	62
1.6. SOX3 gene dosage and X-linked hypopituitarism.....	63
1.6.1. SOX3.....	64
1.6.2. Pituitary function	66
1.6.2.1. Anterior pituitary gland.....	66
1.6.2.2. Posterior pituitary gland.....	67
1.6.2.3. Intermediate pituitary gland.....	68
1.6.2.4. The hypothalamic-pituitary axis	68
1.6.3. Pituitary development	70
1.6.3.1. Signalling molecules and transcription factors in pituitary development.....	71
1.6.3.2. SOX3 in pituitary development	73
1.6.4. Screening for changes in SOX3 gene dosage	73
1.7. Objectives:	74
 2.0. MATERIALS AND METHODS.....	 75
2.1. General materials	75
2.1.1. Products and reagents	75
2.1.2. Buffers, solutions and mixes.....	76
2.1.2.1. PCR Buffers.....	76
2.1.2.2. Solutions	76
2.1.2.3. Media solutions.....	77
2.1.2.4. Plasmid alkaline lysis extraction solutions	78
2.1.2.5. Probe labelling kits	78
2.1.2.6. Polyacrylamide gel solutions	79
2.1.3. Primers	79
2.2. Methods.....	80
2.2.1. Polymerase Chain Reaction (PCR).....	80

2.2.1.1. Long range PCR.....	81
2.2.1.2. Degenerate oligonucleotide primed PCR.....	81
2.2.1.3. UPQFM-PCR.....	81
2.2.1.3.1. UPQFM-PCR dosage analysis.....	82
2.2.1.4. Inverse PCR.....	84
2.2.1.4.1. Genomic DNA restriction digestion.....	85
2.2.1.4.2. Ligation.....	85
2.2.1.4.3. PCR reaction.....	85
2.2.1.4.4. Extraction of DNA from agarose gels.....	86
2.2.2. Electrophoresis.....	86
2.2.2.1. Electrophoresis of PCR products using ABI 377 DNA sequencer.....	86
2.2.2.1.1. Preparation of polyacrylamide gels.....	86
2.2.2.2. Agarose gel electrophoresis.....	87
2.2.3. DNA precipitation.....	88
2.2.3.1. Phenol/chloroform extraction.....	88
2.2.3.2. Ethanol precipitation of DNA.....	88
2.2.4. Sequencing.....	89
2.2.4.1. Sephadex sequencing reaction clean-up.....	89
2.2.4.2. MegaBACE sequencing.....	89
2.2.5. Preparation of FISH probes from human genomic clones.....	90
2.2.5.1. Glycerol stocks.....	90
2.2.5.2. Cosmid/PAC/BAC/plasmid alkaline lysis minipreparation.....	90
2.2.6. Cell culture of lymphoblastoid cell lines.....	91
2.2.6.1. Freezing down of lymphoblastoid cell lines.....	91
2.2.6.2. Preparation of cell suspensions for interphase and metaphase FISH from lymphoblastoid cell line.....	92
2.2.6.3. Slide preparation for DNA fibres.....	92
2.2.7. Probe labelling for FISH.....	93
2.2.7.1. Biotin labelling.....	93
2.2.7.2. Digoxigenin labelling.....	93
2.2.7.3. Direct labelling of probes.....	94
2.2.8. Fluorescence in-situ hybridisation.....	94
2.2.8.1. Slides/suspensions.....	94
2.2.8.2. FISH using probes labelled with biotin and digoxigenin.....	95
2.2.8.2.1. Probe precipitation.....	95

2.2.8.2.2. Slide and probe preparation	95
2.2.8.2.3. Overnight hybridisation	96
2.2.8.2.4. Washes and antibodies	96
2.2.8.3. FISH using directly labelled probes	97
2.2.8.3.1. Probe precipitation	97
2.2.8.3.2. Slide and probe preparation	98
2.2.8.3.3. Denaturation and hybridisation	98
2.2.8.3.4. Washes and mounting	98
2.2.8.4. Viewing of FISH slides and capture of images	99
2.2.8.5. Scoring FISH slides	99
2.2.9. MAPH	100
2.2.9.1. Probe design	100
2.2.9.2. PCR of probe target sequence	100
2.2.9.3. Cloning MAPH probes	100
2.2.9.4. Removal of 3' A residues	101
2.2.9.5. Ligation and transformation	101
2.2.9.5.1. Vector digestion	101
2.2.9.5.2. Ligation	101
2.2.9.5.3. Transformation	102
2.2.9.6. Probe preparation	102
2.2.9.7. MAPH protocol	106
2.2.9.7.1. DNA and filter preparation	106
2.2.9.7.2. Hybridisation	106
2.2.9.7.3. Washes	106
2.2.9.8. MAPH Analysis	107
2.2.10. Bioinformatics	108
2.2.10.1. Sequence database searches and pairwise sequence comparisons	108
2.2.10.1.1. BLASTn	108
2.2.10.1.2. BLAST2	109
2.2.10.1.3. BLASTz	109
2.2.10.1.4. ClustalW and Consensus	109
2.2.10.2. Repeatmasker	110
2.2.10.3. DNA Pattern Finder	110
2.2.10.4. MAR-Wiz	110
2.2.10.5. Oligorep	111

2.2.10.6. Tandem repeats finder.....	111
2.2.10.7. Genome browsers.....	111
2.3. Clinical details	115
2.3.1. Family 1	115
2.3.2. Family 2	115
2.3.3. Family 3	115
2.3.4. Family 4	116
 3.1. IN SILICO ANALYSIS OF PLP1 GENOMIC REGION.....	118
3.2. Repeats specific to the 2Mb region flanking PLP1	118
3.3. Repetitive sequences flanking PLP1.....	119
3.3.1. Repeated flanking sequences close to PLP1	119
3.3.2. More distant repeated sequences either side of PLP1	120
3.4. Repeats distal to PLP1	124
3.4.1. Previously described LCRs distal to PLP1	124
3.4.2. Novel distal LCRs.....	125
3.4.3. Short regions of similarity to the large distal LCRs	125
3.4.4. Sequence similarity between distal repeats and sequence proximal to PLP1.....	126
3.4.5. H2b-like genes	131
3.4.5.1. H2b-like genes in other sequenced genomes	131
3.4.6. Xrep enhancer	132
3.4.7. G+C and interspersed repeat content of distal repeats.....	133
3.5. Repeats proximal to PLP1	135
3.5.1. Proximal repeat group P1.....	135
3.5.2. P2 repeats	140
3.5.2.1. RAB genes	140
3.5.2.2. Sequence characteristics of P2 repeats	141
3.5.3. P3 repeats	144
3.5.3.1. BEX/NADE genes	144
3.5.3.2. Sequence characteristics of P3 repeats	145
3.5.4. P4 repeats	147
3.5.4.1. TCEAL1-like genes	147
3.5.4.2. Sequence characteristics of P4 repeats	147
3.5.5. Similarities between P3 (BEX-like) repeats and P4 (TCEAL1-like) repeats	150
3.6. Comparison of local repeat content near PLP1 with other species.....	152

3.6.1. Regional repeats in the chimpanzee syntenic region	152
3.6.2. Repetitive regions close to the murine Plp1 gene	154
3.6.3. Repetitive regions near the brown rat PLP1 homologue	155
3.7. Discussion	158
3.7.1. LCRs/segmental duplications near PLP1	158
3.7.2. Interspersed repeats	158
3.7.3. Gene families in the PLP1 region	160
3.7.4. Repetitive regions in human and other species	161
3.8. Summary	162
 4.1 DUPLICATION BREAKPOINT MAPPING IN FAMILY 1	 163
4.1.1. Duplication breakpoint mapping – previously published work	163
4.2. Fine mapping of duplication breakpoints in family 1 by interphase FISH and UPQFM-PCR	164
4.2.1. Proximal breakpoint mapping by interphase FISH	165
4.2.2. UPQFM-PCR mapping of proximal breakpoint in family 1	165
4.2.3. UPQFM-PCR mapping of distal breakpoint	166
4.3. Fine mapping of duplication breakpoints by fibre FISH	170
4.3.1. Fibre-FISH to determine normal relationships between clones	170
4.3.1.1. Fibre-FISH for the proximal breakpoint region	171
4.3.1.2. Fibre-FISH for the distal breakpoint region	171
4.3.2. Demonstrating tandem duplication breakpoint junctions by fibre-FISH	172
4.4. Long-range PCR and direct sequencing to span the breakpoint junction	178
4.4.1. Long-range PCR across duplication breakpoint	178
4.4.2. Sequencing the duplication breakpoint in family 1	179
4.5. Interspersed repetitive sequences and G+C content at the duplication breakpoints in family 1	179
4.6. Searching for similarities between the proximal and distal duplication breakpoints	186
4.7. Location of proximal duplication breakpoint relative to PLP1-proximal repeats	186
4.8. Location of distal breakpoint within low-copy repeats	187
4.9. Analysis of sequence from breakpoint regions for recombination and rearrangement-associated motifs	190
4.9.1. 5Kb regions around breakpoints	190

4.9.1.1. Calculating relative occurrence of motifs in the 5Kb regions around breakpoints	190
4.9.1.2. Matrix attachment regions	191
4.9.2. Sequence features found in 100bp regions around breakpoints.....	191
4.9.2.1. Alternating purine/pyrimidines, polypurine and polypyrimidine tracts	191
4.9.2.2. Inverted repeats and secondary structures	192
4.10. In silico analysis of sequence in 5Kb regions around breakpoints	193
4.10.1. Recombination/rearrangement-associated motifs.....	193
4.10.2. Potential MARs near duplication breakpoints in family 1.....	194
4.11. In silico analysis of sequence from 100bp regions flanking breakpoints	194
4.11.1. Alternating purine/pyrimidines, polypurine and polypyrimidine tracts	194
4.11.2. Repeats and secondary structures	195
4.12. Discussion.....	199
4.12.1. Interphase FISH and duplication mapping	199
4.12.2. NHEJ.....	201
4.12.2.1 Mechanisms for DSB formation.....	201
4.12.2.2. Xrep, replication origins and DSBs	202
4.12.3. Origin of the duplication in family 1	203
4.12.4. DSBs, recombination and meiosis	204
4.12.5. DSBs, strand invasion and replication.....	205
4.12.6. Interspersed repetitive elements.....	210
4.12.7. Nucleotide content of breakpoints	211
4.13. Summary.....	211
 5.1. DUPLICATION BREAKPOINT MAPPING IN FAMILY 2.....	212
5.1.1. Previously published data	212
5.2. FISH mapping proximal duplication breakpoint in family 2.....	212
5.2.1. Further mapping of proximal duplication breakpoint by interphase FISH.....	212
5.3. Mapping of proximal duplication breakpoint by UPQFM-PCR.....	213
5.4. Inverse PCR to obtain breakpoint sequences.....	217
5.4.1. Inverse PCR strategy.....	217
5.4.2. Inverse PCR results.....	218
5.4.3. Sequencing of inverse PCR products.....	220
5.5. Verification of cU177E8-1055C14 breakpoint by long-range PCR.....	223
5.6. Determination of sequence copy number around dJ1055C14 in family 2.....	225

5.6.1. UPQFM-PCR.....	225
5.6.2. Interphase FISH around dJ1055C14.....	226
5.7. Analysis of genomic sequence around the two proximal breakpoints in family 2	233
5.7.1. Genes near breakpoint closer to PLP1	233
5.7.2. Genes near proximal breakpoint	233
5.7.3. Proximal region-specific repeats.....	234
5.7.4. Interspersed repetitive elements and G+C content	234
5.7.5. Comparisons of sequence around cU177E8 and dJ1055C14 breakpoints.....	238
5.7.6. Recombination/rearrangement-associated motifs.....	241
5.7.6.1. 5Kb around breakpoint in cU177E8	241
5.7.6.2. 5kb around breakpoint in dJ1055C14.....	241
5.7.7. Matrix attachment regions	242
5.7.8. In silico analysis of sequence from 100bp regions flanking breakpoints	243
5.7.8.1. Alternating purine/pyrimidines, polypurine and polypyrimidine tracts	243
5.7.8.2. Repeats and secondary structures	245
5.8. Discussion.....	251
5.8.1. Sequencing of a breakpoint in family 2 has revealed a complex rearrangement	251
5.8.2. Nature and possible mechanisms of the rearrangement.....	253
5.8.2.1. Insertion of a short sequence at the breakpoint.....	253
5.8.2.2. Recombination/rearrangement associated motifs	254
5.8.2.3. Polypurine tracts and short repeats	254
5.8.2.4. Triplex DNA and secondary structure	255
5.8.2.5. DNA replication and rearrangement in family 2	260
5.8.2.6. Capture of long sequences at DSBs.....	260
5.9. Summary.....	262
 6.1. ANALYSIS OF FAMILY 3	263
6.2. Xq26 PLP1 insertion family mosaic deletion analysis	264
6.2.1. Metaphase FISH.....	264
6.2.2. STS analysis on flow-sorted chromosomes	268
6.2.2.1. Disparity between FISH and STS mapping of deletion breakpoint.....	269
6.2.3. Amplifying deletion breakpoint by LR-PCR.....	273
6.2.4. Sequencing of deletion breakpoint.....	273

6.2.5. Analysis of deletion breakpoints.....	276
6.2.5.1. Interspersed repeats and G+C content	276
6.2.5.2. Comparing sequences near the deletion breakpoints	277
6.2.5.3. Gene content near deletion breakpoints	277
6.2.5.4. Analysis of deletion breakpoint regions for recombination-associated motifs	283
6.2.5.4.1. 5Kb regions around deletion breakpoints	283
6.2.5.4.2. Matrix attachment regions	284
6.2.5.4.3. Detailed analysis of 100bp around each deletion breakpoint	284
6.2.5.4.3.1. Alternating purine/pyrimidines, polypurine and polypyrimidine tracts	284
6.2.5.4.3.2. Inverted repeats and secondary structures	284
6.3. Xq26 insertion mapping.....	288
6.3.1. Metaphase FISH.....	288
6.3.2. Duplication mapping.....	291
6.3.2.1. Metaphase FISH.....	291
6.3.2.2. UPQFM-PCR and fine mapping of proximal duplication breakpoint.....	291
6.3.2.3. Inverse PCR and sequencing of insertion breakpoint	294
6.3.2.3.1. Inverse PCR strategy.....	294
6.3.2.3.2. Sequence results for inverse PCR.....	296
6.3.2.3.3. Confirmation of insertion by PCR.....	296
6.3.3. Analysis of sequences around the duplication and insertion breakpoints.....	299
6.3.3.1. Genomic features near Xq22.2 proximal duplication breakpoint.....	299
6.3.3.2. Genomic features near Xq26.2 insertion breakpoints.....	299
6.3.3.2.1. Genomic features near dJ305B16 inserted sequence.....	299
6.3.3.2.2. Genomic features near dJ197O17 inserted sequence.....	300
6.3.3.2.3. Genomic features near bA453F18 insertion point.....	300
6.3.3.3. Comparison of sequences surrounding breakpoints	300
6.3.3.3.1. 5Kb surrounding breakpoint sequences	300
6.3.3.3.2. 100bp surrounding breakpoints.....	301
6.3.3.4. Recombination/rearrangement associated sequence motifs.....	308
6.3.3.4.1. 5Kb around dJ1055C14 duplication breakpoint	308
6.3.3.4.2. 5Kb around dJ305B16 inserted sequence	309
6.3.3.4.3. 5Kb around dJ197O17 inserted sequence.....	310
6.3.3.4.4. 5Kb around bA453F18 insertion point	311
6.3.3.5. Matrix attachment regions	311

6.3.3.6. In silico analysis of sequence from 50bp regions either side of breakpoints	311
6.3.3.6.1. Purine/pyrimidine content.....	311
6.3.3.6.2. Inverted repeats and secondary structure near breakpoints	312
6.3.3.7. Analysis of genomic features in the wider insertion region	318
6.3.3.7.1. Genes.....	318
6.3.3.7.2. Region – specific repeats	320
6.4. Discussion.....	325
6.4.1. Summary of rearrangement in family 3	325
6.4.2. Matrix attachment regions	327
6.4.3. Sequence motifs, G+C content and other features of the DNA sequence	328
6.4.4. Insertion	329
6.4.5.1. Repetitive regions near Xq26.2 insertion.....	329
6.4.5.2. Mechanism of insertion.....	331
6.4.6. Mild phenotype in family 3 – a possible position effect?.....	332
6.5. Conclusions.....	333
 7.1. MAPH RESULTS.....	 334
7.1.1 MAPH probe mix.....	334
7.1.2. MAPH experiments	336
7.2. Autosomal control probes	336
7.3. Sex-linked control probes	339
7.4. Xq22 probes in control individuals.....	340
7.4.1. PLP1 probes	342
7.4.1.1. Analysis of individual normalised ratios	345
7.4.1.2. Analysis of the significance of individual normalised ratios for MAPH probes plp5 and plp6.....	347
7.4.2. Xq22 probes flanking PLP1	349
7.4.2.1. 198p4.....	349
7.4.2.2. 79p11.....	351
7.4.2.3. 43h13.....	354
7.4.2.4. 240c2.....	356
7.4.2.5. 144a10.....	358
7.5. Discussion.....	360
7.6. Summary	363

8.1 SOX3 DUPLICATION SCREENING AND ANALYSIS OF FAMILY 4	364
8.2 Interphase FISH screening.....	364
8.3. Interphase FISH mapping of Xq27.1 duplication	365
8.4. Further mapping of duplication breakpoints in family 4 using UPQFM-PCR..	365
8.4.1. UPQFM-PCR mapping of proximal duplication breakpoint	366
8.4.2. UPQFM-PCR mapping of distal duplication breakpoint.....	366
8.5. Determining the nature of the Xq27.1 duplication by dual probe interphase FISH	372
8.6. Long-range PCR across the duplication breakpoint junction	374
8.7. Determining the inheritance of the duplication in family 4.....	378
8.8. Discussion	382
8.9. Conclusions.....	384
9.0. DISCUSSION	385
9.1. Gene dosage	385
9.2. MAPH for detecting gene dosage	385
9.3. The importance of duplications in evolution	386
9.4. Duplications involving the PLP1 genomic region	386
9.5. Possible clustering of breakpoints within clone dJ1055C14	387
9.6. Mechanisms for gene duplication	390
9.7. Future work.....	392
REFERENCES	393
APPENDIX A.....	420
APPENDIX B	430
APPENDIX C	433

LIST OF FIGURES

Figure 2.1. Diagram showing the principles of inverse PCR	84
Figure 2.2. Pedigrees showing the families investigated as part of this study.....	117
Figure 3.1. Region-specific repeats in the genomic region surrounding <i>PLP1</i>	121
Figure 3.2. Position of small inverted repeat sequences just either side of <i>PLP1</i>	123
Figure 3.3. Dotplot comparing 250Kb of human genomic sequence against itself..	127
Figure 3.4. Organisation of repeats distal to <i>PLP1</i>	128
Figure 3.5. ClustalW alignment of the H2b-like regions.....	134
Figure 3.6. Dotplot produced by Pipmaker using BLASTz comparison of sequence from 1.1Mb proximal to <i>PLP1</i> against itself.	137
Figure 3.7. Repeat content and other features of the P4 (TCEAL1-like) repeats found in Xq22 proximal to <i>PLP1</i>	149
Figure 3.8. Summary of the local repeat structure, genes and clones proximal to <i>PLP1</i>	151
Figure 3.9. Locally repetitive sequences are present in syntenic regions of other genomes.	156
Figure 3.10. Comparisons of 2Mb of sequence around the <i>PLP1</i> homologue with human genomic sequence in chimpanzee, mouse and rat.....	157
Figure 4.1. Extent of the duplication containing <i>PLP1</i> in family 1, as has been previously published	164
Figure 4.2. Interphase FISH results from the proximal duplication breakpoint using a cell line from the affected boy in family 1.....	167
Figure 4.3. Diagrams showing the proximal (a.) and distal (b.) breakpoint regions in family 1	168
Figure 4.4. Fibre FISH on normal cell lines	174
Figure 4.5. Screenshot from Sanger Institute X chromosome FPC map.....	175
Figure 4.6. Fibre FISH to confirm tandem rearrangement in family 1.....	176
Figure 4.7. Long-range PCR across the tandem breakpoint in family 1.....	181
Figure 4.8. Agarose gel stained using ethidium bromide showing LR-PCR product spanning the duplication breakpoint in family 1.....	182
Figure 4.9. Electropherogram showing sequence from the duplication breakpoint in family 1	182

Figure 4.10. Interspersed repeat content and other sequence features found in the 5Kb regions centred on the proximal and distal duplication breakpoints in family 1	183
Figure 4.11. Comparison of 5Kb around the proximal and distal breakpoints against each other by BLASTz.....	188
Figure 4.12. Alignment of duplication breakpoint junction from family 1 and 100bp flanking the two duplication breakpoints.....	189
Figure 4.13. MAR potential in the 5Kb surrounding the distal duplication breakpoint in family 1 as found by MAR-Wiz.	196
Figure 4.14. Purine and pyrimidine content of 100bp regions surrounding the proximal and distal duplication breakpoints in family 1	197
Figure 4.15. Inverted repeats around the duplication breakpoints in family 1, the duplication junction fragment and the reciprocal rearrangement.	198
Figure 4.16. Diagram showing possible mechanism for <i>PLP1</i> tandem duplications	209
Figure 5.1. Interphase FISH results from the proximal duplication breakpoint using a cell line from the affected boy in family 2.....	214
Figure 5.2. UPQFM-PCR mapping of proximal breakpoint in family 2	215
Figure 5.3. Inverse PCR strategy to clone the proximal duplication breakpoint for family 2	219
Figure 5.4. Electropherogram showing sequence from the cU177E8-dJ1055C14 breakpoint junction in family 2.....	221
Figure 5.5. Diagram illustrating the location of the breakpoints sequenced from the family 2 inverse PCR reaction and a large-scale view of the rearrangement found..	222
Figure 5.6. Confirmation of cU177E8-dJ1055C14 breakpoint by long-range PCR.	224
Figure 5.7. UPQFM-PCR and interphase FISH data mapping two breakpoints within dJ1055C14 and cV857G6.....	229
Figure 5.8. Screenshot from Sanger Institute X chromosome FPC map	232
Figure 5.9. Interspersed repeat content around breakpoints in family 2.....	237
Figure 5.10. Comparison of 5Kb around breakpoints by BLASTz	239
Figure 5.11. Alignment of breakpoint junction and 100bp flanking the sequenced breakpoint in family 2.....	240
Figure 5.12. Position of potential MARs near the proximal breakpoints in family 2	246
Figure 5.13. Nucleotide composition of short regions surrounding the breakpoints sequenced from family 2.	247

Figure 5.14. Location of internal sequence repeats around the breakpoint and junction regions in family 2.	248-249
Figure 5.15. Diagram showing a possible rearrangement in family 2	252
Figure 5.16. Symmetric repeats, Hoogsteen base pairs, and possible triplex DNA configurations.	258
Figure 6.1. FISH mapping of the proximal deletion breakpoint.....	266
Figure 6.2. FISH mapping the distal deletion breakpoint.....	267
Figure 6.3. bA79A21 overlaps substantially with dJ203P18.....	271
Figure 6.4. Screenshot from Sanger Institute X chromosome fingerprinted contig map	272
Figure 6.5 Long PCR across the Xq deletion breakpoint	274
Figure 6.6. Sequence from the deletion breakpoint in the female carrier.....	275
Figure 6.7. Figure showing 5Kb around the proximal and distal deletion breakpoint and the interspersed repeat content of these regions.....	278
Figure 6.8. Alignment of deletion breakpoint junction and 100bp flanking the two deletion breakpoints	281
Figure 6.9. Dotplot output from Pipmaker comparing 5kb regions around the proximal and distal deletion breakpoints	282
Figure 6.10. MAR potential in the 5Kb surrounding the distal deletion breakpoint as found by MAR-Wiz.	285
Figure 6.11. Purine/pyrimidine content in the region immediately surrounding both deletion breakpoints	286
Figure 6.12. Inverted repeats around the deletion breakpoints, the deletion junction fragment and the reciprocal rearrangement	287
Figure 6.13. Metaphase spreads showing dual labelling to map Xq26 PLP1 insertion	289
Figure 6.14. FISH on metaphase chromosomes from 3:4 using probes around the proximal duplication breakpoint.	292
Figure 6.15. Diagram showing strategy for inverse PCR.....	295
Figure 6.16. Inverse PCR run on agarose gel	295
Figure 6.17. The annotated sequence from the MspI iPCR band	297
Figure 6.18. Agarose gel showing long-range PCR from dJ1055C14 to bA453F18	298

Figure 6.19. Diagram showing interspersed repeat content and genes in 5kb regions centred on the breakpoints associated with the Xq26 insertion.....	302
Figure 6.20. Alignment of insertion breakpoint junction and 50bp either side of all the breakpoints or short inserted sequences.....	307
Figure 6.21. Regions of MAR potential in the regions around the duplication and insertion breakpoints in family 3	313
Figure 6.22. Purine/pyrimidine content in the 100bp region immediately surrounding both the insertion/duplication breakpoints and the sequences inserted at the breakpoints.	315
Figure 6.23. Internally repeated sequences around the breakpoints in family 3.....	316
Figure 6.24. Internally repeated sequences around the sequenced breakpoint junctions in family 3	317
Figure 6.25. Genes present in the insertion region	319
Figure 6.26. Dotplot of the region around the Xq26.2 insertion point	322
Figure 6.27. Dotplot output from Pipmaker for the whole sequence of human genomic clone bA453F18.....	323
Figure 6.28 The syntenic regions of the chimpanzee and mouse genomes also contain numerous directly repeated sequence arrays.....	324
Figure 6.29. Summary of X chromosome rearrangements found in family 3	326
Figure 7.1. Position of MAPH probes relative to <i>PLP1</i> and the surrounding genomic region.	335
Figure 7.2. Graph showing the means +/- 1 standard deviation of the normalised ratio for the six autosomal control MAPH probes throughout all experiments.	337
Figure 7.3. Histograms showing the frequency distribution of normalised ratios for the six MAPH autosomal control probes.	338
Figure 7.4. Mean normalised ratios +/- one standard deviation for the Xq12 and SRY MAPH probes for all samples.....	340
Figure 7.5. Bar chart showing means +/- 1 standard deviation for Xq22 probes in normal controls	342
Figure 7.6. Mean normalised ratios for the MAPH probe plp6.....	344
Figure 7.7. Mean normalised ratios for the MAPH probe plp5	344
Figure 7.8. Mean normalised ratios for the MAPH probe 198p4	351
Figure 7.9. Mean normalised ratios for the MAPH probe 79p11	353
Figure 7.10. Mean normalised ratios for the MAPH probe 43h13	355

Figure 7.11. Mean normalised ratios for the MAPH probe 240c2	357
Figure 7.12. Mean normalised ratios for the MAPH probe 144a10	359
Figure 8.1. A duplication including the <i>SOX3</i> gene was detected in family 4 by interphase FISH	367
Figure 8.2. Diagram showing interphase FISH and UPQFM-PCR results around the proximal and distal duplication breakpoints in family 4.	371
Figure 8.3. Interphase FISH using two genomic clones from within the duplication as probes	376
Figure 8.4. Long-range PCR across the tandem duplication breakpoint in family 4.	377
Figure 8.5. Results of genotyping various X-linked polymorphic microsatellite markers in family 4	380
Figure 9.1. Location of sequenced breakpoints within clone dJ1055C14	389

LIST OF TABLES

Table 2.1. Antibody solutions used for indirect detection of probes for FISH.....	97
Table 2.2. Table showing the panel of MAPH probes used in the PMD probe set....	104
Table 2.3. Websites of bioinformatic tools used during the course of this study.....	112
Table 3.1. Summary of the short regions of sequence similarity identified by BLASTz either side of <i>PLP1</i>	122
Table 3.2. Summary of the repeated regions distal to <i>PLP1</i>	129
Table 3.3. Similarity of the repeats that make up LCRs PMD-A and PMD-B to each other as reported from BLASTz alignments.....	130
Table 3.4. Summary of proximal P1 repeats.....	138
Table 3.5. Similarity of the P1 repeats to each other as reported from BLASTz alignments.....	139
Table 3.6. Summary of proximal P2 repeats.....	142
Table 3.7. Similarity of the P2 repeats to each other as reported from BLASTz alignments.....	143
Table 3.8. Summary of proximal P3 repeats.....	146
Table 3.9. Similarity of the P3 repeats to each other as reported from BLASTz alignments.....	146
Table 3.10. Summary of proximal P4 repeats.....	148
Table 3.11. Similarity of the P4 repeats to each other as reported from BLASTz alignments.....	148
Table 4.1. Interphase FISH scores for patient 1:9 near the proximal duplication breakpoint.....	167
Table 4.2. UPQFM ratios for the proximal and distal duplication breakpoint in family 1.....	169
Table 4.3. Interspersed repetitive elements from the 5Kb region surrounding the proximal duplication breakpoint in dJ635G19.....	184
Table 4.4. Interspersed repetitive elements from the 5Kb region surrounding the distal duplication breakpoint in dJ839M11.....	185
Table 4.5. Comparison of 5Kb surrounding the distal duplication breakpoint in family 1 against the related distal repeat units A1b, A2 and A3.....	189

Table 5.1. Interphase FISH scores for 2:9 near the proximal duplication breakpoint.....	214
Table 5.2. UPQFM ratios for the proximal duplication breakpoint in family 2.....	216
Table 5.3. Interphase FISH scores from family 2 near to the second proximal breakpoint.....	228
Table 5.4. UPQFM ratios for the proximal duplication breakpoint in family 2.....	231
Table 5.5. Interspersed repetitive elements from the 5Kb region surrounding the breakpoint in cU177E8.....	235
Table 5.6. Interspersed repetitive elements from the 5Kb region surrounding the breakpoint in dJ1055C14.....	236
Table 6.1. Results of deletion breakpoint analysis.....	265
Table 6.2. Results for STS mapping of Xq deletion breakpoint.....	270
Table 6.3. Repeat content of the 5Kb genomic region centred on the proximal deletion breakpoint within bA346E8.....	279
Table 6.4. Repeat content of the 5Kb genomic region centred on the distal deletion breakpoint within dJ203P18.....	280
Table 6.5. Table showing results from insertion mapping using dual FISH.....	290
Table 6.6. UPQFM-PCR results for primers amplifying sequences within dJ1055C14	293
Table 6.7. Repeat content of the 5Kb genomic region surrounding the proximal duplication breakpoint within dJ1055C14.....	303
Table 6.8. Interspersed repetitive elements from the 5Kb region surrounding the duplicated bases in the rearrangement from dJ305B16.....	304
Table 6.9. Interspersed repetitive elements from the 5Kb region surrounding the duplicated bases in the rearrangement from dJ197O17.....	305
Table 6.10. Interspersed repetitive elements from the 5Kb region centred on position 24956 in clone bA453F18.....	306
Table 6.11. Regions of relatively high MAR potential near the sequences involved in the Xq26.2 insertion.....	314
Table 7.1. Means and standard deviations for 6 MAPH autosomal control probes..	337
Table 7.2. Means and standard deviations (SD) for the X- and Y-linked MAPH probes.....	341

Table 7.3. Means and standard deviations for the different classes of PMD patients and controls for MAPH probes plp5 and plp6.....345

Table 8.1. Interphase FISH scores from the affected males in family 4, for clone bA51C14 and other clones that were close to the boundaries of the duplicated region.....368

Table 8.2. Interphase FISH scores from females in family 4, for clone bA51C14...369

Table 8.3. UPQFM ratios for the duplication breakpoints in family 4.....370

Table 8.4. Appearance of signals in interphase nuclei following hybridisation of two probes from within the duplication region in family 4.....376

LIST OF ABBREVIATIONS

A	Adenine	CO ₂	Carbon dioxide
ACTH	Adrenocorticotrophic hormone	D	Adenine, guanine or thymine
ADH	Antidiuretic hormone	DAPI	4',6-diamidino-2-phenylindole
APS	Ammonium persulphate	dATP	2'-deoxyadenosine 5'-triphosphate
ARS	Autonomously replicating sequence	dCTP	2'-deoxycytidine 5'-triphosphate
AS	Angelman syndrome	DDBJ	DNA Data Bank of Japan
ATP	Adenosine triphosphate	dGTP	2'-deoxyguanosine 5'-triphosphate
BAC	Bacterial artificial chromosome	DIG	Digoxigenin
BLAST	Basic local alignment search tool	DMSO	Dimethyl sulphoxide
bp	Base pair	DNA	Deoxyribonucleic acid
BSA	Bovine serum albumin	DOP-PCR	Degenerate oligonucleotide primed PCR
C	Cytosine	dNTP	2'-deoxynucleotide 5'-triphosphate
C-	Carboxy	DTT	DL-Dithiothreitol
°C	Degrees centigrade	dTTP	2'-deoxythymidine 5'-triphosphate
CCD	Charge coupled device	dUTP	2'-deoxyuridine 5'-triphosphate
cDNA	Complementary deoxyribonucleic acid	DSB	Double-strand breaks
cen	Centromere	EBV	Epstein-Barr virus
CEP	Chromosome enumeration probe	FAM	6-carboxyfluorescein
CGH	Comparative genomic hybridisation	FCS	Foetal calf serum
cM	Centimorgan	FSH	Follicle-stimulating hormone
CMT1A	Charcot-Marie-Tooth disease type 1A	g	Gram
CNS	Central nervous system		

<i>g</i>	Gravitational force	KCl	Potassium chloride
G	Guanine	kDa	Kilodalton
GH	Growth hormone	KV	Kilovolt
GHRH	Growth hormone releasing hormone	LB	Luria-Bertani medium
GnRH	Gonadotrophin releasing hormone	LCR	Low copy repeat
		LH	Luteinising hormone
GTP	Guanosine triphosphate	LINE	Long interspersed nuclear element
H	Adenine, cytosine or thymine	LPA	Linear polyacrylamide
HEX	4,7,2',4',5',7'-hexachloro-6-carboxyrhodamine	LR-PCR	Long range PCR
HgSO ₄	Mercury sulphate	LTR	Long terminal repeat
		M	Molar; adenine or cytosine
HMG	High mobility group	mA	Milliamp
HNPP	Hereditary neuropathy with liability to pressure palsies	MAPH	Multiplex amplifiable probe hybridisation
H ₂ O	Water	MAR	Matrix attachment regions
HR	Homologous recombination	Mb	Megabase
EDTA	Ethylenediamine tetra-acetic acid	MER	Medium reiteration frequency sequence
EMBL	European Molecular Biology Laboratory	mg	Milligram
ER	Endoplasmic reticulum	Mg	Magnesium
EST	Expressed sequence tag	MgCl ₂	Magnesium chloride
FISH	Fluorescence <i>in situ</i> hybridisation	MgSO ₄	Magnesium sulphate
FITC	Fluorescein isothiocyanate	MHC	Major histocompatibility complex
FPC	Fingerprinted contig	μg	Microgram
iPCR	Inverse PCR	μl	Microlitre
K	Guanine or thymine	μM	Micromolar
kb	Kilobase		

μmol	Micromole	PBS	Phosphate buffered saline
MIR	Mammalian-wide interspersed repeat	PCR	Polymerase chain reaction
ml	Millilitre	PDB	Protein Data Bank
MLPA	Multiplex ligation-dependent probe amplification	pmol	Picomole
		PMD	Pelizaeus-Merzbacher disease
mm	Millimetre	PMLD	Pelizaeus-Merzbacher-like disease
mM	Millimolar		
MRI	Magnetic resonance	PNS	Peripheral nervous system
imaging		PRL	Prolactin
MSH	Melanocyte stimulating hormone	PWS	Prader-Willi syndrome
		R	Purine
mW	Milliwatt	RFLP	Restriction fragment length polymorphism
N	Adenine, cytosine, guanine or thymine	RNA	Ribonucleic acid
N-	Amino	rpm	Revolutions per minute
NaCl	Sodium chloride	S	Cytosine or guanine
NAHR	Non-allelic homologous recombination	SD	Standard deviation of the mean
NaH ₂ PO ₄	Sodium dihydrogen phosphate	SDS	Sodium dodecyl sulphate
NaOH	Sodium hydroxide	SHFM3	Split hand-split foot malformation 3
NCBI	National Center for Biotechnology Information	SINE	Short interspersed nuclear element
ng	Nanogram	SMS	Smith-Magenis syndrome
NHEJ	Non-homologous end joining	snRNA	Small nuclear ribonucleic acid
(NH ₄) ₂ SO ₄	Ammonium sulphate	SPG2	Spastic paraplegia type 2
OR	Olfactory receptor	STS	Sequence tagged site
PAC	P1-derived artificial chromosome	T	Thymine

tel	Telomere
TEMED	N,N,N',N'- tetramethylethylenediamine
TET	6-tetrachlorofluorescein
TSH	Thyroid stimulating hormone
U	Unit
UCSC	University of California Santa Cruz
UPQFM-PCR	Universal primer quantitative fluorescent multiplex PCR
URL	Uniform resource locator
UTR	Untranslated region
UV	Ultraviolet light
V	Volt
VCFS	Velo-cardio-facial syndrome
V(D)J	Variable (diversity) joining
W	Adenine or thymine; Watt
WBS	Williams-Beuren syndrome
Y	Pyrimidine

1.0. INTRODUCTION

1.1. Pelizaeus-Merzbacher Disease

Pelizaeus-Merzbacher disease (PMD) is an X-linked recessive dysmyelinating disorder of the central nervous system (CNS). Absence or a severe deficit of white matter is seen on MRI scans, due to the failure of CNS oligodendrocytes to myelinate axons. The symptoms of PMD vary in severity, but often include nystagmus, ataxia and spasticity, with onset of symptoms usually within the first year of life (Hodes *et al.*, 1993). Other symptoms include stridor and seizures. PMD is a progressive disease, and premature death usually occurs from childhood into the third decade of life. The most severely affected cases, where symptoms are apparent from birth (the “connatal” form), are associated with early mortality, often before the first birthday. PMD is allelic with a milder disorder, spastic paraplegia type 2 (SPG2), which exhibits symptoms of progressive spasticity and weakness of the lower extremities. SPG2 has a generally later onset and mildly affected patients may survive into the seventh decade of life, and be able to talk and walk (Saugier-Veber *et al.*, 1994). Severity of PMD/SPG2 varies from the most severe connatal forms, where symptoms are seen soon after birth, with severe developmental delay and seizures, with rapid disease progression leading to a very early death, to the mildest SPG2 phenotype. Affected individuals are usually male, and females are often carriers, with no apparent symptoms, although there is a small subset of affected females, who generally show mild symptoms with a later onset (Hodes *et al.*, 1995; Nance *et al.*, 1996; Hodes *et al.*, 1997; Inoue *et al.*, 2001).

1.2. Myelin

Myelin is a multilammellar spiral membrane structure that surrounds axons in the nervous system in higher vertebrates. Each segment of myelin sheath is separated by a gap, where the axon is exposed, called a node of Ranvier. The insulation of the axons by the myelin sheath facilitates faster conduction of nerve impulses, as the impulse jumps from one node of Ranvier to the next by saltatory conduction. Myelin in the peripheral nervous system (PNS) is formed by Schwann cells, each cell wrapping itself around a single axon to create a single internodal section of myelin sheath. In the CNS, a single oligodendrocyte can send out multiple cytoplasmic processes from the cell body and myelinate several axons at once (Figure 1.1). Electron microscopy of cross-sections of compact myelin shows a layered structure consisting of alternating electron-dense and electron-light layers (Figure 1.1). The major dense line is formed by the cytoplasmic faces of the cell surface membrane in close proximity to each other and the extracellular sides of the membranes form the less intense double intraperiod line (Figure 1.1). The electron-light layers of myelin represent the middle of the cell surface membrane (Figure 1.1). In humans, myelination starts in the spinal cord during the second half of the gestational period, and the peak of myelin formation in the CNS occurs during the first year after birth, although myelination of some areas may not actually finish until the age of 20 (Baumann and Pham-Dinh, 2001).

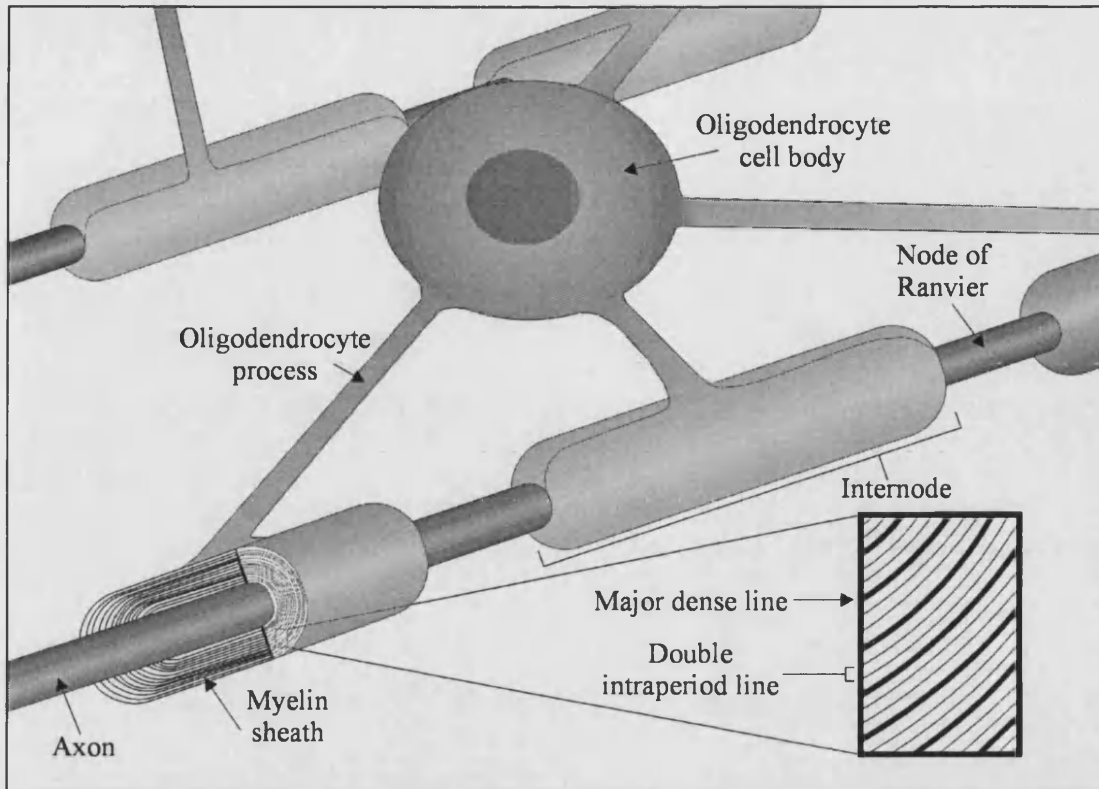


Figure 1.1. Myelin structure in the central nervous system. Axons are encircled by myelin sheath, which is formed from cytoplasmic projections of oligodendrocytes. A magnified view of a cross-section through the myelin sheath is shown. Adapted from (Baumann and Pham-Dinh, 2001).

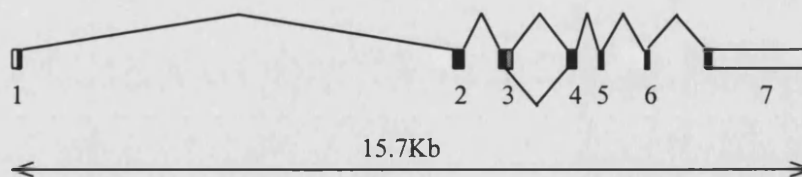


Figure 1.2. Genomic structure of *PLP1*, adapted from the Ensembl genome browser. Coding regions are shaded, exon 3B (*PLP1*-specific) is shaded grey and the rest of the coding region (present in both *PLP1* and *DM20*) is shaded black. 5' and 3' untranslated regions are white. Exons are numbered underneath, and splicing patterns are shown.

1.3. Proteolipid protein and DM20

PMD/SPG2 is caused by mutations involving the proteolipid protein (*PLP1*) gene. *PLP1* is located at Xq22, covers 15.7 kilobases of genomic sequence and has 7 exons, coding for a 276 amino acid, 30kDa protein, known as proteolipid protein 1 (PLP1) (Diehl *et al.*, 1986) (Figure 1.2). *PLP1* also encodes an alternatively spliced isoform, DM20, which is a 241 amino acid, 25kDa protein. DM20 is generated by activation of a cryptic splice site within the third exon that removes the second half of exon 3, termed exon 3B, and causes an in-frame deletion of 35 codons (Figure 1.2). PLP1/DM20 are very hydrophobic integral membrane proteins predicted to have 4 trans-membrane helices, linked together by one intracellular and two extra-cellular loops (Weimbs and Stoffel, 1992) (Figure 1.3). The N- and C- termini are located in the cytoplasm and the 35 amino acids that are absent in DM20 are within the intracellular loop (Figure 1.3). An additional alternatively spliced exon has been found in mice, where an additional exon (exon 1.1) is located in intron 1 (Bongarzone *et al.*, 1999). This exon contains an alternative translational start site which produces PLP1/DM20 proteins containing an extra 12 amino acids at the N-terminus (Bongarzone *et al.*, 1999). The presence of this short leader sequence alters the targeting of the PLP1/DM20 proteins within the cell, and these isoforms are located within the oligodendrocyte cell bodies, associated with trafficking vesicles and endosomes, and not within the myelin sheath, where PLP1/DM20 proteins are usually localised (Bongarzone *et al.*, 1999; Bongarzone *et al.*, 2001). These soma-restricted isoforms of PLP1 and DM20 have been found to be expressed in neurons during mouse development, at higher levels than found in oligodendrocytes (Jacobs *et al.*, 2003). A similar additional exon has not been identified in humans. Together, PLP1 and DM20 proteins are the most abundant proteins in CNS myelin, making up 50% of the total protein content (Baumann and Pham-Dinh, 2001). PLP1 and DM20 also have some post-translational modification in the mature proteins. The N-terminal

methionine residue is removed post-translationally (Milner *et al.*, 1985). There are two disulphide bridges between two pairs of cysteine residues in the large extracellular loop; and some of the cysteine residues on the cytoplasmic face of the protein are acylated (Figure 1.3) (Weimbs and Stoffel, 1992). PLP1 has up to six fatty acids covalently attached to cysteine residues, mostly palmitic acid, and DM20 has four, as two of the acylated cysteine residues are within the PLP1-specific section of the protein, and this difference in acylation may confer different properties on the two isoforms (Weimbs and Stoffel, 1992; Bizzozero *et al.*, 2002) (Figure 1.3). The structural model for PLP1 described in Figure 1.3 is frequently shown in publications, but some slightly different potential topologies have also been suggested. These include having the three acylated residues in the large intra-cytoplasmic loop being inserted into an adjacent cell membrane instead of the lipid bilayer in which the PLP1 protein is located, thereby stabilising the major dense line of myelin (Weimbs and Stoffel, 1992; Sporkel *et al.*, 2002). This uncertainty about the structure of PLP1/DM20 has not yet been resolved by the publication of a structure based on X-ray crystallographic or nuclear magnetic resonance data, which may be partly due to the inherent difficulties in purifying and crystallising these extremely hydrophobic proteins.

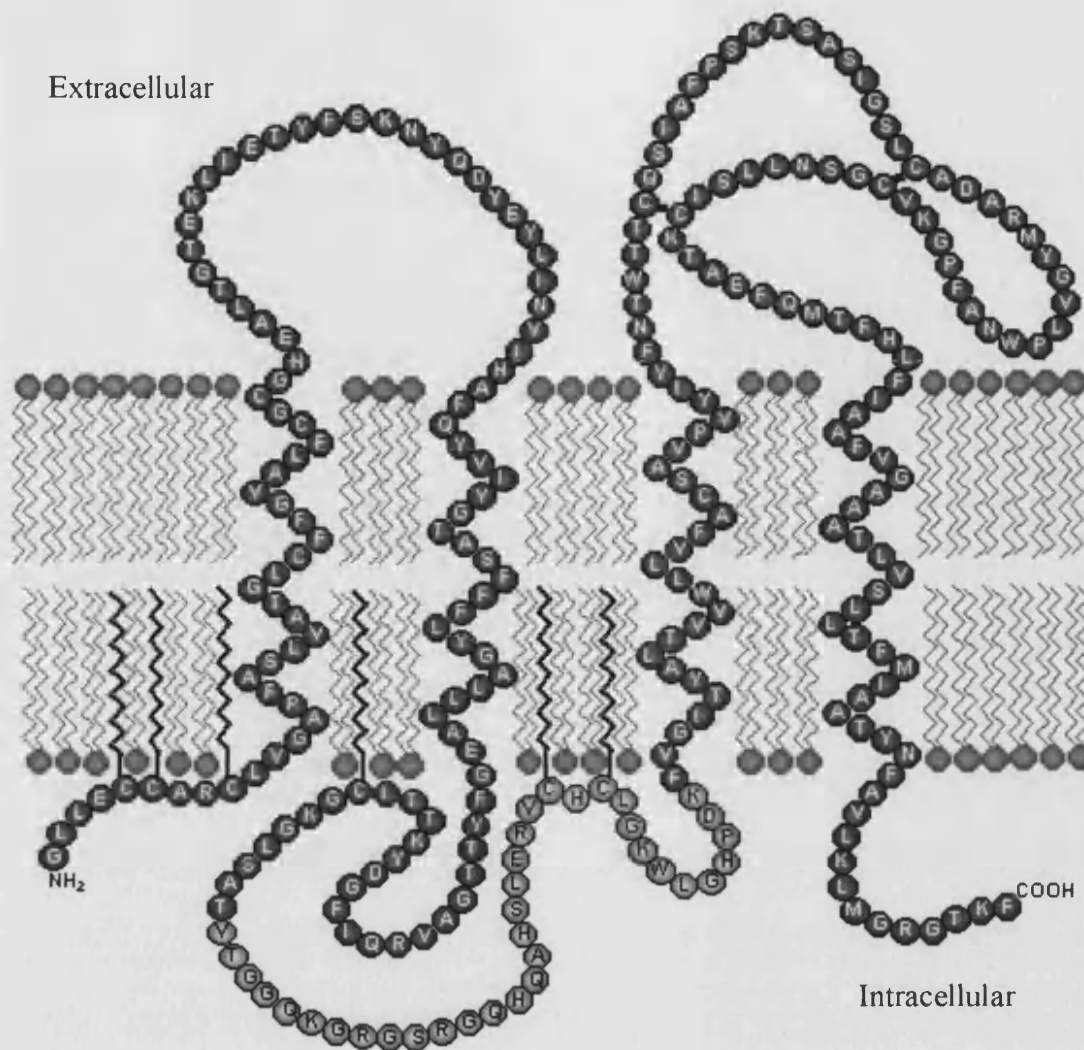


Figure 1.3. Diagram of possible PLP1 structure, showing the position of the protein within the cell surface membrane. The 35 amino acids specific to PLP1 within the intracellular loop are shaded in lighter grey. The two disulphide bonds in the larger extracellular loop are shown as black lines, and the fatty acid groups covalently bonded to cysteine residues in the intracellular parts of PLP1 are shown as zigzags inserted in the phospholipid bilayer. Adapted from (Weimbs and Stoffel, 1992; Garbern *et al.*, 1999; Greer and Lees, 2002).

1.3.1. Evolutionary conservation of PLP1/DM20

PLP1/DM20 is highly conserved between species, with human, mouse and rat having identical amino acid sequences, the dog and rabbit proteins each have only one amino acid change from the human protein sequence and the bovine sequence differs by only two amino acids (Milner *et al.*, 1985; Diehl *et al.*, 1986; Wight and Dobretsova, 2004). PLP1/DM20 is part of a family of proteins found in most vertebrates that also includes the M6a and M6b proteins in mammals, which show 56% and 46% identity with DM20 protein sequence (Yan *et al.*, 1993). The functions of M6a and M6b are unknown, but in mice M6a is neuronal-specific and M6b is expressed in neurons and myelin (Yan *et al.*, 1993). Proteins similar to DM-20 (DM α , DM β and DM γ) are present in cartilaginous fish, but the PLP1-specific segment is not found in these proteins (Kitagawa *et al.*, 1993; Yoshida and Colman, 1996). Invertebrates, such as *Drosophila melanogaster*, have also been found to have genes related to DM20, but without PLP1-specific segment similarity (Stecca *et al.*, 2000). It seems likely that PLP1 is a more recent evolutionary innovation than DM20, and emerged after insertion of the PLP1-specific segment into the DM20 ancestral gene, at the same time as the evolution of amphibians, which do express PLP1 (Gow, 1997).

1.3.2. *PLP1* mutations

Several different types of mutations involving *PLP1* can cause PMD and SPG2. *PLP1* was associated with PMD when a mouse model for the condition, jimpy, was found to have a splicing defect which led to a 74bp deletion in *PLP1* mRNA and was due to a mutation in the 3' splice acceptor site of intron 4 (Nave *et al.*, 1987). Point mutations within the *PLP1* gene were subsequently found in the genomic sequence of patients with PMD, and *PLP1* mutations were also found to cause the milder allelic disorder SPG2 (Gencic *et al.*, 1989; Hudson *et al.*, 1989; Saugier-Veber *et al.*, 1994). Other genetic mechanisms frequently cause PMD, and account for at least half of all

cases. An individual with a cytogenetically visible duplication of Xq13-q22 including *PLP1* was described with the symptoms of PMD, along with other abnormalities (Cremers *et al.*, 1987). It was later recognised that several individuals with PMD also had increased dosage of *PLP1* caused by whole gene duplications, and this is the most common genetic alteration found in PMD patients (Ellis and Malcolm, 1994; Woodward *et al.*, 1998; Sistermans *et al.*, 1998; Mimault *et al.*, 1999). Deletions of the whole *PLP1* gene, although rare, have also been reported in a few PMD patients (Raskind *et al.*, 1991; Inoue *et al.*, 2002). Other functionally null alleles, resulting from truncating point mutations very near the start of the gene, have also been described (Sistermans *et al.*, 1996).

1.3.3. Function of PLP1 and DM20

The function of these proteins is not fully understood, but they seem to play a structural role in the setting up and maintenance of the myelin sheath and formation of the intraperiod line, but may also have additional functions (Griffiths *et al.*, 1998a). There are several different animal models of PMD/SPG2, which have led to many insights into the function of *PLP1*. *PLP1/DM20* are predominantly expressed in CNS oligodendrocytes, but are also expressed in Schwann cells, and at a low level in other tissues, including skin, heart, foetal thymus and spleen (Campagnoni *et al.*, 1992; Carango *et al.*, 1995; Pribyl *et al.*, 1996). *DM20* expression starts before the onset of myelination and declines after the early stages of myelination, when *PLP1* production increases and this becomes the major isoform in oligodendrocytes. *PLP1/DM20* proteins are not necessary for the formation of myelin, as knockout mice with no *PLP1/DM20* expressed are able to form compacted myelin (Klugmann *et al.*, 1997). However, in null mice some ultrastructural changes in myelin structure are visible with electron microscopy, but the reported abnormalities vary, which may be due to differences in the genetic background of the null mutant (Boison and Stoffel,

1994;Klugmann *et al.*, 1997). PLP1/DM20 proteins are both needed for the maintenance of axonal integrity, as knockout mice and PMD patients with null mutations show signs of axonal damage (Griffiths *et al.*, 1998b;Garbern *et al.*, 1999;Garbern *et al.*, 2002). PLP1 and DM20 have been shown to associate during intracellular transport as well as within the myelin sheath and may interact to form an oligomeric complex (Gow and Lazzarini, 1996;McLaughlin *et al.*, 2002). Also, PLP1 (but not DM20) has been found to interact with α_v -integrin in oligodendrocytes, indicating that this isoform may play a role in integrin receptor signalling (Gudz *et al.*, 2002). PLP1 and DM20 have different functions, as mutations that only affect PLP1 still cause symptoms, albeit generally milder, in humans, and mice models that only express DM20 have some myelin abnormalities (Stecca *et al.*, 2000;Sporkel *et al.*, 2002). PLP1 or DM20 transgenes individually were not able to rescue the phenotype of a naturally occurring *Plp1* mutant mouse strain, but when both isoforms were present together, some myelination was seen, suggesting that the two isoforms do have distinct functions, or that some interaction between the two proteins is necessary for normal myelination (Nadon *et al.*, 1994).

1.3.4. Animal models

Naturally occurring mouse mutants have been described, such as the *jimpy*, *rumpshaker* and *myelin synthesis deficient* mice (Nave *et al.*, 1986;Gencic and Hudson, 1990;Schneider *et al.*, 1992). Transgenic mice have also been produced, including *PLP1* knockout mice and mice with increased dosage of *PLP1* (Boison and Stoffel, 1994;Readhead *et al.*, 1994;Kagawa *et al.*, 1994;Klugmann *et al.*, 1997). Other naturally occurring *PLP1* mutants are the *myelin deficient* rat, the dog mutant *shaking pup* and the *paralytic tremor* rabbit (Boison and Stoffel, 1989;Nadon *et al.*, 1990;Tosic M *et al.*, 1994). The severity of the phenotype of these animal models varies with different mutations, as is seen with the human disease, and the phenotype

has also been demonstrated to be dependent on the genetic background in some cases (Yool *et al.*, 2001). Some of the mutations are identical to those described in human patients, and show similar patterns of severity (Yamamoto *et al.*, 1998; Aoyagi *et al.*, 1999). In mice with increased dosage of *PLP1* transgenes, increasing severity of symptoms correlating with increased transgene copy number has been reported (Kagawa *et al.*, 1994; Inoue *et al.*, 1996b; Anderson *et al.*, 1999).

1.3.5. Gene duplication

The most common PMD mutation in *PLP1*, found in 60-70% of cases, is a duplication involving the whole *PLP1* gene (Ellis and Malcolm, 1994; Sistermans *et al.*, 1998). Typically the duplication involving *PLP1* is sub-microscopic and has a size range from 100kb to 4.6Mb (Hobson *et al.*, 2003). *PLP1* is not the only gene present within the duplicated interval, as there are several other genes in the vicinity which may or may not be duplicated depending on the duplication size (Figure 1.4). It is not clear if the size of the duplication is correlated with the severity of disease, but patients with a duplication frequently have a classical PMD phenotype (Sistermans *et al.*, 1998; Inoue *et al.*, 1999). In PMD, the duplication breakpoints have not been found to be similar between patients and are at different positions for almost every family studied (Woodward *et al.*, 1998; Inoue *et al.*, 1999; Hobson *et al.*, 2003). The duplications usually appear to be tandem in orientation and in most cases arise by an intrachromosomal event that occurs during male meiosis, with mothers of affected males almost always being carriers (Inoue *et al.*, 1996; Woodward *et al.*, 1998; Sistermans *et al.*, 1998; Inoue *et al.*, 1999; Mimault *et al.*, 1999). Involvement of duplication of a perfectly normal gene shows that gene dosage is an important mechanism in the pathogenesis of PMD. A few PMD patients have been found to carry (or are suspected to have) a triplication of *PLP1* (Ellis and Malcolm, 1994; Woodward *et al.*, 1998; Wolf *et al.*, in preparation). These individuals with three

times the normal dosage of *PLP1* are particularly severely affected, and have had the early-onset, connatal form of the disease, including seizures and death often within the first year of life (Wolf *et al.*, in preparation). This parallels the correlation of greater disease severity with increased transgene copy number that occurs in transgenic mouse models (Kagawa *et al.*, 1994; Inoue *et al.*, 1996b; Anderson *et al.*, 1999).

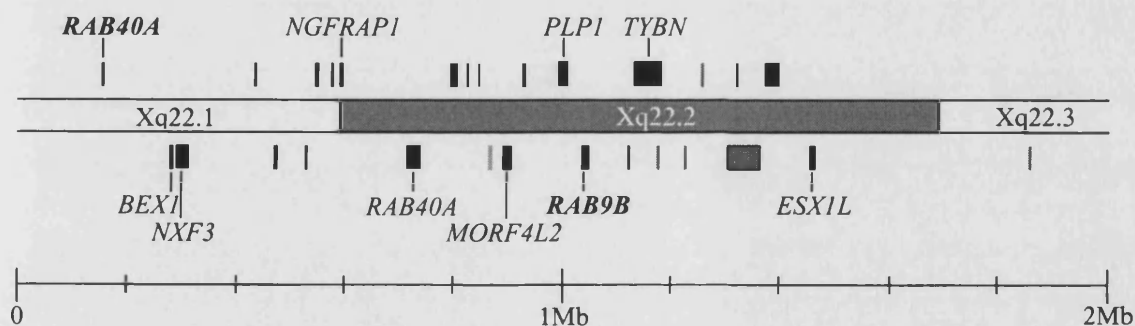


Figure 1.4. Genomic context of the *PLP1* gene. 1Mb either side of *PLP1* is shown, with the approximate boundaries of the chromosomal bands in the region also shown. Genes in the region are shown above or below the chromosome, those above are transcribed from left to right (centromere to telomere), and those underneath the chromosome are transcribed from the opposite strand. Genes with supporting evidence (from EST, cDNA, and protein sequences) are shaded in black, predicted genes just based on nucleotide sequence are dark grey, and pseudogenes are shaded in lighter grey. Adapted from the Ensembl genome browser, release 19.34a1.

1.3.6. Atypical duplications

A small subset of duplications including *PLP1* have been shown to move to a different chromosomal location. In three cases the second copy of *PLP1* has moved to elsewhere on the X chromosome, with two insertions into Xp and one insertion into Xq26 (Hodes *et al.*, 2000). Another case has been found where *PLP1* has been duplicated and inserted on Yq (unpublished data). The mechanism behind these atypical duplications is unknown and must involve at least three chromosomal breakages. One of the Xp duplications also has an inversion of a large section of the X chromosome between the two copies of *PLP1* associated with it (Hodes *et al.*, 2000). The Xq26 insertion has been demonstrated to be associated with mosaicism for a large deletion of the long arm of the X chromosome in a female carrier of the duplication (Woodward *et al.*, 1998; Woodward *et al.*, 2003). In this family, the affected male has inherited the duplication from his mother, and the duplication can be seen by FISH on one of her chromosomes in approximately half of her metaphase spreads in lymphoblastoid cells. Further examination of the mother's cells showed that a proportion of her cells from blood carried a deletion of one of her X chromosomes, so that in these cells there was one normal X chromosome with one copy of *PLP1* visible on metaphase, and an X chromosome carrying a deletion of most of the long arm, including the *PLP1* locus. Interphase FISH results are confusing with respect to the mother due to her mosaicism and quantitative PCR also gave results for the mother that were consistent with her not being a carrier (Woodward *et al.*, 1998). This family had two unusual rearrangements involving the X, an atypical duplication of *PLP1* on Xq26, which appeared to be *de novo* in the mother, and was subsequently shown to have occurred on her paternal X chromosome, which she then passed on to her affected son, and a large deletion of the long arm of the X that had contained the duplication (Woodward *et al.*, 2003). Another case has been reported in the literature where an atypical *PLP1* duplication

was associated with other cytogenetic rearrangements, namely a pericentromeric inversion of the X chromosome (Saito-Ohara *et al.*, 2002). This patient exhibited symptoms of Duchenne muscular dystrophy, and also had profound mental retardation, nystagmus and reduced white matter was found at post mortem examination (Saito-Ohara *et al.*, 2002). Two copies of *PLP1* could be visualised by FISH on the abnormal X chromosome at either end of the inversion, the normal copy at Xq22.2, and the other at Xp21 (Saito-Ohara *et al.*, 2002). The inversion breakpoint within Xq21 was within the *DMD* gene, leading to a deletion of exons 43-60, and within Xq22 a breakpoint was sequenced and located within the promoter region of the more proximal copy of the *RAB40A* gene (Figure 1.4) (Saito-Ohara *et al.*, 2002). Disruption of the expression of *RAB40A* may also have contributed to the severe phenotype seen in this patient along with the *DMD* and *PLP1* abnormalities (see section 3.5.2.1.) (Saito-Ohara *et al.*, 2002).

In addition to these cases where an atypical duplication of *PLP1* is associated with PMD, one individual has been reported with an atypical duplication of *PLP1* that did not result in a PMD phenotype. A male with PMD was demonstrated to have an *Alu-Alu* recombination-mediated deletion of the *PLP1* gene, and examination of other members of the family revealed a translocation of *PLP1* to the subtelomeric region of the long arm of chromosome 19 (Inoue *et al.*, 2002). The mother of the affected boy was a carrier of the translocation and presented in the third decade of life with spasticity, gradual mental deterioration and abnormal white matter upon MRI analysis (Inoue *et al.*, 2002). She was found to carry the deletion on one X chromosome and the *PLP1* translocation to 19q (Inoue *et al.*, 2002). The unaffected brother of the proband had inherited the normal X chromosome but also the chromosome 19 that harboured the translocation, so he carried two copies of *PLP1* (Inoue *et al.*, 2002). In this family it seems most likely that the translocated copy of *PLP1* is either not

expressed or is non-functional, possibly due to sequence changes associated with the rearrangement or position effects relating to its new location.

1.3.7. Point mutations and *PLP1* deletions

Point mutations in *PLP1* including single nucleotide changes, small insertions and deletions, can be found in 15-20% of cases. These are usually missense, nonsense and frameshift mutations that generate abnormal proteins. Most changes are within the coding region of *PLP1*, but non-coding regions can also be important locations for mutations affecting *PLP1* splicing (Hobson *et al.*, 2000). Different point mutations in *PLP1* cause dysmyelinating disease throughout the range of the PMD/SPG2 phenotype. In some cases point mutations cause a relatively mild SPG2-type phenotype (Hodes *et al.*, 1997;Cailloux *et al.*, 2000). This is frequently the case when the point mutation falls within the alternatively spliced exon 3B, which would only affect the PLP1 isoform, leaving DM20 unaffected (Hodes *et al.*, 1997;Cailloux *et al.*, 2000). Other point mutations can lead to a classical PMD phenotype similar to what is typically seen for duplication patients. In other affected individuals, however, point mutations can give rise to the particularly severe congenital phenotype, especially when they change an amino acid that is highly conserved in PLP1/DM20 and related proteins (Cailloux *et al.*, 2000).

In a very small proportion of cases (1-2%) a complete deletion or a null mutation of *PLP1* is found (Raskind *et al.*, 1991;Inoue *et al.*, 2002). Functionally null point mutations have been reported, including missense changes in the initiation codon, which result in a similar phenotype to the deletion (Sistermans *et al.*, 1996;Garbern *et al.*, 1997;Mimault *et al.*, 1999). These patients have mild symptoms, more like SPG2, and some have been found to also show peripheral nerve demyelination (Garbern *et*

al., 1997). This shows that PLP1/DM20 must have a function in the peripheral nervous system as well as the major functions in the CNS.

The remaining 10-15% of patients with a clinical diagnosis of PMD do not have a defined molecular defect involving *PLP1*, with the most likely explanations being either mutations in regulatory sequences or involvement of another locus in these cases (Sisttermans *et al.*, 1998). Mutations in an autosomal gene, *GJA12*, which encodes the gap junction protein $\alpha 12$ (connexin 46.6) have recently been reported in patients with autosomal recessive Pelizaeus-Merzbacher-like disease (PMLD) (Uhlenberg *et al.*, 2004). It is likely that there is at least one more autosomal locus involved in PMD/PMLD, as some individuals with PMD symptoms have been shown not to have mutations involving either *PLP1* or *GJA12* (Uhlenberg *et al.*, 2004).

1.3.8. Disease mechanisms

The mechanisms by which *PLP1* mutations cause PMD/SPG2 are not yet fully understood. It is thought that different types of mechanisms may be involved in the different types of mutations – duplications, point mutations and deletion/null mutations.

1.3.8.1. Duplications

Duplications of the *PLP1* gene, and subsequent over-expression of *PLP1*, will lead to the accumulation of excess protein within the cell (Garbern *et al.*, 1999). Overexpressed PLP1 is missorted within the cell, which has additional effects on oligodendrocyte function (Simons *et al.*, 2002). In cultured oligodendrocytes, PLP1 was found to accumulate in the late endosomal/lysosomal compartment along with cholesterol, instead of being transported to the myelin sheath from the Golgi apparatus (Simons *et al.*, 2002). It is thought that PLP1 may be associated with other

myelin membrane components, such as cholesterol and glycosphingolipids, in lipid rafts which transport their components to the developing myelin sheath (Simons *et al.*, 2000; Simons *et al.*, 2002). Accumulation of PLP1 in an abnormal cellular location may also lead to other membrane components being sequestered, as has been found for cholesterol, and leads to reduced levels of these components in the myelin membrane. This may cause some of the defects in myelination seen in PMD, and may also lead to other myelin proteins being abnormally located in the oligodendrocyte, further affecting function (Simons *et al.*, 2002; Vaurs-Barriere *et al.*, 2003).

1.3.8.2. Point mutations

Point mutations that cause a severe phenotype may do so by a gain of function mechanism, possibly by the misfolded protein not being transported to the cell surface membrane and instead accumulating within the cell, eventually triggering apoptosis (Gow *et al.*, 1998; Garbern *et al.*, 1999). The unfolded protein response pathway has been found to be activated in oligodendrocytes expressing a mutated *PLP1* gene, and may be important in triggering apoptosis in PMD oligodendrocytes (Southwood *et al.*, 2002). Point mutations can affect the properties of the two isoforms differently, which could have an effect on disease severity via differences in trafficking of the two mutated isoforms (Gow and Lazzarini, 1996). Most point mutations in *PLP1* lead to the accumulation of PLP1 within the endoplasmic reticulum (ER) (Gow and Lazzarini, 1996; Tosic *et al.*, 1997; Southwood and Gow, 2001). Only mutations which cause a severe phenotype lead to the sequestering of mutant DM20 within the ER, as in milder forms of the disease mutant DM20 is transported to the cell surface membrane (Gow and Lazzarini, 1996; Tosic *et al.*, 1997; Southwood and Gow, 2001). Missense mutations with a milder phenotype often alter amino acid residues that are less highly conserved in PLP1 and related proteins, and these mutations may lead to a

less severe defect in protein folding and trafficking, which has fewer adverse consequences for the cell, with perhaps some myelination being possible (Gow and Lazzarini, 1996; Gow *et al.*, 1998).

1.3.8.3. Deletions/null mutations

Null mutations of *PLP1*, either rare whole gene deletions or truncating point mutations at the 5' end of the gene, generally result in relatively mild symptoms, and the milder phenotype is presumably a result of the loss of function of the proteins (Garbern *et al.*, 1999).

1.3.8.4. Affected females

As PMD is an X-linked recessive disorder, the majority of affected individuals are males, with females that carry a mutation usually being non-symptomatic carriers. There are a small number of females affected, however. X-inactivation studies have shown that in female carriers of the duplication, the X chromosome carrying the duplication is preferentially inactivated in lymphocytes (Woodward *et al.*, 2000). It has been shown that females with point mutations usually have random X-inactivation (Woodward *et al.*, 2000). Occasionally heterozygous females for point mutations and null alleles have been found to manifest signs of PMD with adult onset leukodystrophies (Hodes *et al.*, 1995; Nance *et al.*, 1996; Hodes *et al.*, 1997). These females are generally from families with a mutation that produces a mild phenotype in males, whereas females that carry a severe *PLP1* mutation are usually phenotypically normal. Assuming that females with point mutations have random X inactivation in their oligodendrocytes as well as in blood, those cells expressing a severe point mutation may have severe PLP1 trafficking problems or other defects and most likely die by apoptosis. The surviving oligodendrocytes would express the normal *PLP1* allele and go on to myelinate normally. In females carrying a milder *PLP1* point

mutation, where the mutation was not so immediately toxic or detrimental to oligodendrocyte function, a mixed population of oligodendrocytes may be able to survive for a while. However, eventually the defect in PLP1/DM20 may cause cell death or failure of myelination, leaving only patchy myelination within the CNS, and causing the later-onset symptoms seen in some female carriers of point mutations. There have been only two cases reported of females with symptoms of PMD that have been shown to have a duplication (Inoue *et al.*, 2001). The two girls had early onset neurological signs and were found to have random X-inactivation patterns, but showed improvement in their symptoms over time (Inoue *et al.*, 2001). This may be because oligodendrocytes expressing the duplicated genes undergo early apoptosis, leaving the cells expressing a normal dosage of *PLP1* able to compensate by myelinating unsheathed axons, leading to the observed clinical improvement (Inoue *et al.*, 2001).

1.4. Gene dosage

Changes in gene dosage are an increasingly significant mutational mechanism in many human genetic diseases. A number of genes within the human genome, and in other organisms, as well as *PLP1*, have been shown to be dosage sensitive. Haploinsufficiency has been recognised as an important factor in the pathogenesis of numerous genetic disorders, both involving single genes or in contiguous gene syndromes, and is generally more frequently associated with a phenotype than increased gene dosage has been. However, increased gene dosage is increasingly being discovered in a small, but growing number of conditions. A common human syndrome where increased gene dosage is a factor is trisomy 21 (Down syndrome), where increased dosage of several genes located on chromosome 21 causes a well-recognised phenotype. Increased dosage of a whole human chromosome is rarely compatible with life, with only two other autosomal full trisomies, trisomy 13 (Patau

syndrome) and trisomy 18 (Edwards syndrome) being viable following gestation and birth, and even then few will survive beyond one year of age. The other human trisomies that are seen in live births are copy number changes involving the sex chromosomes, which have a relatively mild phenotype. Changes in sex chromosome number are less detrimental for two main reasons; X-inactivation will silence the majority of genes on supernumerary X chromosomes, suppressing dosage imbalances, and the Y chromosome is very gene-poor, so very few genes will be present in increased dosage. Of the chromosomes that are involved in viable trisomies, these (13, 18, X and Y) have some of the lowest gene densities of all the human chromosomes, and chromosome 21 is the smallest autosome, so also contains relatively few genes compared to larger chromosomes (Hattori *et al.*, 2000; Venter *et al.*, 2001; Dunham *et al.*, 2004). As increased gene dosage of most chromosomes is not viable in humans, it is likely that most chromosomes harbour at least one gene that will cause failure of normal development when present in increased copy number, or that a combination of several genes from the same chromosome, all present in increased dosage, is severely harmful to development.

Most cases of altered gene dosage involve just one gene or a handful of dosage sensitive genes. Charcot-Marie-Tooth disease type 1A, an autosomal dominant peripheral neuropathy, can be caused by duplication of the *PMP22* gene (Lupski *et al.*, 1991; Raeymaekers *et al.*, 1991). *PMP22* is flanked by two highly homologous 24kb repeat sequences and misalignment of these repeats and recombination between them leads to duplication of a 1.5Mb region including *PMP22*. Most of the *de novo* *PMP22* duplications causing CMT1A are paternal in origin and occur as a result of interchromosomal recombination between the misaligned 24Kb repeated sequences (Lopes *et al.*, 1997). As well as being sensitive to an increase in gene dosage, a reduction in dosage of the functional *PMP22* gene caused by a gene deletion leads to

another peripheral neuropathy, hereditary neuropathy with liability to pressure palsies (HNPP) (Chance *et al.*, 1994). The HNPP deletion is also mediated by the flanking 24Kb repeats, but occurs less frequently and is generally maternal in origin (Lopes *et al.*, 1997). The rearrangement of maternal origin resulting in a deletion is an intrachromosomal event, and rarely *PMP22* duplications are found to be of maternal origin, but in these cases the duplication is also intrachromosomal in origin, in contrast to the more frequent interchromosomal paternal duplications (Lopes *et al.*, 1997; Lopes *et al.*, 1998).

Smith-Magenis syndrome (SMS) is a microdeletion syndrome with a characteristic physical and behavioural phenotype resulting from an interstitial deletion of 3.7Mb of part of the chromosome band 17p11.2 (Smith *et al.*, 1986; Potocki *et al.*, 2000). A milder phenotype has also been described associated with the reciprocal duplication of 17p11.2 (Potocki *et al.*, 2000). The reciprocal deletions and duplications of the SMS region are mediated by flanking highly homologous 200Kb repeats (Chen *et al.*, 1997). Triplication of approximately 2Mb of chromosome 4q including the α -synuclein gene has recently been reported in a family with autosomal dominant Parkinson's disease (Singleton *et al.*, 2003). 500kb tandem duplications within 10q24 have been associated with split hand-split foot malformation 3 (SHFM3) in several individuals, and the duplicated region contains genes involved in limb development (de Mollerat *et al.*, 2003). Duplication of 6p25 has been found in families with autosomal dominant anterior chamber eye defects including iris hypoplasia and glaucoma (Lehmann *et al.*, 2000; Lehmann *et al.*, 2002). 6p25 contains the *FOXC1* gene, in which point mutations have been found in individuals with glaucoma phenotypes, and haploinsufficiency of *FOXC1* has also been described in other individuals with glaucoma and other eye anomalies (Lehmann *et al.*, 2000; Nishimura *et al.*, 2001). Variations in drug metabolism can be caused by alterations in gene copy

number. The *CYP2D6* gene on chromosome 22, which produces a cytochrome P450 protein, the debrisoquine hydroxylase enzyme, has been found to have several extra copies present, in some cases as many as 12 on one allele, in the genome of some individuals who show ultra-rapid metabolism of many commonly used drugs which are metabolised by the *CYP2D6* gene product (Johansson *et al.*, 1993; Agundez *et al.*, 1995; Aklillu *et al.*, 1996). Deficiency of *CYP2D6* activity causes the recessive poor metaboliser phenotype, either by a deletion of the whole gene or due to inactivating point mutations in this gene (Gonzalez *et al.*, 1988; Gaedigk *et al.*, 1991).

1.4.1. Genomic disorders

Many of the conditions that are caused by alterations in gene dosage have been described as “genomic disorders” (Lupski, 1998). Features of the genome can mediate or predispose to sequence rearrangements. Genomic disorders are those where such a rearrangement leads to a pathogenic change, such as deletion or duplication of a dosage sensitive gene. Many disorders in which gene dosage is a substantial factor in pathogenesis have been recognised as genomic disorders. Most commonly the genomic features that mediate the disease-causing rearrangement are low copy repeats (LCRs), repeated sequences specific to the deletion and/or duplication-prone region, which are generally greater than 10Kb in length and exhibit >95% identity (Lupski, 1998; Shaw and Lupski, 2004) . Non-allelic homologous recombination (NAHR) between misaligned LCR sequences during crossing-over leads to rearrangement of the genomic region, giving rise to deletions, duplications and inversions. Genomic disorders frequently cause rearrangements that are submicroscopic, but larger abnormalities that are apparent at the chromosomal level can also occur.

1.4.1.1. Direct repeats

NAHR between directly orientated LCRs creates reciprocal recombinant crossover products, one with the DNA between the repeats deleted out, and the other carrying two copies of the intervening sequence (Figure 1.5). Some recognised genomic disorders where directly repeated LCRs are found include CMT1A/HNPP, SMS, as well as many other conditions. Velo-cardio-facial syndrome (VCFS), also known as DiGeorge syndrome, is commonly caused by a microdeletion of 3Mb from 22q11.2, and this common deleted region is flanked by two 200Kb repeats, LCR22, which contain both directly repeated and inverted sections (Edelmann *et al.*, 1999a). These repeats, of which there are several on chromosome 22, including a copy within the VCFS commonly deleted region may also mediate rarer DNA rearrangements (Edelmann *et al.*, 1999b). Cat-eye syndrome results from a supernumerary chromosome, inv dup(22), with the breakpoints located at one or more of the LCR22 sequences (McTaggart *et al.*, 1998). The reciprocal duplication of the VCFS deletion has now been detected in an increasing number of cases (Edelmann *et al.*, 1999b; Ensenauer *et al.*, 2003; Hassed *et al.*, 2004). The 22q11.2 microduplication results in a milder phenotype, which may be why it has been so rarely reported, and the microduplications are often larger in size than the 3Mb common VCFS/DiGeorge deletion, with the size range of microduplications ranging between 3Mb to approximately 6Mb (Ensenauer *et al.*, 2003). The endpoints of the 22q11.2 microduplication are associated with various LCR22 sequences, including the copies that are most frequently involved in the VCFS/DiGeorge microdeletion (Ensenauer *et al.*, 2003).

Prader-Willi syndrome (PWS) and Angelman syndrome (AS) are disorders with very distinct phenotypes that can result from the same 4Mb deletion of chromosome 15q11-q13 (Magenis *et al.*, 1990). This region contains imprinted genes, which are

either maternally or paternally expressed. Inheritance of a deletion from the mother results in no expression of the maternally expressed genes in the region, resulting in an AS phenotype. A paternally inherited deletion of 15q11-q13 will cause PWS, due to the absence of expression of paternally expressed genes from the maternally inherited normal chromosome 15. The 4Mb deletion is mediated by recombination between LCRs on chromosome 15, which each contain a copy of *HERC2*, a transcribed gene, and are up to 200Kb in size (Amos-Landgraf *et al.*, 1999; Ji *et al.*, 1999). There are several copies of this repeat unit on chromosomes 15 and 16, and there are also other recurrent rearrangements involving chromosome 15 that may be mediated by these LCRs (Amos-Landgraf *et al.*, 1999).

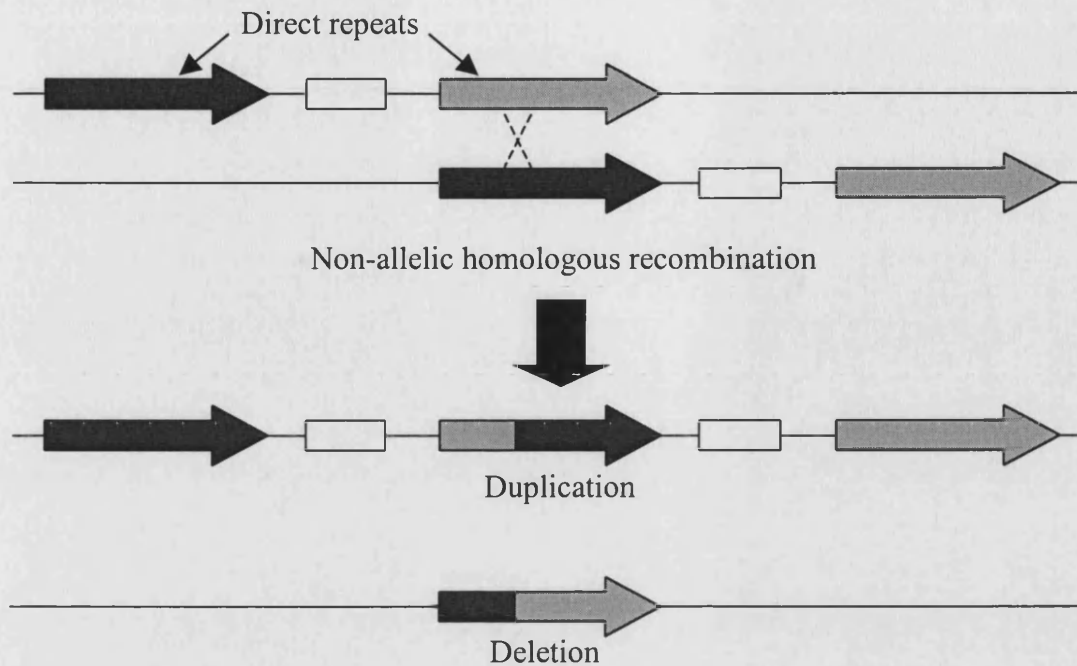


Figure 1.5. Non-allelic homologous recombination between directly repeated sequences can lead to both duplications and deletions. The direct repeats are shown as arrows filled in black and grey to distinguish between the two copies, and a gene between the two repeats (which may or may not be dosage sensitive) is represented by an open box. Misalignment of two homologous repeats, be it either interchromosomal (between homologous chromosomes) or intrachromosomal (between sister chromatids), can then lead to recombination between the two repeats. This can result in either a duplication of the intervening region, or a deletion of the region, which may lead to a phenotype through gene dosage effects. The recombination event creates a chimaeric repeat in both cases, consisting of part of the distal and part of the proximal repeat.

1.4.1.2. Inverted repeats

If the LCRs are inverted in orientation with respect to each other, recombination between the repeats leads to inversion of the sequence flanked by the inverted repeats (Figure 1.6). The factor VIII gene (located at Xq28) is mutated in cases of haemophilia A. Nearly half of severe cases have an inversion of around 600kb including part of the factor VIII gene, and the underlying molecular defect is gene disruption (Lakich *et al.*, 1993). The inversion takes place between two inverted transcribed repeats, known as the A gene (Lakich *et al.*, 1993). One of these is located in exon 22 of the factor VIII gene, with two other A gene sequences 5' to the gene, which are inverted with respect to the intronic A copy (Lakich *et al.*, 1993). Inversion of the factor VIII gene results in an abnormal transcript which contains exons 1-22 of the gene plus additional sequences, without the 3' end of the gene (Naylor *et al.*, 1993). As well as the inversion itself causing the pathogenic genomic change (as in haemophilia A), several polymorphic inversions have been discovered in human populations, which have been found to predispose to subsequent genomic rearrangements, generally in the offspring of individuals who are heterozygous for the inversion. One example of a common polymorphic genomic inversion is found near the *FLN1*/emerin region, also in Xq28. Two 11.3Kb inverted repeats flank 48Kb of sequence containing the *FLN1* and emerin genes, and an inversion of this whole region, mediated by the flanking repeats, is found in 18% of human chromosomes (Small *et al.*, 1997). Deletions of the emerin gene are sometimes found in patients with Emery-Dreifuss muscular dystrophy, and the polymorphic inversion may contribute to deletions in the region, especially in a parent heterozygous for the inversion (Small *et al.*, 1997; Small and Warren, 1998). Another example of this phenomenon has been studied in Williams-Beuren syndrome (WBS). Features of the WBS phenotype include characteristic facial features, vascular stenoses, growth retardation and specific cognitive defects (Osborne *et al.*, 2001). WBS is usually

caused by a heterozygous deletion of 1.55Mb of genomic DNA from 7q11.23, a region containing between 25-30 genes (Ewart *et al.*, 1993; Bayes *et al.*, 2003). The common deleted region is flanked by two LCRs, each about 400Kb in size (Robinson *et al.*, 1996; Osborne *et al.*, 2001). These LCRs contain various duplicated blocks of sequence, some of which are repeated in a direct orientation between the two duplications, and others which are inverted in orientation between the two repeats (Osborne *et al.*, 2001). The 1.55Mb deletion is mediated by NAHR between the directly orientated portions of the flanking LCRs, and an inversion polymorphism of between 1.79-2.56Mb including the WBS common deletion region is caused by NAHR between the inverted sections of the flanking repeats (Osborne *et al.*, 2001; Bayes *et al.*, 2003). The inversion polymorphism has been found in nearly one-third of progenitors who were found to have transmitted the affected chromosome and has also been found in the parents of individuals with atypical deletions and larger chromosome rearrangements involving the region (Osborne *et al.*, 2001; Bayes *et al.*, 2003).

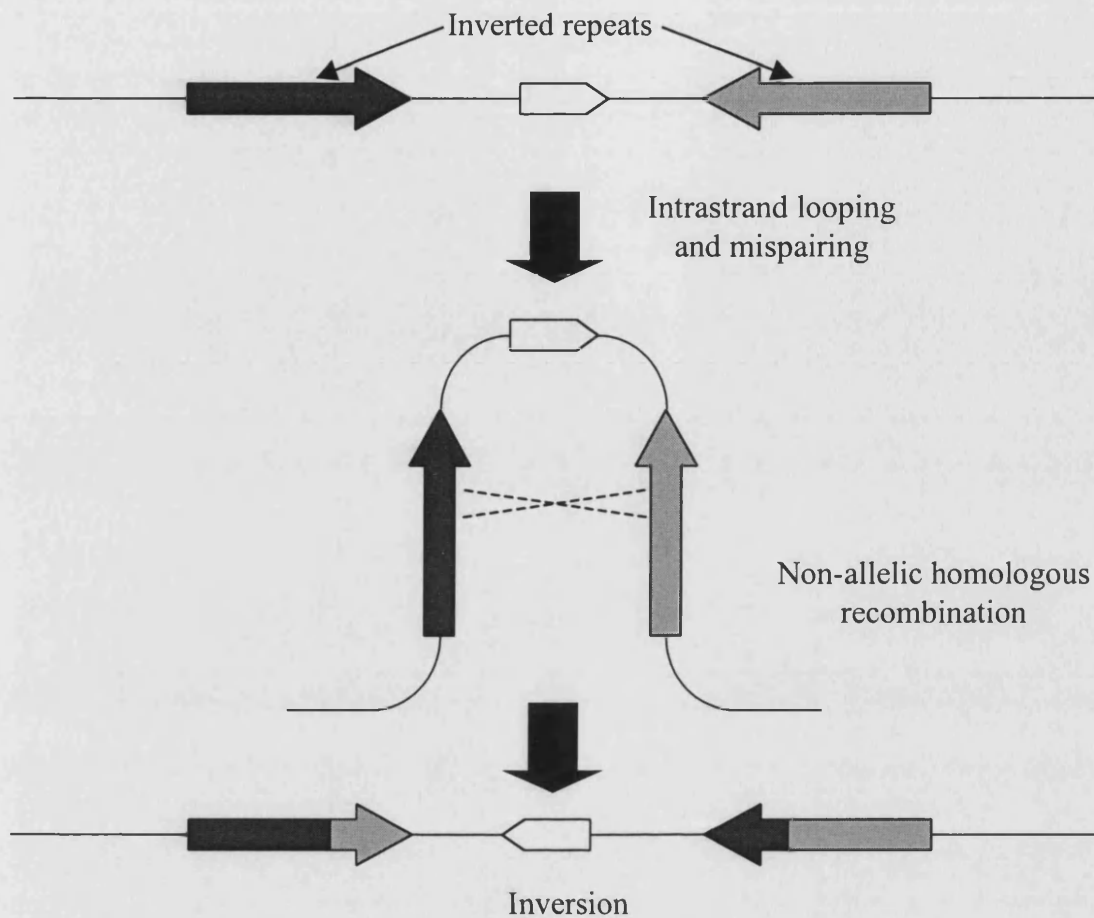


Figure 1.6. Recombination between inverted repeats leads to an inversion of sequences between the repeats, which may disrupt a gene or regulatory sequences, or cause a predisposition to other rearrangements.

1.4.1.3. Gross chromosomal rearrangements

Some large cytogenetically visible chromosomal rearrangements such as translocations have been found to be recurrent events, mediated by repeated regions at the breakpoints. Several recurrent chromosomal rearrangements involving chromosome 8 have been shown to be mediated by LCRs containing clusters of olfactory receptor (OR) genes (Giglio *et al.*, 2001; Giglio *et al.*, 2002). Two 400Kb repeats containing OR genes map to 8p23.1 and 8p23.2, and were found to be involved in the breakpoints of several abnormal chromosome rearrangements, such as inv dup(8p), +der(8)(8p23.1pter), del(8)(p23.1p23.2) and t(4;8)(p16;p23) translocation (Giglio *et al.*, 2001; Giglio *et al.*, 2002). Many of these chromosome 8 rearrangements also involve a polymorphic inversion of a 4.7Mb region flanked by the OR repeats, as this inversion is heterozygous in the parent transmitting the affected chromosome in many cases (Giglio *et al.*, 2001; Giglio *et al.*, 2002; Sugawara *et al.*, 2003). This inversion polymorphism has been reported to be present in the normal population at a frequency of about 26-27% (Giglio *et al.*, 2001; Sugawara *et al.*, 2003). In the case of the 4;8 translocation, the two OR clusters at 4p16 are also associated with a polymorphic inversion of the interstitial 6Mb of sequence, and mothers of translocation cases were found to be heterozygous for both the 8p23 and 4p16 inversions, which presumably predispose to this particular chromosomal rearrangement (Giglio *et al.*, 2002). Another genetic disorder, Kabuki syndrome, has recently been reported to be caused by a 3.5Mb duplication of 8p22-8p23.1, with the duplicated region flanked by OR-containing repeats (Milunsky and Huang, 2003). Additionally, two of the mothers of Kabuki syndrome cases were investigated by FISH and were found to be heterozygous for the inversion polymorphism in this region (Milunsky and Huang, 2003).

1.4.1.4. Partial gene deletions and duplications

Rearrangement of genomic DNA does not always involve a whole gene. Partial deletions and duplications of genes or just individual exons, leading to altered protein sequence, with frameshifts occurring in many cases, are also found in human genetic disease. A notable example of this is in Duchenne and Becker muscular dystrophies, where a deletion or duplication of one or more exons within the enormous dystrophin gene leads to a truncated or interstitially deleted protein being produced (Koenig *et al.*, 1987; Hu *et al.*, 1988). Rearrangement breakpoints within the dystrophin gene are scattered throughout the gene, but some hotspots for deletions and duplications have been reported in particular introns or regions of the gene (den Dunnen *et al.*, 1989; Koenig *et al.*, 1989; McNaughton *et al.*, 1998). Sequenced breakpoints have usually shown no homology to each other, although a minority of breakpoints have interspersed repeat elements such as *Alu* at both ends (McNaughton *et al.*, 1998). Incontinentia pigmenti, an X-linked dominant disorder, is caused by mutations in the *NEMO* (NF- κ B essential modulator) gene (Smahi *et al.*, 2000). 80% of new mutations in the *NEMO* gene are a deletion of exons 4-10, with the breakpoints of the deletion falling within highly similar 870bp MER67B repeats in intron 3 and 3' to exon 10, the final exon of the *NEMO* gene (Smahi *et al.*, 2000). Deletions and duplications of one or more exons within the *BRCA1* gene make a significant contribution to the mutational spectrum of this gene, accounting for between 10 and 30% of germline mutations within *BRCA1* in different populations; these rearrangements are often found to be due to illegitimate recombination between *Alu* elements (Petrij-Bosch *et al.*, 1997; Puget *et al.*, 1999; The BRCA1 Exon 13 Duplication Screening Group, 2000).

1.4.2. Gene dosage in cancer

Gene dosage is not only an important factor in inherited disease; it is also crucial for the development of many tumours. In many human cancers, loss or amplification of most parts of the genome have been found using various techniques to detect changes in gene dosage (Knuutila *et al.*, 2000). Genomic instability, including chromosome aneuploidy, translocations and changes in dosage of smaller regions, is present in many cancer cell genomes (reviewed in (Lengauer *et al.*, 1998; Balmain *et al.*, 2003)). The nature of genomic instability in cancers is still unclear, with uncertainty as to whether genomic instability is a primary factor in early carcinogenesis or just a secondary event in tumours that does not play a great role in cancer development (reviewed in Lengauer *et al.*, 1998; Marx, 2002; Balmain *et al.*, 2003; Pihan and Doxsey, 2003; Duesberg *et al.*, 2004). In general, regions that are recurrently amplified in particular tumours often contain oncogenes that promote the establishment and growth of tumours, such as *MDM2*, *MYC* and *CMYC* (Schwab *et al.*, 1983; Alitalo *et al.*, 1983; Oliner *et al.*, 1992, reviewed in Schwab, 1999). Parts of the genome that are recurrently deleted in cancer will frequently contain tumour suppressor genes, and loss of function due to deletion of both alleles, or deletion of one allele accompanied by a mutation or epigenetic silencing on the other allele, contributes to carcinogenesis (Balmain *et al.*, 2003). It is also becoming apparent that haploinsufficiency of some tumour suppressor genes, such as p27, is also an important factor during tumour development (Philipp-Staheli *et al.*, 2001; Balmain, 2002).

1.4.3. PMD – is it a genomic disorder?

In contrast to most “genomic disorders” described above, such as CMT1A, VCFS and SMS, in PMD patients the duplication containing *PLP1* varies considerably in size between individuals, and no LCRs have been found to date flanking the gene (Woodward *et al.*, 1998; Inoue *et al.*, 1999; Hobson *et al.*, 2003). However, other

features of the genomic sequence within Xq22.2 may be involved in the rearrangements of this region, which include tandem duplications, atypical transposed duplications and deletions (Inoue *et al.*, 2002). Determination of the sequence at and near the breakpoints in PMD patients may help to understand the mechanisms behind these rearrangements. The breakpoints that have been sequenced and reported so far mostly do not show large stretches of homology at the breakpoints, and non-homologous end joining (NHEJ) has been implicated in these rearrangements (Inoue *et al.*, 2002). Many of the breakpoints that have been sequenced have one end located in an interspersed repeat and the other end within unique sequence, and several duplication or deletion endpoints have been located in distal LCRs (see section 3.4.) (Inoue *et al.*, 2002; Hobson *et al.*, 2003; Iwaki *et al.*, 2003; Woodward *et al.*, in preparation). NHEJ is the most commonly used pathway for repair of double-strand breaks in multicellular eukaryotic organisms, and typically the two ends are joined at short regions of microhomology (1-4bp) between the two sequences (reviewed in Lieber *et al.*, 2003).

1.5. Detection of gene dosage

Many methods have been used to detect changes in *PLP1* gene dosage. Interphase FISH is routinely used to detect *PLP1* duplications in a diagnostic setting, and has been found to be reliable for male patients and female carriers, although time-consuming (Woodward *et al.*, 1998). Quantitative fluorescent multiplex PCR and comparative multiplex PCR have also been used to detect dosage changes in *PLP1*, but can give problems with detecting carrier status in females (Inoue *et al.*, 1996a; Wang *et al.*, 1997; Woodward *et al.*, 1998; Sistermans *et al.*, 1998). Southern blotting can also be used to detect dosage changes in *PLP1*, but is not a rapid technique and requires large amounts of DNA (Ellis and Malcolm, 1994). Analysis of RFLP alleles or other polymorphisms within or near *PLP1* can be used if the mother or patient is heterozygous for polymorphisms within the gene. Heterozygosity for polymorphic markers within the duplicated region in males could be used as evidence for duplications of the region, but as the duplication is usually intra-chromosomal in origin, marker analysis is generally uninformative (Inoue *et al.*, 1996a; Woodward *et al.*, 1998).

1.5.1. Real-time PCR

Real-time PCR is a method that can be used to measure DNA sequence copy number. A reaction is designed so that fluorescence increases in proportion with the amount of PCR product present in the reaction, and detection of the amount of fluorescence present throughout the reaction allows quantitation of the product while the PCR is still in the exponential phase (Ginzinger, 2002). Various methods can be used to produce the fluorescence; intercalating dyes can be used that only fluoresce when bound to double stranded DNA, or fluorescent probes that are specific to the PCR product can be used. Real-time PCR has been shown to be a rapid method of detecting changes in gene dosage in many cases, such as the *PMP22* gene (Aarskog

and Vedeler, 2000;Thiel *et al.*, 2003). Real-time quantitative PCR technology does not enable the PCR products to be identified by size, so multiplexing several different sets of primers, as is possible for many conventional quantitative PCR strategies, is limited by the different types of fluorophores available.

1.5.2. MLPA

One method that has been recently developed that can assay dosage at several loci simultaneously (at least 40), is multiplex ligation-dependent probe amplification (MLPA) (Schouten *et al.*, 2002). Each target sequence, which is between 50-70 nucleotides in length, is detected by a pair of probes, each hybridising to one half of the target. Genomic DNA is mixed with the various probes and allowed to hybridise. Following hybridisation, a ligation reaction is carried out, joining together any probes that are hybridised to adjacent sequences, then PCR is carried out using primers that are specific to tag sequences present at the ends of all the probes. Each probe pair produces a differently sized amplification product, so following electrophoresis each probe can be identified and quantified. MLPA has been used to assay gene dosage for a number of applications, such as screening *BRCA1* for copy number changes in exons and detection of aneuploidies (Hogervorst *et al.*, 2003;Slater *et al.*, 2003). An MLPA kit containing all exons of the *PLP1* gene has recently become commercially available from MRC-Holland (Wolf *et al.*, in preparation).

1.5.3. Comparative genomic hybridisation

Most methods for assaying gene dosage only look at the copy number of a limited number of target sequences, but some methods can be used to survey the whole genome for copy number changes. One such method is comparative genomic hybridisation (CGH), which was initially developed as a tool for looking at copy number changes in cancer, but can also be used to detect changes in gene dosage that

may cause genetic disease (Kallioniemi *et al.*, 1992). The principles of CGH are as follows: a test DNA sample and reference DNA sample are each labelled with a different fluorophore, and hybridised to a metaphase spread from a normal individual, along with unlabelled *Cot*-1 DNA. The amounts of each of the two samples relative to each other along all the chromosomes can then be quantified, giving a genome-wide assessment of changes in copy number at a resolution of 20Mb. A later modification of CGH, array-CGH, gave the technique a much higher resolution by hybridising the test and reference DNA to individual genomic clones spotted on a glass slide (Pinkel *et al.*, 1998). Using array-CGH, the resolution is brought down to the level of single genomic clones. Recently, a tiling path DNA microarray used for CGH has been developed, with complete coverage of the human genome, consisting of 32433 BAC clones (Ishkanian *et al.*, 2004). Lower resolution genome-wide arrays are also in use, some that can survey the whole human genome at a 1Mb level, using equally spaced clones, and many other smaller-scale arrays have been developed for specific regions (Vissers *et al.*, 2003; Fiegler *et al.*, 2003). An X-chromosome tiling path BAC array has been shown to detect increased dosage of 3 BAC clones surrounding and including the *PLP1* gene in a sample with a known *PLP1* duplication (Veltman *et al.*, 2004).

1.5.4. MAPH

Multiplex amplifiable probe hybridisation (MAPH) is a technique that can be used to assess sequence copy number in genomic DNA, and has been used as part of this project (Figure 1.7) (Armour *et al.*, 2000). The protocol involves hybridisation of short probes to genomic DNA, which are then recovered and amplified. The probes each consist of sequence complementary to a specific region of genomic DNA (150-450bp), which is flanked by 59bp of sequence common to all probes. Each probe is a different size, so they can be distinguished from each other by their gel mobility

following electrophoresis. After hybridisation to genomic DNA, and washing to ensure sequence-specific binding, probes are amplified using PCR, and the amount of PCR product obtained from the reaction should be directly proportional to the copy number of the sequence in the sample (Armour *et al.*, 2000). The MAPH technique has been shown to be able to detect duplications and deletions of exons of the *DMD* gene, various subtelomeric deletions using a subtelomeric set of probes, a deletion of the *TBX5* gene, and duplications and deletions of *PMP22* (Armour *et al.*, 2000; Akrami *et al.*, 2001; Sismani *et al.*, 2001; White *et al.*, 2002; Akrami *et al.*, 2003; White *et al.*, 2003).

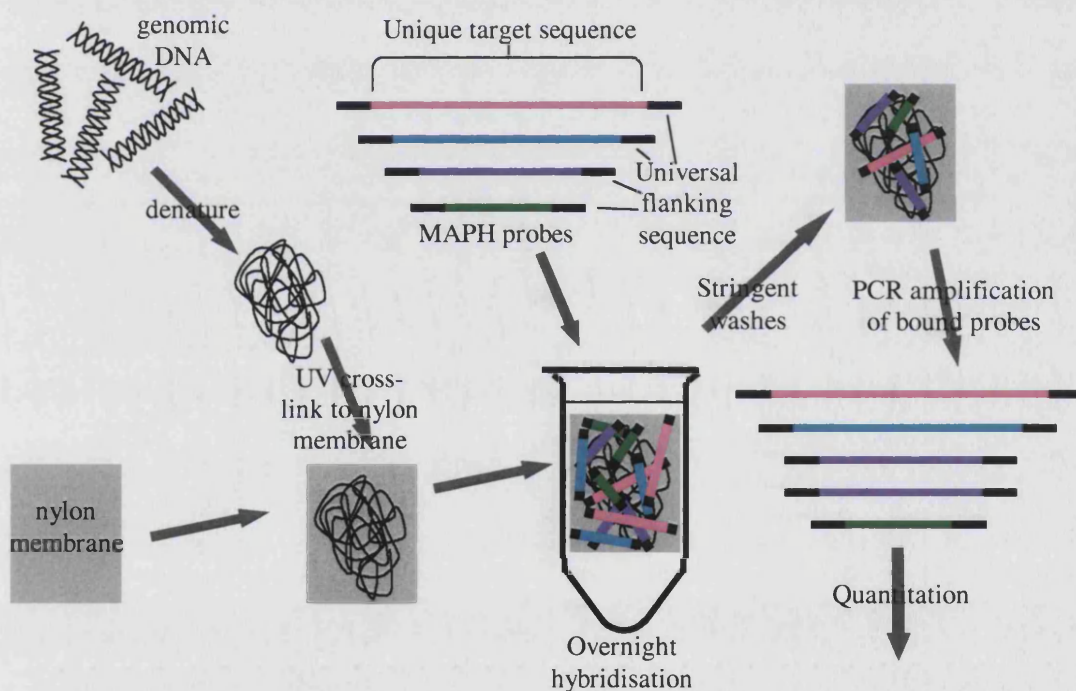


Figure 1.7. Diagram showing principles of MAPH. Adapted from Armour *et al.*, 2000.

1.5.4.1. MAPH and *PLP1* gene dosage

Many different laboratory techniques can be used to detect changes in *PLP1* gene dosage, with interphase FISH being a widely used diagnostic technique, which is used in our laboratory (Woodward *et al.*, 1998). However, alternative techniques may be preferable, by shortening the experimental procedure, requiring less time-consuming analysis, or by other improvements inherent to the methodology. Some techniques may require substantial investment in specialised equipment (Real-time PCR, array-CGH), or may not be amenable to multiplexing of different target loci (Real-time PCR, FISH). Recently developed techniques such as MAPH and MLPA use equipment that is already available to most laboratories, and are able to target several loci (at least 40 in each case) simultaneously. MAPH probes are easily produced by cloning into a specific plasmid vector followed by PCR using flanking vector primers, and MLPA probes are produced from phage M13-derived vectors (Armour *et al.*, 2000; Schouten *et al.*, 2002). Development of a MAPH probe set including *PLP1* probes and also probes from the surrounding region would enable detection of *PLP1* duplications and also detect the extent of the duplication in a single experiment.

1.6. SOX3 gene dosage and X-linked hypopituitarism

Hypopituitarism is characterised by the reduced secretion of one (isolated hypopituitarism), several (partial hypopituitarism) or all (panhypopituitarism) of the various different hormones secreted by the pituitary gland (see sections 1.6.2.1-4) (Pinzone, 2001). Anterior pituitary development is dependent on a complex genetic cascade of signalling molecules and transcription factors. Mutations have been reported in several transcription factors involved in pituitary development in individuals with varying degrees of hypopituitarism, including *HESX1*, *PROT1*, *POU1F1*, *LHX3* and *LHX4* (Pfaffle *et al.*, 1992;Radovick *et al.*, 1992;Dattani *et al.*, 1998;Wu *et al.*, 1998;Netchine *et al.*, 2000;Machinis *et al.*, 2001;Cohen and Radovick, 2002).

The inheritance of hypopituitarism may be autosomal recessive, dominant or X-linked. Recent advances have suggested that X-linked hypopituitarism may be due to abnormalities in gene dosage. One indication that gene dosage might be a factor in X-linked hypopituitarism came with two reports of interstitial duplications within the long arm of the X-chromosome segregating with X-linked hypopituitarism (Hol *et al.*, 2000;Solomon *et al.*, 2002). One report involved a family where a 13Mb duplication of Xq26.1-q27.3 was found in two brothers with X-linked hypopituitarism and spina bifida (Hol *et al.*, 2000). Another family with X-linked hypopituitarism which had previously been linked to Xq25-q26 was then also found to carry a 9Mb duplication of the chromosomal region Xq26.1-q27.3 (Lagerstrom-Fermer *et al.*, 1997;Solomon *et al.*, 2002). As both duplications were of a different size, and had different endpoints, the most likely explanation for the phenotype is that a dosage-sensitive gene located in the common duplicated region was responsible for the pituitary phenotype. Further evidence for this region being involved in pituitary disease came with the report of an expansion of a polyalanine tract (from 15 alanines to 26

alanines) within the *SOX3* gene in a family where the affected individuals showed mental retardation, facial dysmorphism and short stature (Laumonnier *et al.*, 2002).

1.6.1. *SOX3*

SOX3 (*SRY*-related HMG-box gene 3) is a transcription factor that is part of a family of 20 human proteins including the *SRY* gene that all contain a HMG domain (high mobility group), a DNA-binding motif of approximately 79 amino acids, with at least 50% similarity to the *SRY* HMG-box domain (Figure 1.8) (Stevanovic *et al.*, 1993; Schepers *et al.*, 2002). The HMG-box domain binds to DNA in the minor groove and causes bending of the target DNA when bound (Figure 1.8). *SOX3* is one of the B1 group of SOX proteins, along with *SOX1* and *SOX2*, all three of which show a high degree of sequence similarity throughout the protein sequence (Bowles *et al.*, 2000). Many of the other SOX proteins only show homology to *SOX3* within the HMG-box, and have very different N- and C-terminal sequences. The *SOX3* gene is located on the X chromosome, at the chromosomal position Xq27.1, and contains just one exon, 1.3Kb in length, coding for a 446 amino acid protein (Stevanovic *et al.*, 1993). Of all the SOX family of proteins, *SOX3* is the most similar to *SRY*, and it is likely that *SRY* evolved from an ancestral *SOX3* gene (Graves, 1998; Katoh and Miyata, 1999). *SOX3* is well conserved between species, and the mouse and human sequences are 97.2% identical (Stevanovic *et al.*, 1993). *SOX3* is expressed throughout the developing CNS, and in the genital ridge, with an expression pattern that largely overlaps that of *SOX1* and *SOX2*, in developing mouse embryos (Collignon *et al.*, 1996). *SOX1-3* are also involved in lens differentiation and development (Kamachi *et al.*, 1998). *SOX3* is able to bind an *SRY* consensus motif, AACAAAT, but has lower affinity than *SRY* or *SOX2* at binding this sequence (Collignon *et al.*, 1996). SOX-group proteins are thought to bind DNA in conjunction

with other DNA-binding transcription factors, such as POU homeodomain proteins, which gives greater target specificity (Kamachi *et al.*, 2000; Remenyi *et al.*, 2003).

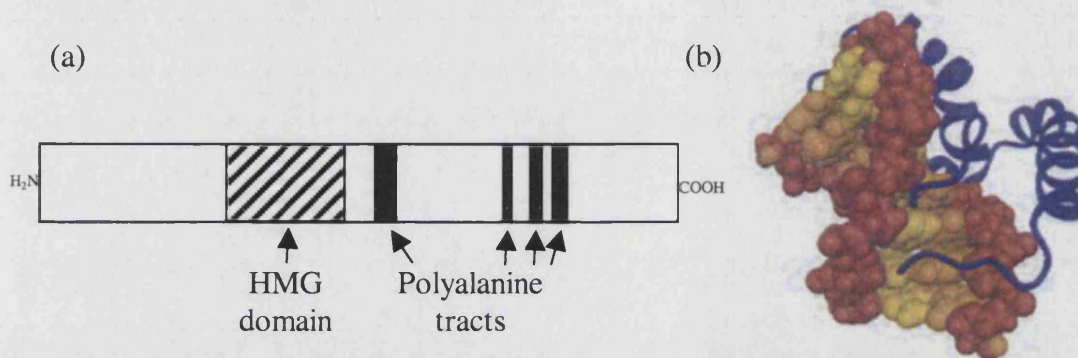


Figure 1.8. SOX3 protein structure. (a) shows the domains present in the SOX3 protein, the HMG-box DNA binding domain is shaded with diagonal lines and four polyalanine tracts are shaded black. (b) shows a three-dimensional representation of the SRY HMG-box domain (which is highly similar to the SOX3 HMG-box) bound to DNA, showing the way the DNA is bent by the bound HMG-box protein domain. SRY HMG-box protein is shown in purple, with the DNA backbone shaded red, and the bases in yellow and orange. From (Remenyi *et al.*, 2003).

1.6.2. Pituitary function

The pituitary gland is a pea-sized midline structure located under the hypothalamus in a bony cavity, the sella turcica, and is directly connected to the hypothalamus by the hypophyseal stalk, or infundibulum (Figure 1.9a) (Thapar *et al.*, 2001). In humans, the gland comprises two lobes, the anterior pituitary (adenohypophysis) and posterior pituitary (neurohypophysis), which are derived from different embryonic origins during development (Figure 1.9b) (Thapar *et al.*, 2001). The pituitary gland is the “master gland” of the endocrine system, and secretes many different hormones into the bloodstream, regulating numerous physiological processes involved in growth, metabolism and reproductive development.

1.6.2.1. Anterior pituitary gland

In the anterior pituitary, each hormone is synthesised and secreted by one of six specialised cell types. Somatotrophs secrete growth hormone (GH), also known as somatotrophin, which has numerous effects on metabolism and growth (Baumann, 2001). Thyrotroph cells secrete thyroid stimulating hormone (TSH), or thyrotrophin, which as the name suggests, stimulates the thyroid gland to secrete thyroid hormones (Cohen, 2001). Adrenocorticotrophic hormone (ACTH), or corticotrophin, is secreted from corticotroph cells in the pituitary as part of the stress response (Torpy and Jackson, 2001). It stimulates the adrenal gland to produce cortisol and other glucocorticoids (Torpy and Jackson, 2001). Lactotroph cells in the pituitary gland produce prolactin (PRL), a hormone that is involved in lactation and has some effects on the immune system (Thapar *et al.*, 2001). The two gonadotrophin hormones, luteinising hormone (LH) and follicle-stimulating hormone (FSH) are synthesized in gonadotroph cells in the pituitary (Thapar *et al.*, 2001; Bremner *et al.*, 2001). LH and FSH have effects in the ovary and testes, stimulating the production of the hormones oestrogen and testosterone, and are also crucial to the maturation of oocytes and

sperm production (Bremner *et al.*, 2001). TSH, LH and FSH are all dimer molecules, with all three hormones containing a common alpha subunit, which is non-covalently bound to a different beta subunit for each different hormone type (Cohen, 2001;Bremner *et al.*, 2001). Melanotroph cells within the human foetal pituitary gland secrete melanocyte stimulating hormone (MSH), which may be involved in regulation of energy homeostasis and body weight (Torpy and Jackson, 2001;MacNeil *et al.*, 2002). MSH and ACTH are both produced from the same precursor peptide, pro-opiomelanocortin, which is differentially cleaved to produce the two different hormones (Torpy and Jackson, 2001).

1.6.2.2. Posterior pituitary gland

The posterior lobe of the pituitary is composed of neural projections from the hypothalamus, and also secretes important hormones (Figure 1.9b) (Thapar *et al.*, 2001). Antidiuretic hormone (ADH), or vasopressin, is secreted by hypothalamic neuroendocrine neurons in the posterior pituitary, and increases the rate of water resorption by acting on the kidney (Robertson, 2001). ADH levels are regulated by osmoreceptor neurons in the hypothalamus, which sense plasma solute concentration and then stimulate the ADH-producing cells if plasma osmolarity is above a threshold level (Robertson, 2001). The other hormone secreted by the posterior pituitary is oxytocin, which again is released by neuroendocrine axonal projections from the hypothalamus into the posterior pituitary (Robertson, 2001). Oxytocin has its main effects in causing the uterus to contract during labour and in stimulating milk flow during lactation (Robertson, 2001).

1.6.2.3. Intermediate pituitary gland

In many species, including almost all mammals, but not in adult humans, there is a distinct third lobe of the mature pituitary between the anterior and posterior lobes, known as the intermediate lobe (Saland, 2001). The intermediate lobe is present during early development in humans, but involutes during development and is not present in adults (Saland, 2001). MSH is secreted from the intermediate lobe in those species in which this lobe is present (Saland, 2001).

1.6.2.4. The hypothalamic-pituitary axis

The hypothalamus secretes many hormones that control the release of pituitary hormones and the two are directly connected by hypothalamic-hypophyseal portal veins, so that factors secreted from the hypothalamus are carried directly to the pituitary gland (Figure 1.9b) (Thapar *et al.*, 2001). Growth hormone releasing hormone (GHRH) is released from the hypothalamus and stimulates the production and release of GH from the pituitary (Baumann, 2001). Somatostatin is also released by the hypothalamus and inhibits the release of GH (Baumann, 2001). Prolactin release is not stimulated by the hypothalamus, but factors including dopamine that are produced in the hypothalamus inhibit prolactin production (Katznelson and Klibanski, 2001). Thyrotrophin releasing hormone from the hypothalamus stimulates the release of TSH from the anterior pituitary (Cohen, 2001). Corticotrophin-releasing hormone, produced by the hypothalamus, stimulates ACTH production in the pituitary gland (Torpy and Jackson, 2001). ACTH production is also stimulated by the action of ADH/vasopressin, secreted from the posterior pituitary (Torpy and Jackson, 2001). LH and FSH production within the pituitary gland is controlled by the action of gonadotrophin releasing hormone (GnRH), which stimulates hormone production from the gonadotrophs, and is secreted by the hypothalamus (Bremner *et al.*, 2001).

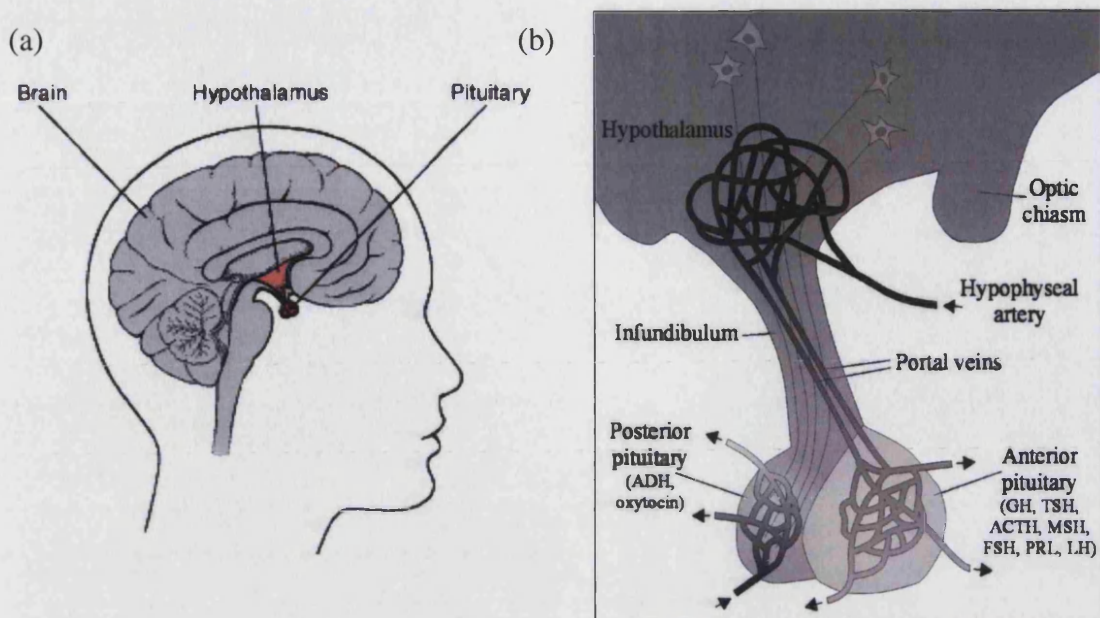


Figure 1.9. Anatomy of the pituitary gland and hypothalamus. (a) Position of the pituitary gland and hypothalamus in the human brain. From the Merck Manual – Second Home Edition (http://www.merck.com/mrkshared/mmanual_home2/fg/fg162_1.jsp). (b) Diagram of the general anatomy of the human pituitary gland, showing the relationship with the hypothalamus. Simplified structures of the blood vessels in the pituitary are shown, including the hypothalamic-hypophyseal portal veins connecting the capillary beds in the anterior pituitary and the hypothalamus. The general location of the hormone-secreting neurons, with axons extending from the cell bodies in the hypothalamus into the posterior pituitary, is shown.

1.6.3. Pituitary development

The development of the pituitary has been studied in detail, particularly in the mouse, but also in other model systems such as *Xenopus* and chick. The anterior and posterior pituitary are derived from different embryological origins. The anterior pituitary is derived from the oral ectoderm, and the posterior pituitary is derived from the neural ectoderm (Thapar *et al.*, 2001). In mouse, pituitary formation is recognised to start at about embryonic day 8.5 (e8.5) when the previously most anterior part of the embryo, the anterior neural ridge, moves ventrally to form the oral roof ectoderm (Dasen.J.S. and Rosenfeld, 1999). From e8.5 to e9.0, thickening and then invagination of part of the oral ectoderm begins, which starts to form a structure known as Rathke's pouch (Figure 1.10a). When Rathke's pouch is forming, evagination of the adjacent neural ectoderm layer also occurs, in a region known as the infundibulum (Figure 1.10a). Over time, Rathke's pouch invaginates further, and by e12 it has closed and separated from the oral ectodermal layer (Figure 1.10b,c). Cells proliferate from the epithelial cells in Rathke's pouch and become the progenitors of the hormone-secreting cell of the anterior pituitary, and each lineage is located in a distinct part of the anterior pituitary initially, although in the mature pituitary gland the hormone-secreting cells are not organised in such discrete domains (Figure 1.10d) (Dasen.J.S. and Rosenfeld, 1999).

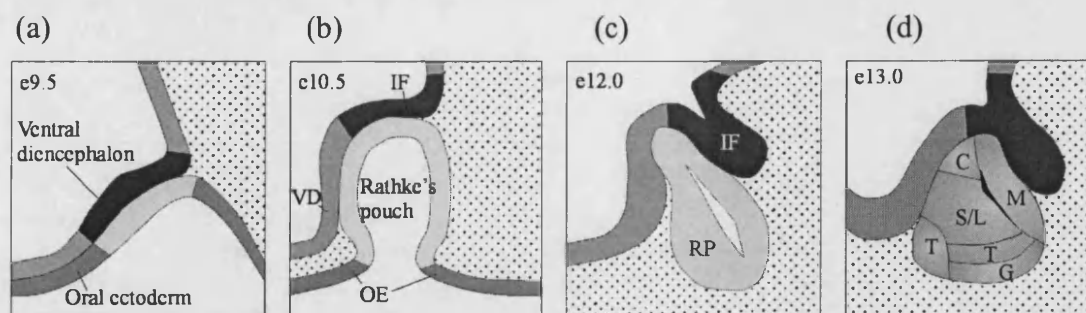


Figure 1.10. Development of the pituitary gland in the mouse. The approximate position of the hormone secreting cell types in the developing anterior pituitary are shown in (d). Abbreviations: IF – infundibulum, VD – ventral diencephalon, OE – oral ectoderm, RP – Rathke's pouch, M – melanotrophs, C – corticotrophs, S – somatotrophs, L – lactotrophs, T – thyrotrophs, G – gonadotrophs. Adapted from (Dasen.J.S. and Rosenfeld, 1999)

1.6.3.1. Signalling molecules and transcription factors in pituitary development

Many of the factors that are important for pituitary development have been studied in various animal models, and mutations in many of the genes that are expressed during pituitary development have been found in individuals with pituitary disorders. Signalling molecules are important in the early development and specification of Rathke's pouch. Around e9.5 in the mouse (equivalent to the third week of gestation in humans), signals from the infundibulum induce the development of Rathke's pouch, including Bmp4 (Bone morphogenic protein 4), along with Wnt5a (Wingless-type MMTV integration site family member 5a) and Fgf8 (Fibroblast growth factor 8) (Figure 1.11a) (Takuma *et al.*, 1998; Treier *et al.*, 1998). In the region of oral ectoderm that forms Rathke's pouch, *Wnt4* and *Bmp2* are expressed as it becomes committed to forming the anterior pituitary (Treier *et al.*, 1998; Savage *et al.*, 2003) (Figure 1.11a). Sonic hedgehog is initially expressed throughout the oral ectoderm, but it is excluded from the *Wnt4*- and *Bmp2*-expressing region as Rathke's pouch begins to form (Treier *et al.*, 1998). The paired-like homeodomain transcription

factors *Pitx1* and *Pitx2* are both expressed in the oral epithelium and endoderm when Rathke's pouch begins to form, and the LIM homeobox genes *Lhx3* and *Lhx4* are expressed in the rudimentary pouch (Figure 1.11b) (Savage *et al.*, 2003). The homeobox gene *Hesx1* has a restricted expression pattern at e9.5, and is only expressed in Rathke's pouch at that time during development although its earlier expression pattern is more widespread (Figure 1.11b) (Hermesz *et al.*, 1996). As the pouch further invaginates, around e10.5, other transcription factors are expressed, including *Prop1* (Prophet of *Pit1*) within Rathke's pouch, *Nkx3.1* in the most dorsal part of the pouch, closest to the infundibulum, and *Pax6* is also expressed in Rathke's pouch (Figure 1.11b) (Dasen.J.S. and Rosenfeld, 1999). As the anterior pituitary differentiates, *Hesx1* expression is turned off, in a ventral to dorsal direction, which coincides with the initiation of expression of *Pou1f1* (*Pit1*) (Hermesz *et al.*, 1996). The distinct spatial and temporal expression patterns of these and other genes within the developing anterior pituitary allow the specification of the different hormone-releasing cell types within the mature pituitary gland.

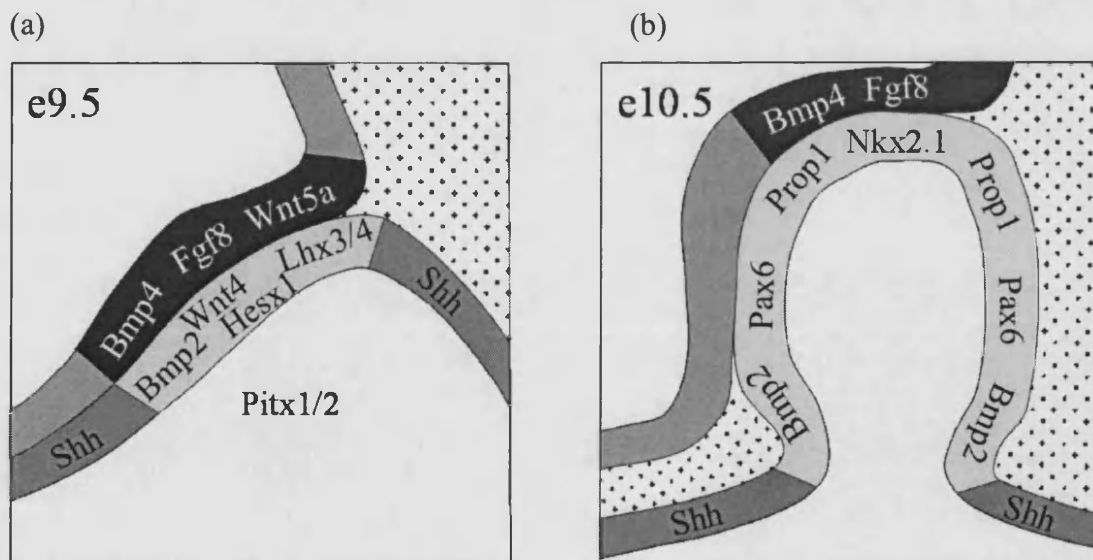


Figure 1.11. Expression of some signalling molecules and transcription factors during early mouse pituitary development. The infundibulum is shaded black, and Rathke's pouch is light grey. Adapted from (Dasen.J.S. and Rosenfeld, 1999;Cohen and Radovick, 2002)

1.6.3.2. *SOX3* in pituitary development

Sox3 is expressed throughout the CNS, including the ventral diencephalon and infundibulum, but not in Rathke's pouch (Rizzoti *et al.*, 2004). Mice that have a targeted deletion of *Sox3* show variable defects in pituitary morphology and function, and defects in CNS midline structures, as well as craniofacial anomalies (Rizzoti *et al.*, 2004). During development in *Sox3* deficient mice, the infundibulum is flattened and Rathke's pouch is expanded dorsally and contains a bifurcated cleft (Rizzoti *et al.*, 2004). Within the infundibulum *Bmp4* and *Fgf8* showed a transient expanded expression domain in the absence of *Sox3* before any changes in Rathke's pouch became apparent, and proliferation is reduced in the infundibulum (Rizzoti *et al.*, 2004). *Sox3* expression was also found to persist postnatally in the ventral hypothalamus, and it may also be involved in the regulation of pituitary hormone secretion (Rizzoti *et al.*, 2004).

1.6.4. Screening for changes in *SOX3* gene dosage

The evidence published so far does suggest that *SOX3* dosage could play a role in the aetiology of X-linked hypopituitarism. Screening a cohort of male patients with, or who are suspected to have, X-linked hypopituitarism for changes in dosage in Xq26-27 may enable further refinement of the X-linked hypopituitarism critical region. These studies may also clarify the role of *SOX3* gene dosage in this disorder.

1.7. Objectives:

To investigate the nature of tandem genomic duplications in Xq22 involving *PLP1*, by mapping the extent of different duplications, and determining the sequences present at the breakpoints, to further understand the mechanisms behind these types of duplications.

To study the complex rearrangements present in the family with insertion of *PLP1* into Xq26, including both the insertion and the mosaic deletion in the carrier mother, to help determine if the insertion and deletion rearrangements are related in some way, and to further understand the events that gave rise to these rearrangements.

To develop MAPH as an alternative technique to interphase FISH for the diagnosis of duplications of *PLP1* and the surrounding genomic region.

To screen a cohort of patients with X-linked hypopituitarism for duplications of Xq26-27 involving the *SOX3* gene, and further investigation of any found, using a similar strategy to that pursued for the *PLP1* duplications.

2.0. MATERIALS AND METHODS

2.1. General materials

2.1.1. Products and reagents

All chemicals were obtained from Sigma-Aldrich or BDH, except for the following:

Genescan-500 Tamra size standard and BigDye Terminator version 3.1 reaction mix was obtained from Applied Biosystems

LB broth base, RPMI 1640 medium, Penicillin/Streptomycin, L-glutamine, KaryoMAX colcemid, herring sperm DNA, agarose, 100bp ladder, λ /HindIII DNA, ϕ X174/HaeIII DNA, BIONICK labelling kits, Zero Background/Kan cloning kits (including plasmid pZErO-2) and One Shot TOP10 competent cells were all supplied by Invitrogen.

New England Biolabs supplied T4 DNA ligase, Klenow fragment, and all restriction enzymes, except for the following: *Mva*I from MBI Fermentas; *Bam*HI and *Rsa*I supplied by Promega; *Eae*I from Roche; and *Eco*TI4I was obtained from Amersham Biosciences. Amersham Biosciences also supplied Hybond N⁺ nylon membrane, MegaBACE Long Read Matrix and LPA buffer .

Foetal Calf Serum was supplied by PAA Laboratories.

Biopro Taq, 10 x NH₄ buffer, MgCl₂ and dNTPs were supplied by Bioline.

HotStarTaq, 10 x PCR buffer, Q solution, Proofstart enzyme, Proofstart 10 x PCR buffer and QIAquick gel extraction kits all were obtained from Qiagen.

Roche Applied Science supplied mouse anti-digoxigenin antibody, DIG-Nick translation mix and *C_{ot}-1* DNA.

Sequagel concentrate, diluent and buffer were obtained from National Diagnostics.

Vysis supplied Spectrum Green and Spectrum Orange dUTP, Nick translation kit and CEP-X centromere probes in both Spectrum Orange and Spectrum Aqua.

Vectashield mounting medium containing DAPI, Texas Red Avidin DCS antibody, Fluorescein Avidin DCS antibody and biotinylated anti-avidin DCS antibody were all supplied by Vector labs.

Vulcanising rubber solution was supplied by Weldtite.

Non fat dried milk was manufactured by Premier Brands.

Shandon coverplates were supplied by Thermo Electron.

Human genomic clones were supplied by the Wellcome Trust Sanger Institute.

2.1.2. Buffers, solutions and mixes

2.1.2.1. PCR Buffers

NH4 10 x reaction buffer: 160mM (NH₄)₂SO₄, 670mM Tris-Cl pH 8.0, 0.1% Tween 20

10 x PCR buffer: Tris-Cl, KCl, (NH₄)₂SO₄, 1.5mM MgCl₂

Proofstart buffer: Tris-Cl, KCl, (NH₄)₂SO₄, 1.5mM MgSO₄, BSA, Triton X-100

2.1.2.2. Solutions

TE: 10mM Tris, 1mM EDTA pH 8.0

5 x TBE: 0.089M Tris, 0.089M Boric acid, 2mM EDTA pH 8.0

20 x SSC 2.9M NaCl, 0.3M sodium citrate

Loading dye	50% glycerol, Orange G
Phenol:chloroform:isoamyl alcohol:	50% buffered phenol, 48% chloroform, 2% isoamyl alcohol
Lysis solution for fibre slides:	50mM NaOH, 28.6% ethanol
FISH hybridisation mix:	50% formamide, 10% dextran sulphate, 2 x SSC
MAPH prehybridisation solution:	0.5M sodium phosphate pH 7.2, 7% SDS, 0.1mg/ml heat-denatured herring sperm DNA

2.1.2.3. Media solutions

LB broth:	1% Bactotryptone, 0.5% Bacto-yeast extract, 1% NaCl
LB agar:	1% Bactotryptone, 0.5% Bacto-yeast extract, 1% NaCl, 15g bacto-agar
SOC medium:	2% Tryptone, 0.5% Yeast extract, 0.05% NaCl, 2.5mM KCl, 10mM MgCl ₂ , 20mM Glucose

2.1.2.4. Plasmid alkaline lysis extraction solutions

Solution 1: 50mM glucose, 25mM Tris-Cl pH 8.0, 10mM EDTA

Solution 2: 0.2N NaOH, 1% SDS

Solution 3: 5M potassium acetate, 11.5% glacial acetic acid

2.1.2.5. Probe labelling kits

BIONICK labelling kit

10x dNTP mix: 0.2mM each dCTP, dGTP and dTTP, 0.1 mM dATP, 0.1mM biotin-14-dATP, 500mM Tris-HCl, pH 7.8, 50mM MgCl₂, 100 mM β-mercaptoethanol, 100μg/ml nuclease-free BSA

Enzyme mix: 0.5U/μl DNA polymerase I, 0.007 U/μl DNase I, 50mM Tris-HCl pH 7.5, 5mM MgCl₂, 0.1mM phenylmethylsulphonyl fluoride, 50% glycerol, 100μg/ml nuclease-free BSA

DIG-Nick Translation kit

DIG-Nick Translation Mix: Stabilised reaction buffer in 50% glycerol (v/v), DNA polymerase I, DNase I, 0.25mM each dATP, dCTP, dGTP, 0.17mM dTTP, 0.08 Digoxigenin-11-dUTP

Nick Translation Kit

10 x nick translation buffer:	500mM Tris-HCl, pH 7.2, 100mM HgSO ₄ , 1mM DTT
Nick translation enzyme mix:	DNA polymerase I and DNase I in 50% glycerol, 50mM MgSO ₄ , 0.1mM DTT, 0.5 mg/ml nuclease-free BSA

2.1.2.6. Polyacrylamide gel solutions

Sequagel concentrate (1 litre):	236.5g acrylamide, 12.5g methylene, 500g 7M urea, pH 8.3, bisacrylamide
Sequagel diluent (1 litre):	500g 7M urea, pH 8.3
Sequagel buffer (200ml):	50% 8.3M urea in 10 x TBE, pH 8.3
Loading buffer for ABI 377:	60% formamide, 20% Genescan-500 Tamra, with 20mM EDTA, 20% loading buffer (contains 50mg/ml blue dextran, 25mM EDTA)

2.1.3. Primers

All oligonucleotide primers were synthesised by MWG-Biotech and all nucleotide sequences and reaction conditions are listed in Appendix A.

2.2. Methods

2.2.1. Polymerase Chain Reaction (PCR)

PCR was carried out using one of two DNA polymerase enzymes. Reactions using Biopro DNA polymerase contained a 1 in 10 dilution of 10 x NH₄ reaction buffer, 200μM each dNTP, 1.5mM MgCl₂, 0.5-1μl of template DNA (50-250ng) and 20pmol each oligonucleotide primer in a total volume of 25μl. Reactions were initially heated to 94°C for five minutes, after which 0.75U Biopro DNA polymerase (0.15μl) was added to the PCR and cycling was carried out, for 30 cycles unless stated otherwise; denaturation at 94°C for 30 seconds, primer annealing at 55°C (unless otherwise stated in Appendix C) for 30 seconds, and elongation at 72°C for between 45 seconds to 1 minute. After 30 cycles, there was a final elongation step at 72°C for 10 minutes.

PCR using HotStarTaq DNA polymerase was carried out in a total reaction volume of 25μl containing a 1 in 10 dilution of 10 x PCR buffer, 200μM each dNTP, 0.5-1μl of template DNA (50-250ng), 20pmol each oligonucleotide primer and 0.75 units of HotStarTaq (0.15μl). Cycling was carried out, with an initial hotstart at 95°C for 15 minutes, followed by 30 cycles (unless otherwise stated) of denaturation at 95°C for 30 seconds, primer annealing at 55°C (or other stated temperature) for 30-45 seconds, then an elongation step at 72°C for 45 seconds to one minute. After cycling, the reaction was held at 72°C for 10 minutes for a final elongation step.

PCR reactions were carried out on either an Omn-E thermal cycler (Hybaid), or a Mastercycler machine (Eppendorf). If a heated lid was not used, PCR reaction mixes were overlaid with 25μl mineral oil.

2.2.1.1. Long range PCR

Long-range PCR was carried out using oligonucleotide primers with a length generally between 25 and 30 nucleotides, in a 10 μ l, 20 μ l or 25 μ l reaction. For a 25 μ l reaction the components were as follows: 0.5 μ l DNA (approximately 50-100ng); 2.5 μ l 10 x PCR buffer; 280mM dNTPs; 5 μ l Q solution; 20pmol each primer; 0.5 μ l (0.2 units) diluted Proofstart enzyme (2 units (0.8 μ l) diluted in 9.2 μ l 1x Proofstart PCR buffer); 2.5 units HotStarTaq and H₂O to a volume of 25 μ l. For a smaller volume such as 20 μ l the quantities were reduced accordingly, e.g. by 20%. Cycling conditions were: a 95°C hot start for 15 minutes, then the following was repeated for 40 cycles - 95°C denaturation for 15 seconds; primer annealing temperature for 30 seconds (typically 65°C); 68°C for 2-15 minutes depending on the length of product expected, allowing approximately one minute for each kilobase of DNA. Then there was a final elongation step at 68°C for 10 minutes before holding at 4°C.

2.2.1.2. Degenerate oligonucleotide primed PCR

A reaction was set up containing 50 flow-sorted chromosomes (in 33 μ l H₂O), 5 μ l 10x PCR buffer, 200mM each dNTP, 100pmol 6MW primer, 2.5U HotStarTaq and water to a volume of 50 μ l. Cycling conditions on an Eppendorf Mastercycler machine were: 94°C for 9 minutes followed by eight cycles of 94°C for 1 minute, 30°C for 1.5 minutes, and 72°C for 3 minutes, then 25 cycles of 94°C for 1 minute, 62°C for 1 minute, 72°C for 1.5 minutes, and a final extension at 72°C for 8 minutes.

2.2.1.3. UPQFM-PCR

Primary reactions were set up in a total volume of 10 μ l with quantities of dNTPs, buffer and HotStarTaq in proportion to the amounts used in a 25 μ l reaction (see above) and 100ng genomic DNA. 1, 2 or 4pmol of each tagged primer was used in each reaction, and

up to 6 pairs of tagged primers were used. Cycling conditions on an Eppendorf Mastercycler were as follows: 95°C hot start for 15 minutes, then 10 cycles of 94°C denaturation for 30 seconds, 56°C primer annealing for 45 seconds and 72°C elongation for 45 seconds, followed by holding at 72°C for 10 minutes.

Secondary PCR reactions were then carried out in a total volume of 20 µl, seeded with 2 µl of the primary reaction as a template. The reaction mix also contained 2 µl 10x PCR buffer, 200 µM each dNTP, 20 pmol each universal primer and 1.6U HotStarTaq. PCR conditions were similar to the primary reaction conditions, but 20 cycles were carried out.

2.2.1.3.1. UPQFM-PCR dosage analysis

2-4 pairs of UPQFM-PCR primers were used alongside two pairs of control UPQFM-PCR primers in each experiment, one from exon 6 of *PLP1* (PLP1) and one from the *CFTR* locus on chromosome 7 (CF). Dosage of the sequence amplified by each primer pair was calculated from the ratio obtained by dividing the ratio of the fluorescent signal from each primer pair against each control primer by the same ratio in a group of normal, sex-matched controls. For a male with *PLP1* duplication, the expected value for a single-copy sequence when compared against this primer pair is 0.5, and the expected value for a duplicated sequence is 1. For a female *PLP1* duplication carrier, who has three copies of *PLP1*, the expected ratio from a non-duplicated sequence (two copies) is 0.67, and the expected value for a duplicated target sequence is 1. For ratios against the autosomal control CF primer pair, the male ratio expected for a non-duplicated sequence is 1; for a duplicated target sequence a ratio of 2 is expected. Female ratios against CF are expected to be 1 for normal copy number and 1.5 for one extra copy of a target sequence. To classify UPQFM-PCR results as either normal copy number or duplicated, a threshold that was halfway between the two expected ratios was used as the boundary between normal and increased copy number. Thus for males, when dosage ratios are calculated

relative to the PLP1 primers, sequences were classed as single copy when ≤ 0.75 , and duplicated when > 0.75 . Similarly for male/CF ratios, single-copy ≤ 1.5 and duplicated > 1.5 . The ratios for females were classified as follows: against PLP1 normal copy number ≤ 0.83 , duplicated > 0.83 , against CF normal copy number ≤ 1.25 and duplicated > 1.25 .

2.2.1.4. Inverse PCR

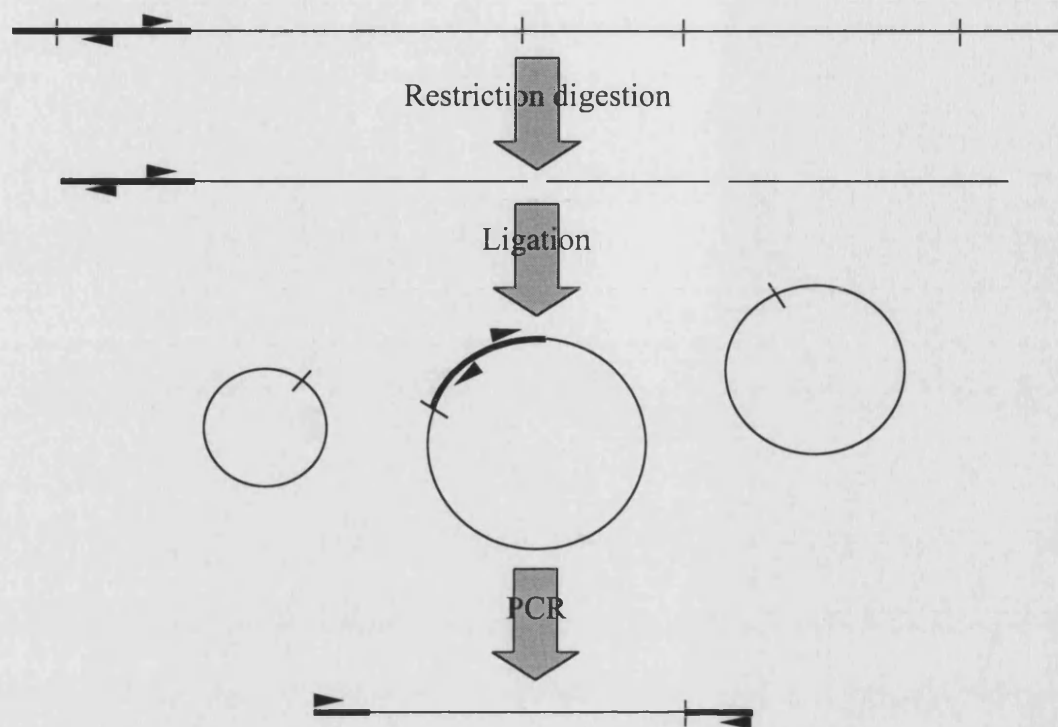


Figure 2.1. Diagram showing the principles of inverse PCR. Bold lines show known sequences, PCR primers (and orientation) are shown by arrows, and vertical lines show restriction sites. In summary, DNA is digested with a restriction enzyme, and the fragments are then ligated at a low concentration, which promotes circularisation. PCR primers used for the inverse PCR are in the opposite orientation to normal PCR primers, so would not normally be able to amplify. Following digestion and ligation, the primers should then be able to amplify the sequences that surround the restriction sites. If the junction between known and unknown sequence is located between a primer binding site and a restriction site then it will be amplified by the inverse PCR, as well as the junction created by the circularisation.

2.2.1.4.1. Genomic DNA restriction digestion

2µg genomic DNA, or DNA obtained from a human genomic clone, was digested with 20-40 units of the appropriate restriction enzyme, in a total volume of 40µl with 4µl of the appropriate enzyme buffer (10 x stock concentration), for 4.5 hours in a waterbath at the temperature recommended by the manufacturer. The reaction was stopped by heating to 68°C for 15 minutes.

2.2.1.4.2. Ligation

200ng genomic DNA that had been completely digested by a restriction enzyme was incubated at 16°C in a reaction mixture containing 1500 units T4 DNA ligase and 10µl 10 x T4 ligase buffer in a total volume of 100µl, to ensure a low concentration of DNA (2ng/µl) and promote circularisation. The reaction was stopped after 16 hours by heating to 94°C for 15 minutes. The ligation reaction was then ethanol precipitated (see 2.2.3.2.). The pellet was completely air-dried, and resuspended in 5µl H₂O.

2.2.1.4.3. PCR reaction

The 5µl of resuspended ligation reaction was used in a 25µl PCR reaction using HotStarTaq (see section 2.2.1.) in 0.5ml thin walled PCR tubes. The reaction was incubated on an Eppendorf Mastercycler machine, initially at 94°C for 15 minutes, then taken through 35 cycles of 94°C for one minute, 57°C for one minute and 72°C for three minutes, followed by 72°C for 10 minutes, and the reactions were held at 4°C afterwards. PCR products were electrophoresed on a 1% agarose gel (see section 2.2.2.2.). Altered size bands were cut out of the gel and DNA was extracted from the agarose slice prior to sequencing (see section 2.2.1.4.4.).

2.2.1.4.4. Extraction of DNA from agarose gels

DNA was extracted from excised agarose gel slices using the QIAquick gel extraction kit protocol recommended by the manufacturer. Each gel slice was placed in a 1.5ml tube and weighed, three gel volumes of buffer QG were added (3µl per 1mg of gel) and incubated in a water bath at 50°C for 10 minutes. For DNA fragments <500bp and >4Kb, one gel volume of isopropanol was added and mixed. The dissolved sample was added to a QIAquick spin column, which was placed in a 2ml collection tube, and spun at 13000rpm for 10 minutes in a microcentrifuge. The flow-through was discarded and 0.5ml buffer QG was added to the column and spun at 13000rpm for 10 minutes in a microcentrifuge to remove all traces of agarose. 0.75ml buffer PE was added to the spin column, which was allowed to stand for 2-5 minutes before spinning at 13000rpm for 10 minutes in a microcentrifuge. The flow-through was discarded and the spin tube and collection tube were re-centrifuged under the same conditions to remove all residual buffer. The column was then placed in a clean 1.5ml tube and 30µl 10mM Tris-Cl or H₂O was added to the centre of the QIAquick membrane, allowed to stand for one minute, then spun at 13000rpm for 10 minutes in a microcentrifuge to elute the DNA.

2.2.2. Electrophoresis

2.2.2.1. Electrophoresis of PCR products using ABI 377 DNA sequencer

For analysis and quantitation of gene dosage, by UPQFM-PCR and MAPH, and genotyping polymorphic microsatellite markers, fluorescently labelled PCR products were electrophoresed on an ABI 377 automated sequencer (Applied Biosystems).

2.2.2.1.1. Preparation of polyacrylamide gels

A 4% polyacrylamide gel solution was used. For 12cm plates, 16.7ml of gel mix was prepared by mixing 12.3ml Sequagel diluent, 2.7ml Sequagel concentrate and 1.7ml

Sequagel buffer. If longer 36cm plates were used, a greater volume (40ml) of gel mix was made, consisting of 29.6ml diluent, 6.4ml concentrate and 4ml buffer. Just prior to pouring the gel, 133µl 10% APS and 8µl TEMED were added to the mixture (320µl 10% APS and 16µl TEMED for the 40ml gel solution). Plates were washed with anionic detergent, rinsed thoroughly with purified water, and then air-dried. Plates were assembled in the gel cassette with 0.2mm plastic spacers. The gel solution was drawn into a 50ml syringe and quickly poured into the horizontal prepared plates. The flat end of a plastic shark's-tooth comb was inserted into the top of the gel, between the two glass plates, and the top of the gel was held together with metal clips. Gels were left to set for at least one hour before use. Once the gel was set, the clips and flat comb were removed and any excess dried acrylamide was wiped from the outside of the plates. A paper shark's-tooth comb, with either 36 or 48 wells, was placed in the top of the gel and then the gel cassette was loaded in the ABI 377 DNA sequencer. 1 or 2µl of product was mixed with 2µl loading dye/Tamra size standard and denatured by heating to 95°C for 1-3 minutes before chilling on ice. The buffer tanks were filled with 1 x TBE and 2µl of product and loading solution was loaded in each lane. The gel was electrophoresed until all PCR products and size standard had been detected. Electrophoresis conditions were 3KV, 60mA and 200W, the gel was kept at a constant temperature of 51°C, and the laser power was 40mW. Gels were analysed using Genotyper software (Applied Biosystems).

2.2.2.2. Agarose gel electrophoresis

1g, 0.5g, or 0.4g electrophoresis grade agarose was dissolved in 50ml 1 x TBE in a glass bottle or flask, then melted by heating in a microwave for 1-2 minutes, to make agarose gels of concentrations 2%, 1% and 0.8% respectively. 2µl ethidium bromide (10mg/ml) was added to the melted gel solution, which was mixed and allowed to cool slightly before pouring into a minigel tank, with appropriate combs inserted (Flowgen). The gel

2.2.4. Sequencing

Sequencing reactions were carried out using 5µl of gel extracted PCR product in a total reaction volume of 10µl containing 4µl BigDye Terminator version 3.1 reaction mix and 3.2pmol primer. Cycle sequencing was carried out as follows: 96°C for 1 minute, then 35 cycles of 96°C for 10 seconds, 50°C for 10 seconds and 60°C for 4 minutes, after which the reactions were stored at 4°C.

2.2.4.1. Sephadex sequencing reaction clean-up (carried out by Kerra Pearce)

Prior to electrophoresis, unincorporated dye terminators, small molecule contaminants and salts were removed from the sequencing reactions by gel filtration. 45µl dry Sephadex G-50 Superfine beads was loaded into each well of a 96-well MultiScreen HV plate (Millipore). 300µl H₂O was added to each well and left for 3 hours at room temperature to allow the resin to swell. The HV plate was centrifuged at 910 x g for 5 minutes, after being placed on top of a standard 96-well microplate, held in place with a centrifuge alignment frame, to pack the columns. The columns were pre-rinsed with 150µl H₂O, which was removed by centrifuging at 910 x g for 5 minutes. Sequencing reactions were made up to 20µl by the addition of 10µl H₂O, and applied to the centre of each well. The HV plate was then placed on top of a skirted 96 well plate and centrifuged at 910 x g for 5 minutes.

2.2.4.2. MegaBACE sequencing (carried out by Kerra Pearce)

Sequencing reactions were electrophoresed on the MegaBACE 1000 capillary electrophoresis system (Amersham Biosciences), according to the manufacturers instructions. 6 tubes of MegaBACE Long Read Matrix (at room temperature) were spun at 4000rpm in a microcentrifuge for 4 minutes. The capillaries were injected with MegaBACE Long Read Matrix, and prerun with LPA buffer. The samples were

was left to set completely, then the combs were removed and 50ml of 1x TBE buffer was added to the tank. 3-5µl loading dye was added to 5-20µl of each product to be run on the gel, and loaded into the wells. Molecular weight markers appropriate to the expected sizes of product were used; for smaller products (up to 2000bp) a 100bp ladder was used and for larger fragments a λ DNA/*Hind*III marker was used. 0.5µg molecular weight marker was used in each marker lane. Electrophoresis was carried out at a voltage between 75-125V for between 20 minutes to one hour. Gel imaging and capture was carried out using the ChemiDoc gel documentation system and Quantity One software (version 4.4.0) (BioRad Laboratories).

2.2.3. DNA precipitation

2.2.3.1. Phenol/chloroform extraction

Phenol/ chloroform extraction of DNA was carried out by addition of an equal volume of phenol:chloroform:isoamyl alcohol mixture, vortexing until thoroughly mixed, then separating the aqueous and organic layers by spinning at 13000rpm for 10 minutes in a microcentrifuge. The upper aqueous layer was then removed to a clean 1.5ml tube.

2.2.3.2. Ethanol precipitation of DNA

DNA was ethanol precipitated by addition of 1/10th volume 3M sodium acetate pH 5.2 and 2.5 times the total volume of cold 100% ethanol, then leaving at -80°C for at least one hour or -20°C overnight. The precipitated DNA was pelleted by spinning at 13000rpm in a microcentrifuge for 10 minutes, and the supernatant was removed. The pellet was washed by adding 0.5ml cold 70% ethanol and spinning in a microcentrifuge for 5 minutes at a speed of 13000rpm. The supernatant was completely removed, and the pellet allowed to completely air-dry, before it was resuspended in the appropriate volume of H₂O or TE (pH 8.0).

denatured by heating to 95°C for two minutes and then cooled on ice. Samples were electrokinetically injected into the capillary from a skirted with an injection voltage of 3KV and an injection time of 1 minute. The sequencing samples were then electrophoresed at 9KV for 100 minutes.

2.2.5. Preparation of FISH probes from human genomic clones

2.2.5.1. Glycerol stocks

Human genomic clones were initially received as stabs growing on LB agar. These were streaked on LB plates, with the appropriate antibiotic selection (either 30µg/ml kanamycin for cosmids, 25µg/ml kanamycin for PACs or 20µg/ml chloramphenicol for BACs) and grown at 37°C overnight. Single colonies were inoculated into 10ml LB broth, with appropriate antibiotic selection as before, and grown at 37°C overnight with 225rpm shaking. 0.5ml of the growing culture was mixed with 0.5ml 30% glycerol in a 1.5ml tube and stored at -80°C.

2.2.5.2. Cosmid/PAC/BAC/plasmid alkaline lysis minipreparation

A single bacterial colony, or 1µl from a glycerol stock, was inoculated into 10ml LB broth in a 50ml tube and incubated at 37°C overnight with 225rpm shaking. Antibiotic selection was used, as in section 2.2.5.1. The following day, the overnight cultures were centrifuged at 2000rpm at 4°C for 10 minutes, before discarding the supernatant and gently resuspending the bacterial pellet in 100µl ice-cold solution 1. At this point the suspension was transferred to a 1.5ml tube. 200µl freshly made solution 2 at room temperature was added, the tube was inverted to mix the contents and then left on ice for 5 minutes. 150µl ice-cold solution 3 was added, the contents were mixed by inversion, and stored on ice for a further 10 minutes. The tubes containing the lysed cells were then spun at 13000rpm in a microcentrifuge for 10 minutes, and the supernatant was

transferred to a fresh 1.5ml tube. The supernatant was phenol extracted (see 2.2.3.1) and ethanol precipitated (see 2.2.3.2.) and the pellets allowed to dry thoroughly, either at room temperature or at 37°C, before being resuspended in 30µl TE (pH 8.0). 20µg RNase A was added to each resuspended pellet which was then incubated in a 37°C water bath for 10-30 minutes.

2.2.6. Cell culture of lymphoblastoid cell lines

EBV transformed human B cell lymphoblastoid cell lines (ECACC) were grown in RPMI 1640 medium containing additional 10% foetal calf serum, 2mM L-Glutamine, 5000 units Penicillin and 5000µg Streptomycin. Vials of cells were removed from storage in liquid nitrogen and quickly thawed by holding in a plastic beaker containing water at 37°C.

Cells were pelleted by spinning at 1500rpm in a centrifuge for 5 minutes and the freezing medium was removed from the pellet, which was gently resuspended in 2ml pre-warmed growth medium and transferred into a flask already containing 3ml pre-warmed medium. The flasks containing the cells were then placed in a 37°C incubator with 5% CO₂. Every 2-3 days, the medium was removed and replaced with fresh pre-warmed medium, and growing cell suspensions were split as necessary.

2.2.6.1. Freezing down of lymphoblastoid cell lines

15ml of healthy growing lymphoblastoid culture was pelleted by centrifuging at 1500rpm for 5 minutes. The growth medium was removed and replaced with 70% RPMI 1640 (including FCS, L-Glutamine and Penicillin/Streptomycin, as above), 20% FCS and 10% DMSO to a volume of 1.8ml. The cells were resuspended in this freezing medium, and transferred into 2ml cryotubes. The tubes were placed in an isopropanol bath and placed in a -80°C freezer for at least 24 hours, to allow gradual cooling at a rate of 1°C per minute, after which they were stored in liquid nitrogen.

2.2.6.2. Preparation of cell suspensions for interphase and metaphase FISH from lymphoblastoid cell line

For primarily interphase FISH cell suspensions, lymphoblastoid cell lines were starved for up to five days, to reduce the number of dividing cells present. For a cell line that was going to be used for mainly metaphase FISH experiments, healthy growing cells were used. 150µl KaryoMAX colcemid (10µg/ml) was added to 15ml cell suspension, which was then incubated at 37°C for 1 hour, before being spun in a centrifuge at 1500rpm for 5 minutes. The supernatant was removed and the cell pellet gently flicked to resuspend. 0.075M KCl was steadily added to the cells, dropwise up to a volume of 10mls. Cells were incubated at 37°C until 15 minutes had elapsed since the first drop was added, to prevent nuclear lysis, and then centrifuged at 1500rpm for 5 minutes. The procedure was then repeated three times, but instead of using 0.075M KCl, a mixture of 3:1 methanol:acetic acid was used, and no incubation step was required. On the final repetition of the procedure the cell suspension was not centrifuged and the 15ml fixed cell suspension was stored at -20°C until required.

2.2.6.3. Slide preparation for DNA fibres

3ml of cell suspension from a growing lymphoblastoid cell line was centrifuged at 1200 rpm for 5 minutes. The supernatant was removed and the cell pellet resuspended in 1ml PBS, and centrifuged at 1200 rpm for 5 minutes. The supernatant was removed again and the cell pellet resuspended in PBS, to give a concentration of approximately 2×10^6 /ml. 10µl of this cell suspension was spread over a 1cm diameter circle on the upper part of a clean microscope slide, and allowed to dry for 30 minutes. Microscope slides that had been previously stored in 96% ethanol and then air-dried were used. A slide was then fitted onto a plastic Shandon coverplate, and held together vertically with the dried cell spot at the top of the slide facing the chamber. 150µl lysis solution was added to the gap

at the top of the slide chamber and allowed to flow through the slide chamber until the meniscus had reached the top of the cell spot. 150µl 96% ethanol was then applied to the top of the slide chamber and allowed to drain for about 30 seconds. The slide was then pulled slowly away from the slide chamber, and allowed to air-dry. Slides were fixed in acetone for 10 minutes, and left at room temperature for at least 24 hours before hybridisation.

2.2.7. Probe labelling for FISH

2.2.7.1. Biotin labelling

DNA from human genomic cosmid, PAC or BAC clones grown in *E.coli* and obtained by an alkaline lysis minipreparation protocol (see 2.2.5.2.) was labelled with biotin-14-dATP by nick translation in a 50µl reaction using a BIONICK kit. 500ng DNA from each miniprep was used in each labelling, along with 5µl 10x dNTP mix and 5µl enzyme mix. The labelling reactions were incubated at 16°C in a waterbath for 2 hours, and then placed on ice. 5µl from each labelling reaction was electrophoresed on a 2% agarose gel alongside a 100bp ladder size standard (see 2.2.2.2.) to check that the fragments were within the optimum size range (300-500bp) before the reaction was stopped by addition of 2.5µmol EDTA.

2.2.7.2. Digoxigenin labelling

DNA from human genomic cosmid, PAC or BAC clones grown in *E.coli* and obtained by an alkaline lysis mini-preparation protocol (see 2.2.5.2.) was labelled with digoxigenin-11-dUTP by nick translation using the DIG-Nick Translation Mix. 1µg DNA was made up to a volume of 16µl with sterile water, and 4µl 5x DIG-Nick Translation Mix was added. The labelling reaction was mixed, and incubated at 16°C for 1.5 hours in a water bath, then chilled on ice. 2µl from each labelling reaction was electrophoresed on a 2%

agarose gel alongside a 100bp size standard (see 2.2.2.2.) to check that the fragments were within the optimum size range (300-500bp), before the reaction was stopped by the addition of 0.5µmol EDTA and incubation at 70°C for 10 minutes.

2.2.7.3. Direct labelling of probes

Cosmid, PAC or BAC DNA miniprep containing cloned human genomic DNA was directly labelled with either Spectrum Green dUTP or Spectrum Red dUTP using a Nick Translation Kit. 500ng miniprep DNA was used in each labelling reaction, to which was added 2.5µl 10x nick translation buffer, 500pmol each dATP, dCTP and dGTP, 250pmol dTTP, 5µl nick translation enzyme mix, and 250pmol Spectrum Green or Spectrum Orange labelled dUTP. The contents of each reaction were thoroughly mixed by pipetting up and down repeatedly, and the nick translation reactions were incubated at 16°C for 16 hours, before being placed on ice. 5µl from each labelling reaction was run on a 2% agarose gel stained with ethidium bromide with a 100bp size standard ladder (see 2.2.2.2.) to check that the size of the fragments was around the optimal 300bp. The reaction was stopped by heating to 70°C for 10 minutes.

2.2.8. Fluorescence *in-situ* hybridisation

2.2.8.1. Slides/suspensions

Slides were prepared from cell suspensions in 3:1 methanol:acetic acid obtained from lymphoblastoid cell lines or peripheral blood cultures (see section 2.2.6.2.). Glass slides (dimensions 76 x 26 x 1-1.2mm) were kept at -20°C until required, then removed from the freezer and moistened by breathing on the slide. A single drop of cell suspension was dropped onto the middle of the slide and allowed to partially air-dry while the extent of the spot was marked on the underside of the slide using a diamond pen. Then the slide was flooded with a mixture of 3:1 methanol:acetic acid, immediately drained and then completely air-dried. Once dry, the slides were held horizontally and flooded with 70%

acetic acid for a couple of minutes, then drained and completely air-dried. Slides were stored at 4°C for up to a week until used, and were made at least one day before they were used.

2.2.8.2. FISH using probes labelled with biotin and digoxigenin

2.2.8.2.1. Probe precipitation

The probes labelled by nick translation were ethanol precipitated with excess competitor DNA. For PAC and BAC probes, 30µg *C_ot*-1 DNA and 20µg herring sperm DNA were added to 200ng of biotin or digoxigenin labelled probe and ethanol precipitated (as above). For cosmid probes, the same amounts of biotin or digoxigenin labelled probe were used as for PAC and BAC probes, but 10µg *C_ot*-1 DNA and 20µg herring sperm DNA were added, and then ethanol precipitated (see section 2.2.3.2.). The precipitated and thoroughly dried pellet was resuspended in 10µl hybridisation mix, if just a single probe was used on the slide, or otherwise in 5µl hybridisation mix (if two probes were going to be used in the hybridisation).

2.2.8.2.2. Slide and probe preparation

Resuspended probes were denatured by heating in a 75°C waterbath for 5 minutes and they were then transferred to a 37°C waterbath for 30-90 minutes to allow repetitive sequences to preanneal. Slides were washed in 2 x SSC at 37°C for 30-60 minutes. The slides were dehydrated through a series of 70%, 90% and 100% ethanol baths for five minutes each, then drained and allowed to air-dry. Slides were then denatured at 75°C in prewarmed 70% deionised formamide/2 x SSC for 3 minutes. They were transferred to ice-cold 70% ethanol for two 3 minute washes, further dehydrated through 90% and 100% ethanol baths at room temperature for three minutes each, and drained and left to air-dry. A CEP-X centromere probe directly labelled with Spectrum Orange or Spectrum

Aqua was used in all interphase and metaphase FISH experiments. For each slide 0.15µl X centromere probe was added to 2µl hybridisation mix. This was denatured by heating at 75°C for 3 minutes and then cooled on ice.

2.2.8.2.3. Overnight hybridisation

Each 10µl denatured and preannealed probe suspension was mixed with 2.15µl diluted and denatured X centromeric probe and applied to a 22mm x 22mm glass cover slip. The slide was then gently lowered onto the coverslip, ensuring no bubbles were formed. The edges of the cover slip were sealed using vulcanising rubber solution. Slides were incubated overnight at 37°C or 42°C in a moist atmosphere.

2.2.8.2.4. Washes and antibodies

On the following morning the rubber sealant and coverslips were removed, then the slides were washed at 45°C in 50% formamide for 15 minutes with gentle agitation, changing the wash solution every 5 minutes. Slides were washed at 60°C in 0.1 x SSC for 15 minutes, changing the wash solution every 5 minutes. Slides were subsequently transferred to 4 x SSC/0.1% Tween 20 and washed with moderate shaking at room temperature for 5 minutes. The slides were incubated at room temperature for at least 20 minutes in 4 x SSC/5% non-fat dried milk. Excess liquid was drained from the slides, which were incubated for 20 minutes with solution 1 under a 22mm x 50mm glass cover slip (see Table 2.1 for details of antibodies in each detection antibody solution). Unbound antibody was removed by carrying out three 5-minute washes with agitation using 4 x SSC/0.1% Tween 20. Slides were then drained and the hybridisation and washes repeated sequentially with solutions 2 and 3. Slides were protected from light during these stages. After the final wash, slides were drained and mounted in 25µl Vectashield mounting medium containing 1.5µg/ml DAPI counterstain.

	Biotin labelled probe only	Digoxigenin labelled probe only	Both biotin and digoxigenin probes
Solution 1	1µg Fluorescein avidin DCS	0.1µg Mouse anti-digoxigenin	1µg Texas Red Avidin DCS
Solution 2	0.5µg biotinylated anti-avidin D	1µg Goat anti-mouse FITC	0.5µg biotinylated anti-avidin D, 0.1µg Mouse anti-digoxigenin
Solution 3	1µg Fluorescein avidin DCS	-	1µg Texas Red Avidin DCS, 1µg Goat anti-mouse FITC

Table 2.1. Antibody solutions used for indirect detection of probes for FISH. Quantities shown are for detection of signal on one slide. Antibodies were diluted in 100µl 4 x SSC/5% non-fat dried milk per slide

2.2.8.3. FISH using directly labelled probes

2.2.8.3.1. Probe precipitation

For cosmid clones, 10µl from the labelling reaction (200ng) was mixed with 10µg *C_{ot}-1* DNA and 20µg herring sperm DNA, and this was ethanol precipitated (see section 2.2.3.2.). For PAC and BAC clones, 10µl from the labelling reaction (200ng) was mixed with 30µg *C_{ot}-1* DNA, 20µg herring sperm DNA, then the mixture was ethanol precipitated (see section 2.2.3.2.). Precipitated probes were then resuspended in 5µl (if two probes were going to be used in the hybridisation) or 10µl hybridisation mix, if just a single probe was used on the slide.

2.2.8.3.2. Slide and probe preparation

Slides were washed at 37°C in 2 x SSC for 30-60 minutes, dehydrated through a series of 70%, 90% and 100% ethanol baths for five minutes in each, then drained and allowed to air-dry. CEP-X centromere probe directly labelled with Spectrum Orange was used where a single clone (labelled using Spectrum Green) was used for the hybridisation, and if the single clone had been labelled using Spectrum Red, a CEP-X Spectrum Aqua labelled centromere probe was used in the hybridisation. In all dual probe experiments Spectrum Aqua labelled CEP-X centromere probe was used (where one probe had been labelled using Spectrum Green dUTP and the other with Spectrum Red dUTP). 0.15µl centromere probe was diluted in 2µl hybridisation mix, and added to the 10µl hybridisation mix containing the appropriate preannealed and ethanol precipitated probe(s). Hybridisation proceeded under a 22mm x 22mm glass cover slip and the edges were sealed using vulcanising rubber solution.

2.2.8.3.3. Denaturation and hybridisation

The slides were placed on the heated block of the Omnislide *in situ* hybridisation system (Hybaid) and a simulated slide control programme was used which allowed denaturation at 70°C for 1 minute and then reduced the temperature to 42°C and incubated for at least 16 hours, in a moist atmosphere.

2.2.8.3.4. Washes and mounting

After removal of the seal the slides were washed in a prewarmed solution of 0.4 x SSC / 0.3% NP-40 at 73°C for 2 minutes with a short initial gentle agitation for 1-3 seconds. This was followed by a wash in 2 x SSC / 0.1% NP-40 at room temperature for 1 minute. Excess fluid was drained from each slide, which was then mounted in 25µl of Vectashield mounting medium containing 1.5µg/ml DAPI counterstain.

2.2.8.4. Viewing of FISH slides and capture of images

All FISH slides were examined using a Zeiss Axiophot fluorescent microscope with a triple band pass filter and separate Aqua filter with images recorded by Photometrics CCD KAF1400 camera (Photometrics) and controlled with SmartCapture imaging software (Digital Scientific).

2.2.8.5. Scoring FISH slides

FISH was carried out on interphase nuclei that had been obtained from lymphoblastoid cell cultures that had been starved for 5 days, to enrich for non-dividing cells. Interphase nuclei from males were scored when one red signal from the X centromere was present, to minimise scoring G₂ and S phase nuclei where replication had already occurred. Similarly, when nuclei from females were scored, only those with two visible X centromere signals were scored. However, as the whole X chromosome does not replicate simultaneously, when cells are in S phase it is likely that the region of interest may have replicated before the X centromere, especially as the X centromere region in males has been reported to replicate late during S phase (Ten Hagen *et al.*, 1990). It is therefore possible that if nuclei are in early S phase a pattern of hybridisation may be observed that is consistent with a duplication of the target sequence (such as a single X centromere signal and a pair of signals from the target sequence), which may in fact just be a result of replication of the target region prior to replication of the centromere. However, it has been reported that the X chromosome as a whole on average replicates in the later stages of S phase so this may reduce the numbers of pseudo-duplicated signals seen (Woodfine *et al.*, 2004). Where replication of a bona fide duplicated region has occurred before replication of the X centromere sequence, this could result in signal patterns such as four closely associated signals from the genomic clone probe (when both copies of the duplicated sequence have replicated) or three copies of a probe signal (where just one of the two duplicated sequences has been replicated).

2.2.9. MAPH

2.2.9.1. Probe design

Probe sequences were chosen from publicly available sequence from the human genome project, genome databases, and from exons of the *PLP1* gene (Figures 2.3. and 2.4.). Primer design was carried out using Primer3 (Table 2.3.). Each potential probe sequence was compared against sequences at the NCBI using BLASTn (see section 2.2.10.1.1.). Only sequences with no similarities greater than 30 nucleotides in length to other submitted sequences were used. Probe G+C content varied from 39.6-54%. The size range of probes was from 244 - 390bp and each probe differed from the nearest one in size by at least 4 base pairs (Table 2.2.).

2.2.9.2. PCR of probe target sequence

Probe sequences were amplified using PCR (for conditions see sections 2.1.3., 2.2.1., and Appendix A) from normal human genomic DNA, and *Xenopus* genomic DNA for the non-human control probe. The PCR products were run on a 2% agarose gel, and the bands were cut out the DNA extracted from the gel (see section 2.2.1.4.4.).

2.2.9.3. Cloning MAPH probes

The PCR products from the target sequences were cloned into the *EcoRV* site of the pZErO-2 plasmid (Figure 2.2.). This disrupts a fusion gene carried on the plasmid, *lacZα-ccdB*, which is lethal to *E.coli* when expressed (Bernard *et al.*, 1994). pZErO-2 also contains a kanamycin resistance gene, so after transformation only those cells carrying the plasmid, but with the lethal *lacZα-ccdB* gene disrupted, should be able to grow on media containing kanamycin.

2.2.9.4. Removal of 3' A residues

To remove extra untemplated 3' A residues from PCR products and allow blunt-end end cloning, the purified products were incubated with Klenow polymerase which has 3' to 5' exonuclease activity. 14µl H₂O, 5µl 10 x EcoPol buffer and 1µl Klenow fragment were added to 30µl gel-purified PCR product (see section 2.2.1.4.4.). Following incubation at 37°C for 5 minutes, 4pmol of each dNTP was added to the mixture, which was then incubated at room temperature for 20 minutes. Following treatment with Klenow DNA polymerase, DNA was purified by phenol extraction (see section 2.2.3.1.) and ethanol precipitated (see 2.2.3.2.). The pellet was completely dried and then resuspended in 15µl 1 x TE.

2.2.9.5. Ligation and transformation

2.2.9.5.1. Vector digestion

The plasmid vector p-ZErO-2 was linearised by digestion with the blunt-end cutting restriction enzyme, *EcoRV*. 1µg supercoiled p-ZErO-2 was digested with 20U *EcoRV* in a total volume of 10µl with a 1 x concentration of NEB buffer 3 and 100µg/ml BSA. The linearised vector was extracted using phenol/chloroform/isoamyl alcohol, ethanol precipitated (see sections 2.2.3.1-2) and resuspended in 90µl TE buffer, giving a concentration of 10ng/µl.

2.2.9.5.2. Ligation

Ligations were set up using 10ng of linearised pZErO-2, 7.5µl of gel-purified insert PCR product, and 2U T4 DNA ligase, with a 1 x concentration of T4 ligase buffer in a total volume of 10µl. Ligations were incubated at 16°C for one hour.

2.2.9.5.3. Transformation

2µl from each ligation reaction was chemically transformed into *E.coli* TOP10 chemically competent cells. For each ligation to be transformed, one 50µl vial of One Shot TOP10 cells was thawed on ice. 2µl from each ligation reaction was added directly to each vial of cells, which were mixed by tapping gently and then incubated on ice for 20 minutes. The cells were incubated at 42°C in a water bath for 30-45 seconds and then removed and swiftly placed on ice. 250µl of SOC medium was added to each vial, and the transformations were incubated at 37°C with shaking at 225rpm for one hour, with the vials placed on their side. 50µl of each transformation was spread on an LB-kanamycin agar plate, with a kanamycin concentration of 25µg/ml, and incubated at 37°C overnight. Controls used in the transformation procedure were a cells only plate, no DNA in the ligation reaction, a vector-only ligation, and a ligation using test inserts provided with the kit. The following day, colonies containing recombinant pZErO-2 were grown in 5ml LB broth with 25µg/ml kanamycin antibiotic selection at 37°C with shaking overnight, and then the plasmids were recovered by alkaline lysis mini-preparation (see section 2.2.5.2.). If a colony was found to contain the required recombinant plasmid, a glycerol stock was made (see section 2.2.5.1.).

2.2.9.6. Probe preparation

PCR was carried out on the recombinant plasmids using primers that flanked the insertion site, PZA and PZB (see section 2.2.1.). The structures and orientation of the inserts within the probes were checked using restriction digestion with one or two enzymes (Table 2.2.). 7µl of PCR product was digested in a reaction mix containing 1.5µl restriction enzyme and 1.5µl of the appropriate 10 x enzyme buffer in a total volume of 15µl, by incubating at 37°C for at least 2 hours. Restriction digested PCR products were

electrophoresed alongside undigested PCR products on an agarose gel together with a 100bp ladder size standard, to check that the fragments produced following restriction digestion were of the expected sizes (Table 2.2.). PCR products were gel-extracted to purify the MAPH probes (see section 2.2.1.4.4.). Gel-extracted probes were electrophoresed in an agarose gel along with DNA of known concentration, to estimate the concentration of each probe. The probes were combined together to make a probe mix, containing approximately 100pmol of each probe in 1µl of probe mixture.

Probe	Probe size (bp)	G+C (%)	Enzyme(s)	Sizes expected after digest
ch7q31	244	44.9	<i>MvaI</i>	129+115 / 140+104
			<i>BamHI</i>	124+120 / 135+109
XLnkx	253	48.7	<i>MvaI</i>	106+146 / 117+135
			<i>MwoI</i>	169+83 / 180+72
plp5	258	45.7	<i>MvaI</i>	112+146 / 101+157
144a10	264	43.4	<i>HinFI</i>	144+120 / 155+109
ch17p13	268	44	<i>RsaI</i>	185+83 / 174+94
240c2	273	50	<i>AluI</i>	132+141 / 143+130
79p11	282	40.4	<i>AluI</i>	170+112 / 159+123
ch1q24	289	46	<i>MwoI</i>	108+181 / 97+192
plp6	298	41.8	<i>MvaI</i>	111+187 / 100+198
			<i>RsaI</i>	254+44 / 243+55
ch6p24	322	45.6	<i>MspI</i>	212+110 / 201+121
SRY	330	50.2	<i>MvaI</i>	77+253 / 88+242
ch4q26	335	44	<i>RsaI</i>	189+146 / 200+135
Xq12	340	48	<i>EcoT14I</i>	113+227 / 102+238
			<i>BamHI</i>	253+87 / 242+98
198p4	352	39.6	<i>AluI</i>	240+112 / 251+101
43h13	359	46.3	<i>AluI</i>	34+165+106+54 / 34+165+95+65
plp3	368	54	<i>MwoI</i>	210+158 / 199+169
plp7	381	49	<i>EcoT14I</i>	237+124 / 246+135
			<i>MwoI</i>	311+70 / 300+81
ch17q21	390	51	<i>RsaI</i>	133+277 / 122+268

Table 2.2. Table showing the panel of MAPH probes used in the PMD probe set. Sizes of cloned and amplified probes are given in base pairs. The restriction enzymes that were used to confirm the structure of the probes, together with the fragment sizes that were expected, are also shown.



Figure 2.2. Features of the pZErO-2 plasmid. The multiple cloning site is near the start of the *lacZ α -ccdB* fusion gene, and the position of the kanamycin resistance gene is also shown. Taken from the Zero Background/Kan Cloning kit manual (Invitrogen).

2.2.9.7. MAPH protocol

2.2.9.7.1. DNA and filter preparation

1µg genomic DNA was denatured by heating to 95°C then placed on ice. 1µl 1M NaOH was added to the genomic DNA and the mixture was spotted onto an individual nylon filter (Hybond N+) with approximate dimensions 2mm x 4mm. Cutting the filters into unique shapes identified individual filters. The denatured genomic DNA was irreversibly bound to the filter by UV irradiation, 50mJ to each side of the filter (UV Stratalinker 2400, Stratagene). The filters were then prehybridised together in 1ml of prehybridisation solution at 65°C for at least 2 hours. This prehybridisation solution was replaced with 200µl of prehybridisation solution with the addition of 10µg/ml *Cot*-1 DNA that had been boiled for two minutes, and incubated at 65°C for 30-60 minutes.

2.2.9.7.2. Hybridisation

1µl of probe mix was mixed with 1µg *Cot*-1 DNA, 7µg *E.coli* DNA that had been digested with *Hae*III, 0.5µg ϕ X174/*Hae*III, 20pmol of each blocker primer PZAX and PZBX, and denatured by addition of 2µl 1M NaOH. The mixture was heated to 37°C for 1 minute, neutralised by the addition of 3µl 1M NaH₂PO₄, and then added to the hybridisation, which was left at 65°C overnight.

2.2.9.7.3. Washes

Filters were then washed to remove unbound and non-specifically bound probes, initially with two 1ml changes of hybridisation solution, after which the filters were transferred to a 50ml tube. Washing was carried out using a total of 500ml 1 x SSC/1% SDS at 65°C for 15-20 minutes, then using a total of 500ml 0.1 x SSC/0.1%SDS for the next 30-40 minutes at 65°C, with frequent changing of the wash solution. Following the washes,

filters were air-dried and transferred into 50µl PCR reactions and amplified for 5 cycles of 95°C for one minute, 60°C for one minute and 70°C for one minute with PZA and PZB primers. 1µl from this “primary” PCR was used to seed a secondary 25µl PCR with FAM labelled PZA and PZB, for 20 cycles. To increase signal on the gel, 20µl from the secondary PCR was precipitated by adding 20µl 0.4M NaCl and 100µl 100% ethanol. This was left at room temperature for 15 minutes, then spun at 13000rpm in a microcentrifuge for 10 minutes. Excess ethanol was removed and the pellet was air-dried, then 3µl loading buffer, containing Genescan 500 Tamra size standard, was used to redissolve the pellet. The loading mixture was denatured by heating to 95°C, and then placed on ice. 2µl of the precipitated and resuspended product was electrophoresed on a 4% polyacrylamide gel using an ABI 377 DNA sequencer and analysed with Genotyper software (see section 2.2.2.1).

2.2.9.8. MAPH Analysis

For analysis of each DNA sample, the area underneath the peaks of each probe was used to obtain normalised ratios for the probes. Initially, the area under each peak was divided by the sum of the nearest sized 4 autosomal control peaks. In later analyses the area under each peak was divided by the sum of four autosomal control probe peaks, but not using the two control probes with the highest variability, ch7q31.2 and ch17q21.32, in calculating these ratios. An average ratio for each probe was calculated across all samples for the normal control individuals for each experiment. Normalised ratios were then calculated for each probe for each sample by dividing the ratio for each peak by the corresponding mean ratio. For autosomal probes a normalised ratio of 1 should correspond to a diploid copy number. For probes whose target sequences were located on the X chromosome, the mean ratios were calculated taking into account the different copy numbers of the X chromosome in males and females. For an X-linked probe, if the mean

ratio was being calculated from a panel of three males and three females, the total was divided by 4.5, to give the mean ratio expected for two X chromosomes. Hence for the individual X-linked normalised ratios, a value of 1 indicates that there are two copies of that sequence present, and a value of 0.5 is expected if there is only one copy of the sequence, as for normal males. For Y-linked probes, the mean ratio was calculated using only the control males, so a value of 1 corresponds to there being one copy of the target sequence present.

2.2.10. Bioinformatics

All the websites used for analysis of sequence characteristics and other bioinformatic applications are listed in Table 2.3.

2.2.10.1. Sequence database searches and pairwise sequence comparisons

BLAST-based (Basic Local Alignment Search Tool) sequence comparison methods were used for most sequence comparisons, either when searching for sequence similarities to one query sequence within a sequence database, or when comparing two sequences to one another (Altschul *et al.*, 1990).

2.2.10.1.1. BLASTn

BLASTn was used to search databases for significant similarities to a query sequence (Altschul *et al.*, 1990; Altschul *et al.*, 1997). In most instances, BLASTn was used to search the nr database at NCBI, which contains all GenBank, RefSeq Nucleotides, EMBL, DDBJ and PDB sequences. In most cases, the BLASTn search was restricted to sequences of human origin within the nr database, and default settings were used (low complexity filter, expect = 10, word size = 11).

2.2.10.1.2. BLAST2

BLAST 2 Sequences (BLAST2) was used for pairwise sequence comparisons, either between two nucleotide sequences (blastn) or between one nucleotide and one protein sequence (tblastn) (Tatusova and Madden, 1999). Default parameters were used unless otherwise stated (match = 1, mismatch = -2, open gap = -5, gap extension = -2, gap_x dropoff = 50, expect = 10, wordsize = 11, filter on).

2.2.10.1.3. BLASTz

BLASTz is an algorithm that is designed for aligning long DNA sequences, using a gap scoring system that tolerates longer gaps in aligned sequences than most sequence alignment tools (such as BLASTn) (Schwartz *et al.*, 2000). Output from a BLASTz based comparison is given as both a percentage identity plot and a dotplot-type output (Schwartz *et al.*, 2000; Ovcharenko *et al.*, 2004). For all of the BLASTz comparisons in this study the output was visualised in the form of a dotplot, which shows an overview of the relationships between the two sequences by plotting one sequence on the x axis and one on the y axis. Any significant alignments between any regions of the two sequences are shown as shaded areas on the dotplot, with short regions of similarity visible as dots and longer homologies shown as diagonal lines. Two web-based BLASTz comparison tools were used in this study, Pipmaker and zPicture (Schwartz *et al.*, 2000; Ovcharenko *et al.*, 2004). In all comparisons, default settings were used (match = 1, mismatch = -1, gap open = -6, gap extension = -0.2) (Schwartz *et al.*, 2003).

2.2.10.1.4. ClustalW and Consensus

ClustalW was used for producing nucleotide sequence alignments, and a consensus sequence was derived from the ClustalW alignment where necessary using the Consensus program (Thompson *et al.*, 1994). ClustalW is a multiple alignment method which uses a

distance matrix based on individual alignments to calculate a guide tree, then the sequences are progressively aligned according the branching order of the guide tree (Thompson *et al.*, 1994). The Consensus program uses the multiple alignment output from ClustalW to calculate a consensus sequence.

2.2.10.2. Repeatmasker

Repeatmasker was used to mask out common interspersed repetitive elements from genomic DNA sequences. In most cases the program was run from the EMBL website, using Repeatmasker version 2002/07/13, using Repbase update version 7.4 (Table 2.3.) (Jurka, 2000). When large amounts of repeat masked genomic sequence was needed, the appropriate region was downloaded from the UCSC genome browser after repetitive sequence was masked based on June 2003 versions of Repeatmasker and the repeat libraries.

2.2.10.3. DNA Pattern Finder

Short sequence motifs were searched for by the DNA Pattern Find program, which can search for several different sequence motifs simultaneously in a given DNA sequence, providing an output showing the position and orientation of each occurrence of each motif (Table 2.3.).

2.2.10.4. MAR-Wiz

Potential matrix attachment regions (MARs) in genomic DNA sequence were searched for using the MAR-Wiz program, which searches for various sequence characteristics associated with MARs and combines these into a MAR-potential value which is calculated within a sliding window throughout the region in question (Singh *et al.*, 1997). The parameters used to calculate the MAR-potential are: motifs associated with origins of

replication, T+G richness, the presence of sequences known to curve or kink DNA, topoisomerase II recognition sites, A+T content, the presence of an MAR consensus sequence motif, and the presence of stretches of 20 or more A, T or C nucleotides (Wang *et al.*, 1995; Singh *et al.*, 1997)

2.2.10.5. Oligorep

Repeated regions (direct, inverted, symmetric and complementary repeats) within a DNA sequence were found using the Oligorep program, with the minimum length of repeats set at 6bp, maximum at 30bp, and a maximum mismatch of 4 (Babenko *et al.*, 1999).

2.2.10.6. Tandem repeats finder

Tandemly repeated sequences were predicted using the Tandem repeats finder program, using the default settings (Benson, 1999).

2.2.10.7. Genome browsers

Frequent use of genomic databases was made during this study, for applications such as checking the position of genes and genomic clones within the human genomic sequence, interspersed repeat and G+C content, recombination rates, the position of polymorphic markers, and comparison of syntenic regions between species. The Ensembl genome browser was most frequently utilised for these purposes, with substantial use also made of the UCSC genome browser (Table 2.3.) (Karolchik *et al.*, 2004; Birney *et al.*, 2004).

Program	Website address
BLASTn	http://www.ncbi.nih.gov/BLAST/Blast.cgi
BLAST2	http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html
BLASTz	http://pipmaker.bx.psu.edu/cgi-bin/pipmaker?basic http://zpicture.dcode.org/
ClustalW	http://www.ebi.ac.uk/clustalw/
Consensus	http://www.bork.embl-heidelberg.de/Alignment/consensus.html
Repeatmasker	http://woody.embl-heidelberg.de/repeatmask/
DNA Pattern Finder	http://bioinformatics.ccr.buffalo.edu/ToolBox/seqsuite/dna_pattern.html
MAR-Wiz	http://www.futuresoft.org/MAR-Wiz/
Oligorep	http://wwwmgs.bionet.nsc.ru/mgs/programs/oligorep/InpForm.htm
Primer3	http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi
Ensembl	http://www.ensembl.org
UCSC genome browser	http://genome.ucsc.edu
Rebase Update	http://www.girinst.org/Rebase_Update.html
Sanger Institute	http://www.sanger.ac.uk
Tandem repeats finder	http://tandem.bu.edu/cgi-bin/trdb/trdb.exe?taskid=1
NCBI homepage	http://www.ncbi.nih.gov/

Table 2.3. Websites of bioinformatic tools used during the course of this study.

Motif	Sequence(s)	References
χ element	GCTGGTGG	(Smith <i>et al.</i> , 1981)
Ade-M26 heptamer	ATGACGT	(Schuchert <i>et al.</i> , 1991)
LTR-IS	TGGAAATCCCC	(Edelmann <i>et al.</i> , 1989)
Retrotransposon LTR	TCATACACCACGCAGGGGTAGAGGACT	(Zimmerer and Passmore, 1991)
XY32	AAGGGAGAARGGGTATAGGGRAAGAGGGAA	(Rooney and Moore, 1995)
Human minisatellite core	GGGCAGGARG	(Jeffreys <i>et al.</i> , 1985)
Human hypervariable minisatellite recombination sequence	GGAGGTGGGCAGGARG (1), AGAGGTGGGCAGGTGG (2)	(Jeffreys <i>et al.</i> , 1985; Wahls <i>et al.</i> , 1990)
Pur binding site	GGNNGAGGGAGARRRR	(Bergemann and Johnson, 1992)
Translin binding sites	GCNCWSSWN ₀₋₂ GCCCWSSW (1), MTGCAGN ₀₋₄ GCCCWSSW (2), ATGCAG (3), GCCCWSSW (4)	(Aoki <i>et al.</i> , 1995)
Human replication origin consensus	WAWTTDDWWWDHWGWHMAWTT	(Dobbs <i>et al.</i> , 1994)
<i>S. cerevisiae</i> ARS	WTTTATRTTTW	(Broach <i>et al.</i> , 1983)
<i>S. pombe</i> ARS	WRTTTATTAW	(Maundrell <i>et al.</i> , 1988)
Scaffold attachment region consensus	AATAAAYAAA (1), TTWTWTTWTT (2), WADAWAYAWW (3), TWWTDTTWWW (4)	(Gale <i>et al.</i> , 1992)
Topoisomerase II binding site	GTNWAYATTNATNNR	(Sander and Hsieh, 1985)
tetranucleotide repeats from mouse MHC recombination hotspots	(TCTG) ₄₋₆ (1), (CAGG) ₇₋₉ (2)	(Kobori <i>et al.</i> , 1986; Uematsu <i>et al.</i> , 1986)
Vertebrate/plant topoisomerase I consensus cleavage sites	CAT (1), CTY (2), GTY (3), RAT (4)	(Been <i>et al.</i> , 1984)
Vaccinia topoisomerase I consensus cleavage site	YCCTT	(Shuman, 1991)
Vertebrate topoisomerase II consensus cleavage site	RNYNNCNGYNGKTNINY	(Spitzner and Muller, 1988)
Heptamer recombination signal	CACAGTG	(Early <i>et al.</i> , 1980)
Nonamer recombination signal	ACAAAAACC	(Early <i>et al.</i> , 1980)
Immunoglobulin heavy chain class switch repeats	GAGCT (1), GGGCT (2), GGGGT (3), TGGGG (4), TGAGC (5)	(Rabbitts <i>et al.</i> , 1981; Ohno, 1981)
Human minisatellite conserved sequence/ χ -like sequence	GCWGGWGG	(Krowczynska <i>et al.</i> , 1990)
Mariner transposon-like element (3' end)	GAAAATGAAGCTATTTACCCAGGA	(Reiter <i>et al.</i> , 1996)
Deletion hotspot consensus sequence	TGRRKM	(Krawczak and Cooper, 1991)
Mouse parvovirus recombination hotspot	CTWTTR	(Hogan and Faust, 1986)
Murine MHC deletion hotspot	(CAGR) _n	(Steinmetz <i>et al.</i> , 1986)
Murine LTR recombination hotspot	TGGAAATCC	(Edelmann <i>et al.</i> , 1989)
DNA polymerase α pause site core sequences	GAG (1), ACG (2), GCS (3)	(Weaver and DePamphilis, 1982)
DNA polymerase arrest site	WGGAG	(Weaver and DePamphilis, 1982)

Table 2.4. Continued on next page.

Motif	Sequence(s)	References
DNA polymerase α frameshift hotspots	TCCCCC (1), CTGGCG (2)	(Kunkel, 1985b)
DNA polymerase β frameshift hotspots	ACCCWR (1), TTTT (2)	(Kunkel, 1985a)
DNA polymerase α/β frameshift hotspots	TGGNGT (1), ACCCCA (2)	(Kunkel, 1985a;Kunkel, 1985b)
DNA bending	TTTAAA	(Crothers <i>et al.</i> , 1990)

Table 2.4. (continued) Table showing some DNA motifs that have been associated with rearrangements, mutation and cleavage. Where more than one sequence is in the same motif category, each separate motif is identified by the number in brackets after each sequence.

2.3. Clinical details

2.3.1. Family 1

The proband (1:9) was diagnosed at the age of 8 with probable PMD (Figure 2.2.). Nystagmus had been noted during the first month of life, which became less noticeable with age. Development was delayed, particularly motor development and he could crawl and stand at age 2, but he never achieved ambulation. He was wheelchair-bound by age 8 and exhibited dystonic quadriplegia and ataxia. MRI at age 8 showed absence of myelin in much of the brain. He has no speech, but shows reasonable understanding. He has two younger siblings, a sister (1:10) and brother (1:11) who are unaffected and there is no other family history of PMD (Figure 2.2.).

2.3.2. Family 2

The proband (2:9) presented at age 3 with motor developmental delay, ataxia and nystagmus, which had become apparent at about 2 years of age (Figure 2.2.). MRI showed greatly reduced myelination for his age. He has some speech, and can communicate in two-word sentences. There was a possible family history of PMD; his great uncle (2:3) had shown normal development until 2 years old, after which he had shown developmental regression (Figure 2.2.). He was diagnosed with cerebral palsy, and died aged 40. Another great uncle (2:4) had died aged 1, and may have had some motor developmental delay having never sat unsupported. The younger brother (2:10) of the proband is unaffected.

2.3.3. Family 3

The proband (3:6) presented at 13 months of age with slight developmental delay, titubation and some ataxia (Figure 2.2.). Nystagmus had been noticed at 2-3 months of age, and MRI scans showed very little myelination. He has a fairly mild PMD phenotype

and can talk and walk with assistance, although this is slowly deteriorating with time.

There is no previous family history of PMD, and a recent prenatal diagnosis was made for his younger sister (3:7), who was found to carry the same mutation as her brother (Woodward *et al.*, 2003).

2.3.4. Family 4

The proband (4:6) initially presented with short stature at the age of 7 years (Figure 2.2.). He had been noted to have hypoglycaemia and hyponatraemia in the neonatal period. Investigations revealed GH insufficiency with normal TSH, cortisol and gonadotrophin secretion. His MRI scan revealed anterior pituitary hypoplasia with hypoplasia of the infundibulum, an absent posterior pituitary and dysgenesis of the corpus callosum. He has been treated with recombinant human growth hormone and has progressed satisfactorily through puberty unaided. His half-brother (4:7) presented with neonatal hypoglycaemia and was diagnosed as having severe cortisol, TSH, GH and probable gonadotrophin deficiencies. His MRI scan revealed hypoplasia of the anterior pituitary gland and infundibulum with an ectopic/undescended posterior pituitary. The corpus callosum is normal. He has subsequently been treated with GH, hydrocortisone and thyroxine and is currently 2.5 years old. There is no other family history of pituitary disorders known.

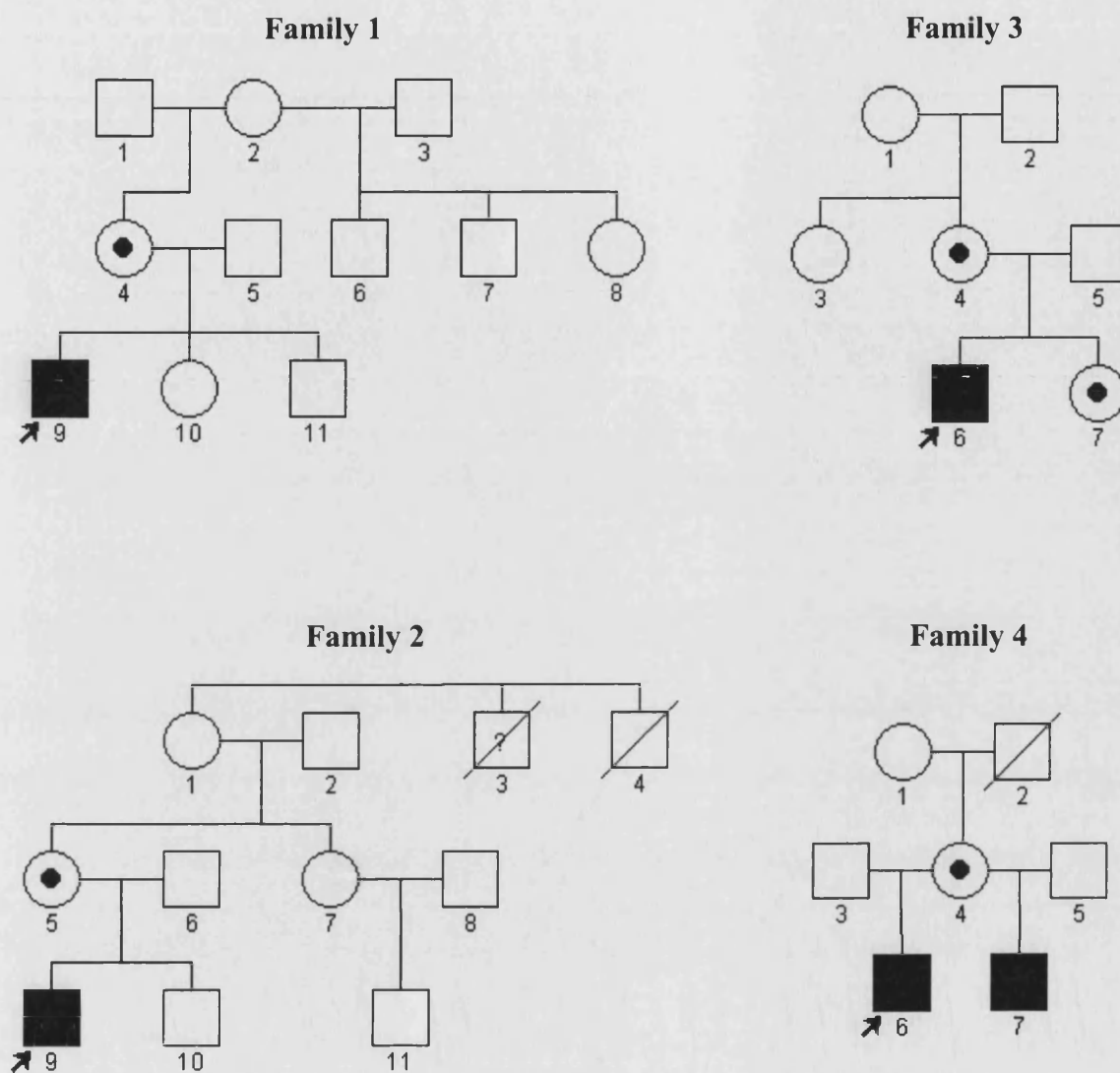


Figure 2.2. Pedigrees showing the families investigated as part of this study.

3.1. *IN SILICO* ANALYSIS OF *PLP1* GENOMIC REGION

Duplications and deletions of dosage sensitive genes, in many genomic disorders, have been found to be mediated by non-allelic homologous recombination between flanking low-copy repeats (see section 1.4.1.) (Lupski, 1998). Following the completion of the sequencing of the human genome, the genomic sequence of the region surrounding *PLP1* is freely available (Lander *et al.*, 2001). To determine if there were any low-copy repeats flanking *PLP1* that could be mediating the *PLP1* duplications and deletions, a detailed analysis of any repeats present in this region was carried out, using a variety of bioinformatic tools.

3.2. Repeats specific to the 2Mb region flanking *PLP1*

Initially, 2Mb of human genomic sequence surrounding *PLP1* was investigated looking for the presence of flanking repeats. Most duplications and deletions involving *PLP1* are shorter than 2Mb, with the duplication endpoints falling within this region (Woodward *et al.*, in preparation). Therefore it seems most likely that any sequence features involved in producing genomic rearrangements involving *PLP1* will be within this region, close to the gene. 2Mb from the human genomic sequence, centred on the transcription start site of *PLP1*, and masked for interspersed repetitive elements by the UCSC genome browser, was compared against itself using the BLASTz algorithm by the Pipmaker program, and the dotplot output is shown in Figure 3.1. (see section 2.2.10.1.3.) (Schwartz *et al.*, 2000). Although no obvious large repeat regions were found on either side of *PLP1*, it was apparent that there were some local repetitive elements within this 2Mb region (Figure 3.1.). Most of the repetitive regions shown on the dotplot were found either only proximal to *PLP1*, mostly within Xq22.1 (bottom left-hand quadrant of the dotplot), or distal to *PLP1* within Xq22.2 (top right-hand quadrant of the dotplot in Figure 3.1.). There were a few very small regions of similarity on both sides of *PLP1* (Figure 3.1.). One of the repeat regions either side of

PLP1 (indicated by an oval outline on the dotplot) appeared to have similarities with parts of a large block of mainly distal-specific repeats (see section 3.4.4., Figure 3.4., Tables 3.2. and 3.3.).

3.3. Repetitive sequences flanking *PLP1*

Further investigations of the small repetitive regions found proximal and distal to *PLP1* were carried out using BLASTn, BLAST2 and also using more dotplots of the relevant regions from Pipmaker using BLASTz (see sections 2.2.10.1.1-3.) (Altschul *et al.*, 1990; Tatusova and Madden, 1999; Schwartz *et al.*, 2000; Schwartz *et al.*, 2003). Comparison was also made with chained BLASTz alignments annotated on the UCSC genome browser for verification of repeated regions (Karolchik *et al.*, 2003; Schwartz *et al.*, 2003). Chained alignments have been derived by comparing the human genome against itself using BLASTz, after which trivial alignments of a sequence to itself are filtered out of the chained alignment, along with known interspersed repetitive elements (Karolchik *et al.*, 2004).

3.3.1. Repeated flanking sequences close to *PLP1*

One repeated region was located relatively close to *PLP1*, with two copies present 69Kb apart and *PLP1* located in between (the similarity between these two regions is picked out by the diamond outline in Figure 3.1.). Combining the information generated from all these sources showed that the extent of the similarity found between the sequence at the proximal end in genomic clone cU35G3, extended for over 3Kb, although the regions of similarity were fragmented and only 555 nucleotides were included in the alignment in total (Figure 3.2. and Table 3.1.). The distal repeat sequence was similarly fragmented, and although it covered nearly 3Kb of sequence in human genomic clone cV698D2, just 585bp was included in the alignments (Figure 3.2., Table 3.1.). Much of the unaligned intervening sequence

consisted of interspersed repeat sequences, with 73.29% of the proximal repeat region mapping to cU35G3 and 50.26% of the distal repeat sequence from cV698D2 classed as interspersed repetitive elements by Repeatmasker (Repbased update version 7.4). Of the nucleotides actually included in the alignment, there was 87% identity between the two sequences. The two copies of this repeated sequence were inverted with respect to each other. Database searches using BLASTn did not reveal any similarities between the rest of the human genome and these short repeated sequences.

3.3.2. More distant repeated sequences either side of *PLP1*

There were two other very small hits seen using Pipmaker between the regions proximal and distal to *PLP1* (regions of similarity on the dotplot are indicated by the grey circle and square in Figure 3.1.). The exact position of these two regions was checked by carrying out further comparisons using Pipmaker between individual genomic clones in the regions where the similarity was found, and verified by BLAST2 comparisons. Both of these matches found by BLASTz and shown on the dotplot were very short and did not have very high percentage identity between the two sequences (Figure 3.1., Table 3.1.). Both of these small repeats were inverted in orientation. As both of these regions, which had been picked up by the BLASTz algorithm used by Pipmaker, had fairly low levels of similarity and were very short the BLAST2 comparisons were unable to confirm the alignments initially, but after adjustments of the parameters from the default values to make them less stringent (Match:2; wordsize:7), BLAST2 was also able to detect the alignments (see section 2.2.10.1.2.).

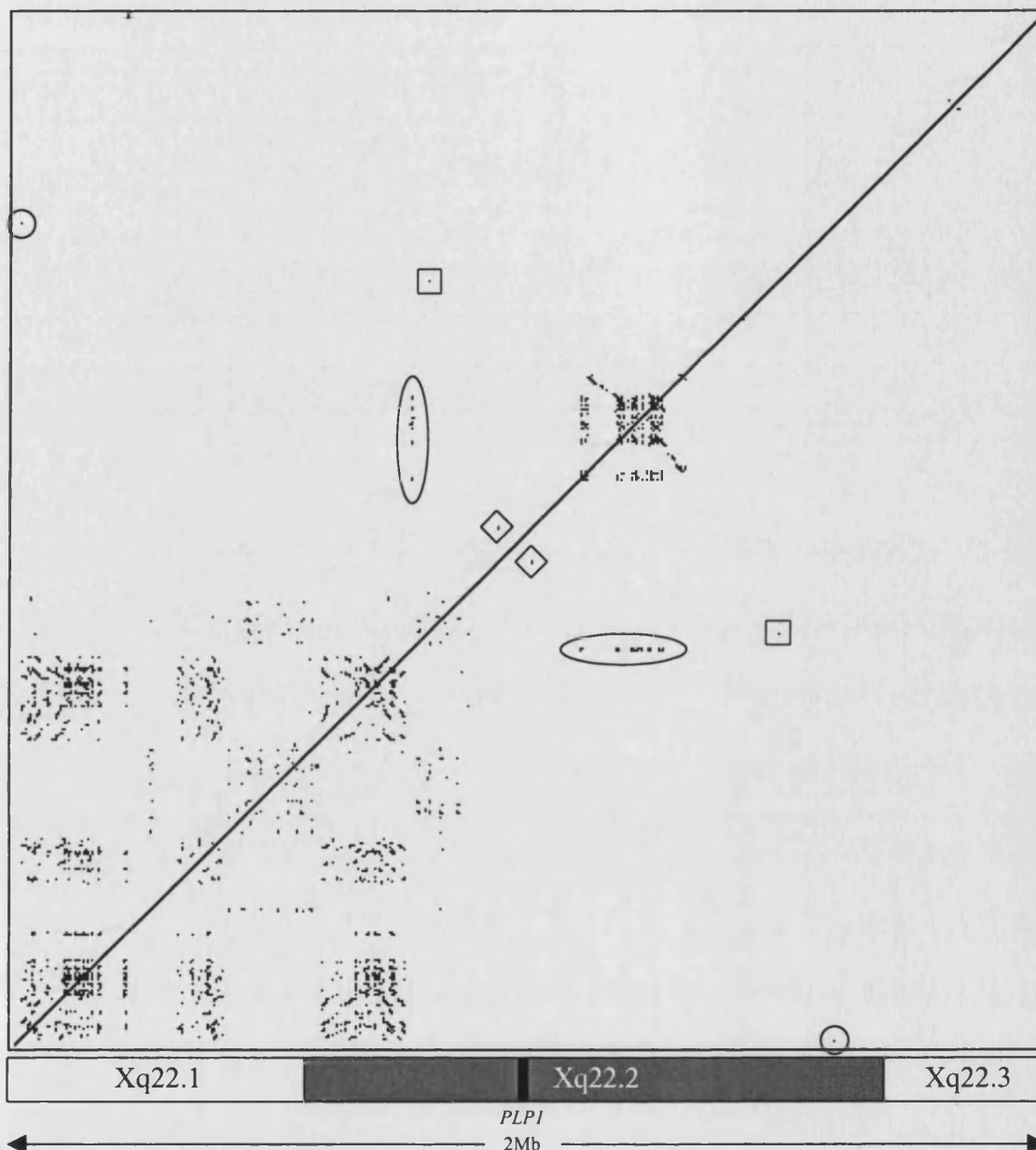


Figure 3.1. Region-specific repeats in the genomic region surrounding *PLP1*. 2Mb of human genomic sequence was masked for interspersed repetitive elements (UCSC genome browser, July 2003 release) and compared against itself using Pipmaker (BLASTz), with the resulting dotplot shown above. Matches are shown as black areas on the dotplot. Repeats within this region are shown as diagonal lines on the dotplot, upwards-facing diagonals (/) indicate directly repeated sequences, and inverted repeated sequences are downwards-sloping diagonals (\). The small regions of similarity found either side of *PLP1* are identified by the grey outlines. The approximate location of the chromosome bands in the region are shown underneath the dotplot, and the position of the *PLP1* gene is indicated by the black box. Regions of similarity either side of *PLP1* are outlined (see Table 3.1.).

Clone	Position within clone	Total size	% identity	Distance from <i>PLP1</i>
◇ cU35G3	22975-26171bp	3197bp	17.3%	(+) 32.4Kb
◇ cV698D2	36996-39898bp	2903bp	20.2%	(-) 30.6Kb
□ cV857G6	9469-9554bp	86bp	68.6%	(+) 154.0Kb
□ dJ513M9	111144-111231bp	88bp	67.0%	(-) 513.6Kb
○ bA522L3	26638-26549bp	90bp	65.6%	(+) 969.2Kb
○ bA541I12	102438-102527bp	90bp	65.6%	(-) 622.9Kb

Table 3.1. Summary of the short regions of sequence similarity identified by BLASTz either side of *PLP1*. The symbols next to each entry correspond to the ones used to highlight these regions in Figure 3.1. The position of each repeat sequence within the sequence for each genomic clone is shown, and the percentage of identical nucleotides within each alignment is also shown. The distance from the transcription start site of *PLP1* is indicated, (+) is next to sequences proximal to *PLP1*, (-) is next to the distal sequence.

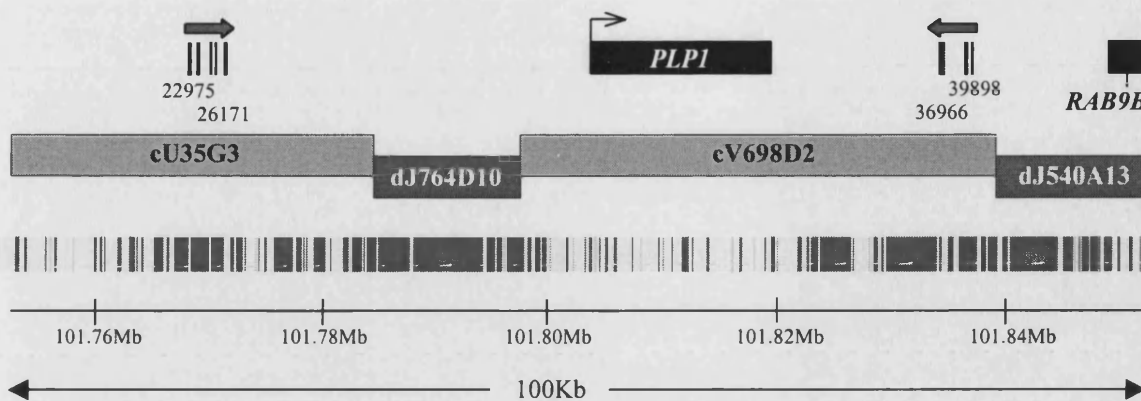


Figure 3.2. Position of small inverted repeat sequences just either side of *PLP1*. 100Kb of human genomic sequence including the *PLP1* gene is shown, based on data from the Ensembl human genome browser. Distances shown are in Mb from the Xq telomere (Ensembl release 19.34b.2). Interspersed repeat content of the region is shown by the grey bars above the scale bar, and is taken from the Ensembl genome browser. Sequenced human genomic clones are shown (grey boxes) as well as the positions of genes in the region, represented by black filled boxes (*PLP1* and the 3' UTR of *RAB9B*). The location of the short sections of inverted repeats are shown as grey arrows above the genomic clones, with black lines underneath the arrows indicating where the actual regions of similarity lie. The start and finish of the regions of similarity is also shown as the distance in base pairs from the start of the appropriate genomic clone, underneath the arrows representing the position of the repeats.

3.4. Repeats distal to *PLP1*

Within the 1Mb of sequence distal to *PLP1*, a 200Kb section starting approximately 125Kb telomeric to *PLP1* appeared to contain several repeats as shown by the initial dotplot of the whole region (Figure 3.1.). Further analysis of this repetitive region was carried out, initially creating a dotplot of a 250Kb region containing all the repeats (Figure 3.3.). This showed that there was a complicated repeat structure, comprising both inverted and directly repeated regions (Figure 3.3.). Further analysis, using BLASTn, BLAST2, BLASTz, ClustalW, and Repeatmasker was carried out to determine the boundaries of the various repeat elements (see sections 2.2.10.1.1-4. and 2.2.10.2.) (Thompson *et al.*, 1994).

3.4.1. Previously described LCRs distal to *PLP1*

Some of the low-copy repeats within this region have been described as consisting of two LCRs, LCR-PMDA and LCR-PMDB (Figure 3.4.) (Inoue *et al.*, 2002). Each of the LCRs has been found to contain a pair of inverted repeats, A1a and A2 in LCR-PMDA, and A3 and A1b in LCR-PMDB (Figure 3.4.) (Inoue *et al.*, 2002). LCR-PMDA and LCR-PMDB are mainly contained within the human genomic clones dJ839M11 and cU240C2 (Figures 3.3. and 3.4.). The four copies of the repeat unit (A1a, A1b, A2 and A3) that make up the LCRs PMDA and PMDB all show a high degree of sequence identity, particularly the repeats A1a and A1b, which have over 99% of their sequence in common (Table 3.3.). The sequences from repeats A2 and A3 are both between 89-90% identical to repeats A1a and A1b, but only show 62.91% similarity with each other when the percentage identity is calculated from the number of identical bases out of the entire length of the shortest member of the pair (Table 3.3.). However, this low level of identity is partly because these two repeat units only share homology over some of the sequence as they are similar to different overlapping sections of the A1a/A1b consensus (Figure 3.4.). When the percentage similarity between repeats A2

and A3 is calculated just out of the total sequence included in the alignment, not including the most proximal part of both repeats, the percentage identity within the alignment increases to 88.95%. This is similar to the degree of similarity found between either of A2 or A3 and A1a or A1b (Table 3.3.)

3.4.2. Novel distal LCRs

As well as these previously described LCRs, there are additional novel LCRs present in this region. A pair of inverted repeats, here designated LCR-PMDC and LCR-PMDD, which are 35.6Kb and 27.7Kb in length respectively, lie on either side of LCR-PMDA and LCR-PMDB (Figures 3.3. and 3.4., Table 3.2.). This pair of large inverted repeats are also very similar to each other, with over 25Kb of identical sequence shared between the two, as found by comparison with BLASTz (92.43% identity). No annotated genes map to either LCR-PMDC or LCR-PMDD., and both contain a high proportion of interspersed repetitive elements (65.38% and 54.40%, see Table 3.2.). The major differences between these two regions are two regions of sequence within LCR-PMDC that are not present in LCR-PMDD, which both contain numerous interspersed repeats.

3.4.3. Short regions of similarity to the large distal LCRs

There are also a few small regions of similarity to LCR-PMDA and LCR-PMDB lying proximal to the main distal LCR region, but still distal to *PLPI*, within genomic clones bA370B6 and cU116E7 (Figures 3.3. and 3.4., Tables 3.2. and 3.3.).

3.4.4. Sequence similarity between distal repeats and sequence proximal to *PLP1*

A small section of the distal repeats (A1a, A1b and A2) is also present proximal to *PLP1* (Figures 3.1. and 3.4., Tables 3.2. and 3.3.). The matched sequence similarities between the short proximal sequence and the distal repeats is marked by the oval outlines on the dotplot shown in Figure 3.1. The similarity extended for just 104bp, with 73.08% identity, within sequence contained in human genomic clone cV857G6 (Figures 3.1. and 3.4., Table 3.2.). This short sequence was located approximately 164Kb proximal to the start of the *PLP1* gene.

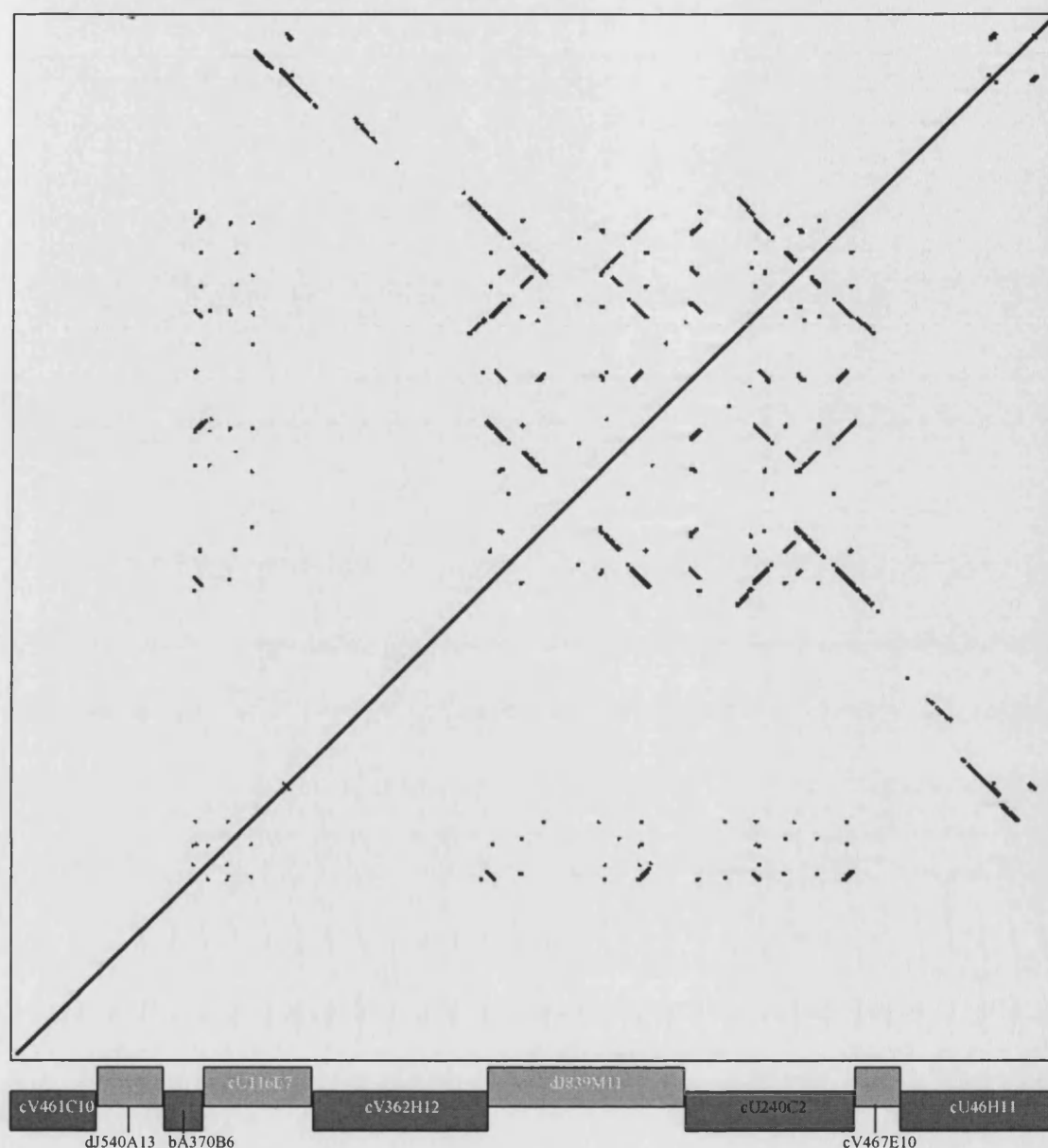


Figure 3.3. Dotplot (using Pipmaker software) comparing 250Kb of human genomic sequence, starting 70Kb distal to *PLP1* and including the large repetitive region, against itself. Repeats within this region are shown as diagonal lines on the dotplot, upwards-facing diagonals (/) indicate directly repeated sequences, and inverted repeated sequences are downwards-sloping diagonals (\). The sequence was masked for interspersed repeats and obtained from the UCSC genome browser (July 2003 version). The human genomic clones in the region are aligned underneath the dotplot.

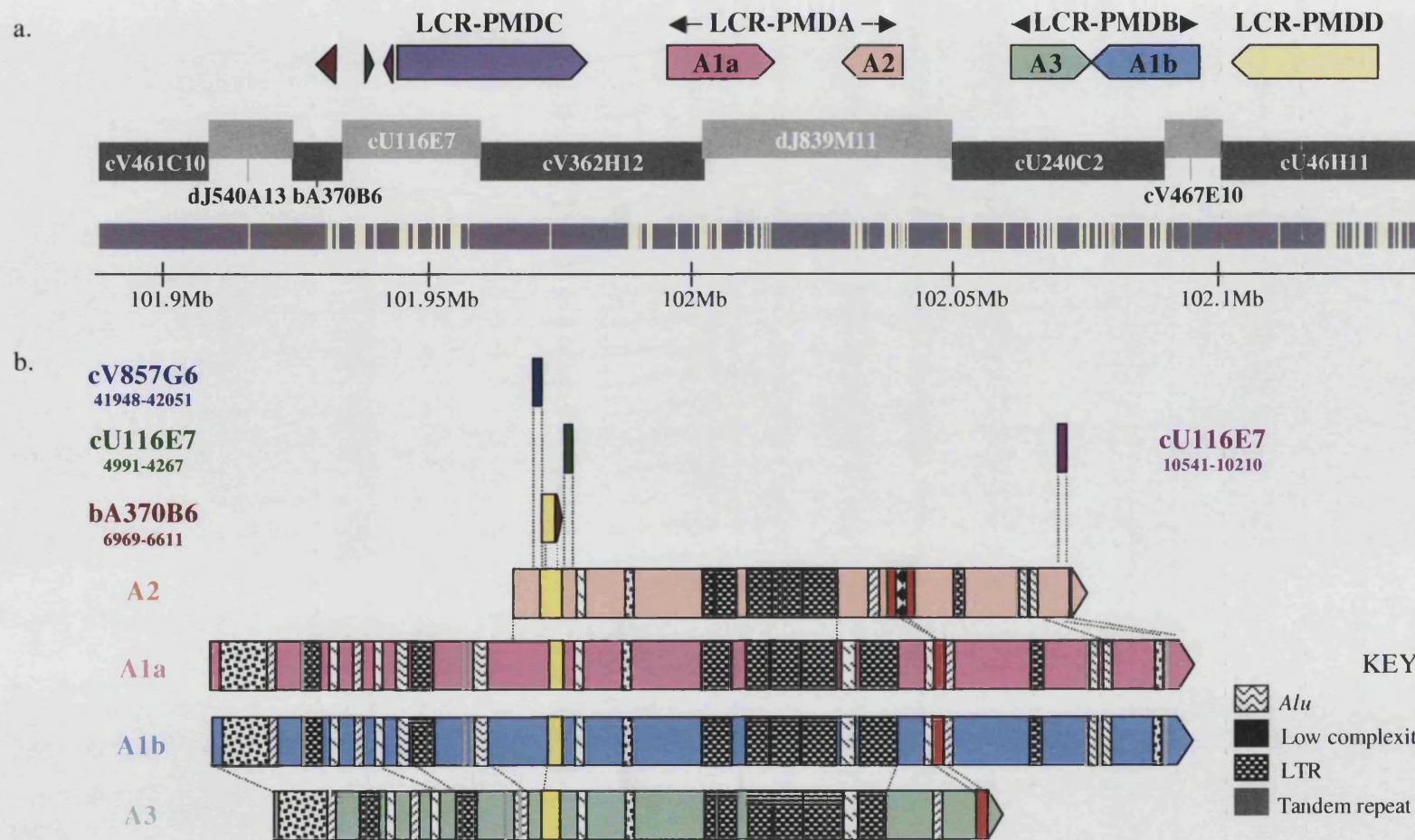


Figure 3.4. Organisation of repeats distal to *PLP1*. For legend, see next page

Figure 3.4. Legend.

(a) shows the position of the various repeat elements relative to genomic clones in the region. Interspersed repeats (as annotated for this region on Ensembl) are shown above the scale bar as grey bars. (b) shows all the separate repeats within LCRs PMDA and PMDB aligned together, with dotted lines between the repeats to more clearly define the alignment. The position of interspersed repetitive elements in each repeat is shown by the different patterned boxes, the H2b-like genes are shown in yellow and the Xrep enhancer-like sequences are shown in red.

Repeat/ Clone	Size (bp)	Start point in (clone)	End point in (clone)	G+C content	Interspersed repeat content	Orientation
cV857G6	104	41948	42051	53.85%	0%	-
bA370B6	356	6611	6969	58.15%	0%	-
cU116E7	277	4991	5267	38.71%	29.60%	+
cU116E7	332	10210	10541	38.55%	56.63%	-
PMDC	35585	10858 (cU116E8)	19867 (cV362H12)	41.84%	65.38%	+
A1a	20338	35373 (cV362H12)	13821 (dJ839M11)	43.78%	41.31%	+
A2	11751	26748 (dJ839M11)	38498 (dJ839M11)	47.43%	33.69%	-
A3	14858	28716 (cU240C2)	13859 (cU240C2)	44.43%	46.63%	+
A1b	20340	13858 (cU240C2)	6580 (cV467E10)	43.75%	41.15%	-
PMDD	27712	2029 (cU46H11)	29740 (cU46H11)	42.46%	54.40%	-

Table 3.2. Summary of the repeated regions distal to *PLP1*. The positions of the start and end of each repeat are given from the start of the genomic clone in which it is contained. Where repeats start and finish in sequence covered by different genomic clones the name of the relevant clone is given underneath the position within that clone. Interspersed repeat content is as found using Repeatmasker. (+) orientation is going from centromere to telomere, (-) is the opposite direction.

							cV857G6 (41948-42051)
						bA370B6 (6969-6611)	–
					cU116E7 (4991-5267)	–	–
			cU116E7 (10541-10210)	–	–	–	–
	A1a	79.52% (264)	69.31% (192)	76.40% (272)	73.08% (76)		
	A2	89.69% (10539)	74.70% (295)	71.84% (199)	78.93% (281)	73.08% (76)	
	A3	62.91% (7393)	89.55% (13305)	–	71.84% (199)	77.53% (276)	–
A1b	89.89% (13286)	89.42% (10508)	99.31% (20198)	81.02% (269)	70.04% (194)	76.40% (272)	73.08% (76)

Table 3.3. Similarity of the repeats that make up LCRs PMD-A and PMD-B (along with other related sequences) to each other as reported from BLASTz alignments. The pairwise similarity between sequences is expressed as the percentage of identical bases out of the total length of the shorter of the two sequences in the alignment. The number of identical nucleotides included in each alignment is given in brackets underneath the percentage similarity for each comparison. Where there was no significant similarity found between two sequences by the BLASTz algorithm, this is shown by a dash (–).

3.4.5. H2b-like genes

Part of the sequence making up the repeats in LCRs PMDA and PMDB has similarity at both the nucleotide and protein level to the H2b family of histone genes. Two of these H2b-like sequences have been annotated on the Ensembl genome browser (Ensembl gene IDs: ENSG00000101812; ENSG00000123569), in repeats A2 and A3 respectively, as predicted genes both with similarity to H2b genes at the N-terminal half of the translated sequence. The part of the distal repeat block that has similarity to a section of clone bA370B6 contains H2b-like sequence, and this has also been annotated as a gene on the Ensembl genome browser (Ensembl gene ID: ENSG00000172450) and has been named BA370B6.1 (Figures 3.4. and 3.5.). As with the other two annotated genes, BA370B6.1 also has similarity to H2b protein sequences at the N-terminus. Analysis of the nucleotide sequence of the inverted repeats A1a and A1b and comparison with the other repeats A2 and A3, as well as the consensus H2b, showed that a part of these repeats also had similarity to the H2b genes (Figure 3.5.). There are several copies of all of the five types of histone genes within the genome, which are distributed in clusters, with the largest on chromosome 6 containing 55 histone genes (Marzluff *et al.*, 2002). Histone mRNAs are unique in that they do not contain a poly(A) tail and instead all end with a conserved sequence that forms a stem-loop secondary structure (Marzluff *et al.*, 2002). The five H2b-like gene/nucleotide sequences were all compared against the published consensus H2b protein sequence using BLAST2 and ClustalW, and also against a published consensus stem-loop sequence, showing that there was a large degree of similarity between the sequences (Figure 3.5.) (Marzluff *et al.*, 2002).

3.4.5.1. H2b-like genes in other sequenced genomes

Some other mammalian genome sequences with synteny to this region of the human X chromosome were examined for the presence of H2b-like genes by comparing sequence from the human transcripts in this region against the syntenic sequence by BLAST2.

There was no similarity found between the human H2b-like genes and either the mouse or rat syntenic genomic sequence, and as these two species do not seem to have a similar large repeated region distal to *PLP1* this finding was not surprising (see section 3.6.). Comparison of H2b-like transcripts with the chimpanzee genome found one short sequence with similarity to the H2b-like genes, which appeared to be orthologous to the BA370B6.1 gene based on its position relative to other orthologous chimpanzee genes annotated in the region. The sequence had not been annotated as a gene in its own right in the chimpanzee sequence (NCBI Build 1 version 1). It is quite possible that the chimpanzee may harbour further orthologous H2b-like genes distal to *PLP1*, as the available sequence at the time of writing contained a large gap in the region that was syntenic to the major human distal repeats (see section 3.6.1. and Figure 3.9.).

3.4.6. Xrep enhancer

Part of the LCR-PMDB and PMDB repeats also consists of a 973bp sequence that has been described as “Xrep”, and has been found to stimulate plasmid growth in *E.coli* and *S. cerevisiae*, either by increasing the rate of replication or increasing the stability of Xrep-containing plasmids (Riley *et al.*, 1986). Xrep was found to contain sequences very similar to the replication origins of some human DNA viruses, such as the human BK virus, as well as viral enhancer-like sequences (Rosenthal *et al.*, 1983; Riley *et al.*, 1986). It has been suggested that, given the similarities between Xrep and viral replications of origin, that Xrep may function as an X chromosome replication origin (Riley *et al.*, 1986). However, Xrep sequences only showed very weak replicative activity in yeast, and the Xrep sequence was found not to contain a modular sequence element associated with several origin regions in eukaryotes (Riley *et al.*, 1986; Dobbs *et al.*, 1994). The position of the Xrep-related sequences within the four repeat units A1a, A1b, A2 and A3 is indicated by the red shading in Figure 3.4b.

3.4.7. G+C and interspersed repeat content of distal repeats

The 1Mb region distal to the transcription start site of *PLP1*, which contains all the distal repeats, has a G+C content of 40.43% and an interspersed repeat content of 55.27% (Repeatmasker, Repbase version 7.4). The G+C and interspersed repeat content of many of the distal repeats differ from these average values (Table 3.2.). The repeated regions within LCRs PMDA and PMDB had a slightly higher G+C content (the mean of A1a, A1b, A2 and A3 is 44.85% G+C), and a lower mean incidence of interspersed repetitive elements (40.70%) (Table 3.2.). The two other main LCRs PMDC and PMDD, however, had G+C contents closer to the regional mean, but a relatively high interspersed repeat content (mean of 59.89%) (Table 3.2.). The shorter stretches of similarity to the large distal repeats had very divergent G+C and interspersed repeat sequence content, which is because these are such short stretches of sequence, so are likely to have a skewed sequence content by chance and some of which are also associated with coding sequences (Figure 3.4. and Table 3.2.).

H2b consensus	E S Y S V Y V Y K V L K Q V H P D T G I S S K A M G I M N S
A3	TTCTGGG GACAGCT TTCA CC CCCTATTTCC CC CGG GTGCTGAAGCAGGTT CAC CCAGGGCC TCAGC CTT TCC CCAG GAGGCCG TCAGT GT CA TG GAT TCT
A1b	TTCTGGG GACAGCT TTG CC GCCTATTTCC CC CGG GTGCTGAAGCAGGTT CAC CCAGGGCC TCAGC CTT TCC CCAG GAGGCCG TCAGT GT CA TG GAT TCT
A1a	TTCTGGG GACAGCT TTG CC GCCTATTTCC CC CGG GTGCTGAAGCAGGTT CAC CCAGGGCC TCAGC CTT TCC CCAG GAGGCCG TCAGT GT CA TG GAT TCT
A2	CGCTGGG GACAGCT TCG CC ACCTATTTCC CC CGG GTGCTGAAGCAGGTT CAC CCAGGGCC TCAGC CTT TCC CCAG GAGGCCG TCAGT GT CA TG GAT TCT
370B6 6969	GG GACAGC TTCA CC ACCTATTTCC CC CGG GTGCTGAAGCAGGTT CAC CCAGGGCC TCAGC CTT TCC CCAG GAGGCCAG TGGAT GT CA TG AT TCC
H2b consensus	F V N D I F E R I A G E A S R L A H Y N K R S T I T S R E I Q T
A3	ATGATCCAT GAC ATATTG GAC CGCATCGCCACCGAGGCTGGT CAG CTGGCCCA TAC ACCAAGCGCGTGACCATCACCTCCCGG GAC ATCCAGATG
A1b	ATGGTT CGT GACATACTG GAC CGCATCGCCACCGAGGCTGGT CAC CTGGCCCA CTAC TCC AAG TGCGTGACCATCACCTCCCGG GAC ATT CG ATG
A1a	ATGGTT CGT GACATACTG GAC CGCATCGCCACCGAGGCTGGT CAC CTGGCCCA CTAC TCC AAG TGCGTGACCATCACCTCCCGG GAC ATT CG ATG
A2	TTGGTT CAT GACATACTG GAC CGCATCGCCACCGAGGCTGGT CGC CTGGCCCGCTCCACCAAGCGCGCAGACCATCACT GCC TGGGAGACC CG ATG
370B6	TT CGTT CAC GAC ATCTTG GAG CACATCGCCACCAAGGCCGGCCACT TG GCCCA CTAT ACCAAGTGC ACC ACCATCACCTCCTGCGAGATGCAGACC
H2b consensus	A V R L L L P G E L A K H A V S E G T K A V T K Y T S S K X
A3	GCCGTGCGACTGCTGCTGCCGGGG AAG ATGGCAAGCTCGCCGAG GCCCAG GGC CAC GAATGCCGCCCTCAGGTACACCAAAAGCAAG TG AGCTGTC
A1b	GCCGTG TGC CTGCTGCTGCCGGGG AAG ATGGCAAGCTCGCTGAGTCT CAG GGC CAC GAATGCCACCCTCAGGTACACCAAAAGCAAG TG AGCTATC
A1a	GCCGTG TGC CTGCTGCTGCCGGGG AAG ATGGCAAGCTCGCTGAGTCT CAG GGC CAC GAATGCCACCCTCAGGTACACCAAAAGCAAG TG AGCTATC
A2	GCTGTGCGCTGCTGCTGCCGGGG CAG ATGGCAAGCTCGCCGAGTCCGAAGGCACGAAGGCTGTCTCAGGTACACCAGGAGCAAGTGCCTCCC
370B6	ATTGCGCTGATGTTT---CCGGGG CAG ATGGCAAGCACGCCATATCCAGGGGCTCCAAGACTCTGCTCCACTACACCAGGAGCAAG TG AGCTGCC
3'end histone mRNAs	N ₁₇₋₇₂ ccAAAGGcuCUUuUcAGaGCCacc-cacnuuuucnnaaaagagcugu
A3	TCAGGAGCGCCTGAGCACCTGGGAAACCCAAAGGCTCTATT CAGA ACCACCGCCCGTGG-CCCTATAGGCCAGTGGCCCGCCAAGGGGGGA 5950
A1b	TCAGGAGCGCCCGAGCACCCGGGAAAGCCAAATGCTCTGTTC CAGA ACCACCACACGTGG-CCCTAAAGACCAGTGGCCTGCCAAGGAGGGGA 7327
A1a	TCAGGAGCGCCCGAGCACCCGGGAAAGCCAAATGCTCTGTTC CAGA ACCACCACACGTGG-CCCTAAAGACCAGTGGCCTGCCAAGGAGGGGA 7326
A2	TCAGGAGCACCCGAGCACACGGGAACCCAAAGGGTCTTTTCAGAGCCACTGAATGTGG-CCTGAAAAGACCAGTGGCTCGCCAAGG--GGGA 1058
370B6	TCAGGAGCGCCTGAGCACCA---AAACCCAAAGCTCTTTTCAGAGCCACCGCACTTGG-CTTGAAAAACCTGTAGCCC 6611

Figure 3.5. ClustalW alignment of the H2b-like regions. Legend on next page

Figure 3.5. Legend.

ClustalW alignment of the H2b-like regions contained within the various repeated regions distal to *PLP1*, with the consensus H2b protein sequence (from Marzluff *et al.*, 2002) also shown. Sequence that codes for an identical amino acid to the consensus is highlighted in red, sequence coding for a similar class of amino acid is shaded grey, and stop codons are shaded maroon. Positions from the start of the various repeats are shown, except for the sequence from clone bA370B6, where the numbers refer to the distance in base pairs from the start of the sequence of this clone. The region similar to the 3' end of histone mRNAs (from Marzluff *et al.*, 2002) in the repeats is highlighted in green, the consensus is shown above, with invariant bases shown as capital letters in the consensus.

3.5. Repeats proximal to *PLP1*

The region upstream of *PLP1* appeared to contain numerous short locally repeated sequences, which seemed to fall into two major alternating blocks, creating a checkerboard-like pattern when this region was compared against itself using BLASTz by the Pipmaker program and shown as a dotplot (Figures 3.1. and 3.6.). The nature of these repeats was investigated using a combination of BLASTn and BLASTz sequence comparison tools to delineate the borders of the repeats. Four different types of repeated DNA sequence were found in this *PLP1*-proximal region, which have been here designated as types P1-P4, and a summary of all the repeats is shown in Figure 3.8. Some of these repetitive sequences were also found to be associated with genes.

3.5.1. Proximal repeat group P1

The set of repeats arbitrarily designated as “P1” included 9 different sections of sequence in the 1.1Mb region examined proximal to *PLP1* (Figure 3.8., Table 3.4.). The average length of the P1 repeat sequences was 8640bp, and the mean G+C content of the repeats

(40.71%) was close to the G+C content of the 1.1Mb region proximal to the start of *PLP1*, which contained all of the proximal repeated sequences considered in this chapter (40.67%) (Table 3.4.). The interspersed repeat content of the various P1 repeats varied, with a mean of 41.94%, which was considerably lower than the average (55.67%) for the whole region (Tables 3.4. and 3.5.). Most of the alignments amongst the P1 repeats only included part of each repeat unit, and alignments were often fragmented due to the presence of other sequences, such as common interspersed repetitive elements, within the boundaries of the P1 repeats (Table 3.5.). The P1 repeats were the most widely dispersed of all the repeats found in the *PLP1*-proximal region, with the two most proximal and distal repeats in the group located almost 800Kb apart, and none were found to be associated with annotated genes (Figure 3.8.).

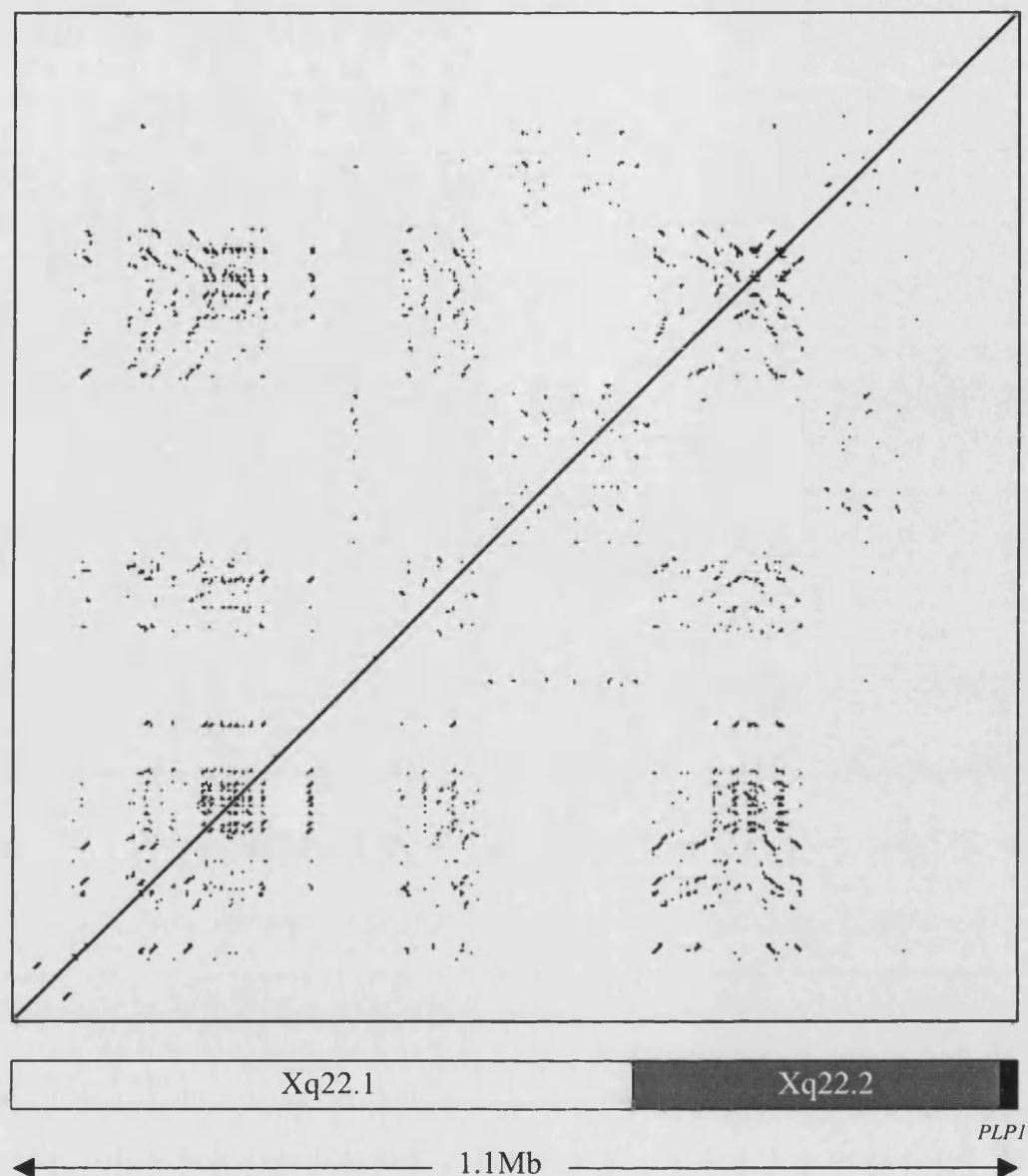


Figure 3.6. Dotplot produced by Pipmaker using BLASTz comparison of sequence from 1.1Mb proximal to *PLP1* against itself (finishing at the end of *PLP1*). Regions of similarity are shown as black areas on the dotplot. Repeats within this region are shown as diagonal lines on the dotplot, upwards-facing diagonals (/) indicate directly repeated sequences, and inverted repeated sequences are downwards-sloping diagonals (\). The location of the chromosomal bands in the region (taken from the Ensembl genome browser) are shown underneath the dotplot. The position of *PLP1* is indicated by the black shaded area in Xq22.2. Genomic sequence (taken from the UCSC genome browser) was masked for human interspersed repetitive elements before the BLASTz comparison was made.

Clone	Size (bp)	Start	End	G+C (%)	Repeat content (%)	Orientation on chromosome
dJ769N13	11350	139630	150979	41.80	38.17	+
bA522L3	7961	58300	66260	40.52	30.13	+
cU157D4	15210	6231	21440	41.60	43.89	–
dJ635G19	1925	130	2824	42.41	52.65	+
dJ635G19	3695	33007	36901	35.51	57.37	+
bB349O20	11556	35037	46592	42.13	35.63	+
dJ421I20	7605	15230	22834	40.88	39.91	+
dJ43H13	9137	9332	196	41.70	35.04	–
cU246D9	9322	9328	7	39.87	44.65	–

Table 3.4. Summary of proximal P1 repeats. The positions of the start and end of each repeat are given from the start of the sequence from the genomic clone in which it is contained. Interspersed repeat content is as found using Repeatmasker. (+) orientation is going from centromere to telomere, (-) is the opposite direction.

								dJ769N13 139630-150979
							bA522L3 58300-66260	60.92% (3368)
						cU157D4 6231-21440	74.48% (3792)	74.62% (4645)
					dJ635G19 130-2824	71.22% (693)	71.53% (515)	69.66% (675)
				dJ635G19 33007-36901	67.26% (452)	82.60% (1182)	70.08% (923)	68.61% (483)
			bB349O20 35037-46592	66.74% (1240)	74.88% (903)	92.22% (5100)	65.48% (3893)	68.48% (6950)
		dJ421I20 15230-22834	69.70% (3612)	70.61% (370)	73.27% (551)	70.38% (3712)	70.78% (2807)	59.60% (2400)
	dJ43H13 9332-196	77.04% (3420)	60.75% (4346)	69.59% (1301)	75.36% (923)	74.81% (6795)	76.13% (3965)	73.81% (4503)
cU246D9 9328-7	75.45% (1718)	70.78% (2807)	75.49% (3480)	73.53% (1711)	77.46% (976)	64.13% (5173)	75.05% (3478)	75.69% (2923)

Table 3.5. Similarity of the P1 repeats to each other as reported from BLASTz alignments. The pairwise similarity between sequences is expressed as the percentage of identical bases out of the total number of base pairs included in each alignment. The size ranges given for each repeat are for the maximum extent of the repeat based on comparisons to all of the other repeats in the group, the percentage given is only for the parts of two sequences that are aligned by BLASTz. Not all of an individual repeat unit is necessarily aligned in a particular pairing; the number of base pairs included in each alignment is given underneath the percentage similarity in brackets. Where at least half of either sequence is included in the alignment, the figures are in bold type.

3.5.2. P2 repeats

This group of repeated sequences, referred to as “P2” repeats in this study, proximal to *PLP1*, were found to contain coding sequences in a minority of the repeat units, by comparisons with the annotation of the genomic sequence (Ensembl genome browser). BLASTz and BLASTn searches were used to determine the extent of the similarity between the different sequences (Tables 3.6. and 3.7.). The P2 group of repeats contained the largest number of individual repeats out of the four classes of repeat types, with 15 whole or partial copies present (Figure 3.8., Table 3.6.). The P2 group of repeats also contained the most sequence in total out of the four repeat groups, 125.1Kb (Table 3.6.).

3.5.2.1. *RAB* genes

Two of the repeats contained genes belonging to the *RAB* gene family of small GTP-binding proteins. These two genes were both annotated as *RAB40A* on the Ensembl genome browser and were located in regions with very high sequence similarity, and both the DNA and protein sequence were 97% identical to each other. RAB proteins comprise the largest subgroup within the Ras superfamily of GTPases, and the human genome contains at least 60 *RAB* genes (Bock *et al.*, 2001). RAB proteins contain various conserved motifs, including a GTP-binding domain and a prenylation motif at the C-terminal end of the protein (Pereira-Leal and Seabra, 2000). There is also another gene from the RAB protein family located within Xq22.2, *RAB9B*, which is just proximal to *PLP1* (Figure 1.4.). However, *RAB9B* is not particularly closely related to the *RAB40A* genes within the wider *RAB* gene family (Pereira-Leal and Seabra, 2001). There is only 31% identity at the protein level between RAB9B and either of the two RAB40A copies. No significant similarity was found between the transcript sequences at the nucleotide level using BLASTn, and no similarity was found between the sequences within and around these three sequences using BLASTz comparisons (Figure 3.1.). The RAB40A proteins differ from many of the other RAB proteins (including RAB9B) in the

prenylation motif present in the protein. The RAB40A isoforms have a C-terminal CAAX box as a target for prenylation modification of the protein, whereas most RAB proteins, including RAB9B, contain a different prenylation motif consisting of one, or usually two, cysteine residues at the very end of the protein (Pereira-Leal and Seabra, 2000;Pereira-Leal *et al.*, 2001). RAB proteins are prenylated by the covalent addition of a geranylgeranyl isoprenoid group to a C-terminal cysteine residue, by the protein RAB geranylgeranyl transferase (RGGT), which recognises RAB proteins only when bound to the 95KDa RAB escort protein (REP) (reviewed in Pereira-Leal *et al.*, 2001). The RAB40A proteins (as well as the closely related proteins RAB40B and RAB40C, present in genomic locations 17q25.3 and 16p13.3 respectively) also differ from other RAB proteins, including RAB9B, in the presence of a SOCS box (suppressor of cytokine signalling) protein motif near the C-terminus of the protein, a motif which may have a function involving the regulation of protein turnover (Hilton *et al.*, 1998;Pereira-Leal and Seabra, 2000;Kile *et al.*, 2002). The *RAB40A* isoform located in genomic clone cU237H1 has been found to be expressed most highly in the brain, and is also expressed in heart, skeletal muscle, kidney and liver (Saito-Ohara *et al.*, 2002). Immunocytochemistry on this RAB40A protein showed intracellular localisation of the protein to the mitochondria (Saito-Ohara *et al.*, 2002).

3.5.2.2. Sequence characteristics of P2 repeats

Most of the P2 repeat units did not contain sequences coding for *RAB* genes, however, and most of the similarities found were between non-coding sequences. As found with the P1 repeats, many of the alignments between different members of the P2 family only included part of each sequence, and many of the alignments were fragmented by interspersed repetitive elements (Table 3.7.). Although the mean interspersed repeat content (as found by Repeatmasker) was below the average for the region, at 42.67%, the interspersed repeat content of the different repeats varied widely between P2 sequences

with some being much higher than the regional average of 56.67% (Tables 3.4. and 3.7.).

The G+C content of the P2 repeats was not as variable, and the average G+C content (39.42%) was close to the regional average of 40.67% (Tables 3.4. and 3.7.).

Clone	Size (bp)	Start	End	G+C (%)	Repeat content (%)	Orientation on chromosome	Gene name
dJ1054G24	5045	5268	224	37.62	20.61	–	
cU237H1	9243	29180	19938	37.91	54.40	+	
cU237H1	1842	13364	15205	37.68	6.95	–	
cU237H1	10283	10285	3	41.97	20.32	+	<i>RAB40A</i>
cU235H3	9693	9963	271	38.37	51.20	+	
cU61F10	4306	39008	34703	38.16	27.71	+	
cU73E8	12976	25387	12412	40.26	69.22	+	
cU101D3	2931	10528	7598	41.28	62.71	+	
dJ635G19	3819	14246	18064	39.28	58.34	+	
dJ635G19	8440	26734	18295	43.53	46.88	–	
dJ635G19	10455	46750	57204	38.27	80.87	+	
cU250H12	11433	37916	26484	38.62	53.04	+	
cU250H12	21841	3169	25009	39.70	33.41	–	<i>RAB40A</i>
dJ43H13	5004	9573	14576	39.91	20.00	+	
dJ43H13	7789	14573	22361	38.80	34.37	+	

Table 3.6. Summary of proximal P2 repeats. The positions of the start and end of each repeat are given from the start of the genomic clone in which it is contained. Interspersed repeat content is as found using Repeatmasker. (+) orientation is going from centromere to telomere, (-) is the opposite direction.

														dJ1054G24 5268-224
													cU237H1 29180-19938	73.03% (685)
												cU237H1 13364-15205	77.09% (773)	74.47% (455)
											cU237H1 10285-3	65.23% (1272)	75.57% (1262)	74.92% (1146)
										cU235H3 9963-271	66.77% (1879)	76.46% (1023)	75.18% (4478)	76.49% (1402)
									cU61F10 39008-34703	73.32% (1154)	64.34% (1721)	77.38% (335)	69.19% (1534)	71.50% (961)
								cU73E8 25387-12412	64.33% (2047)	68.66% (4743)	72.49% (2443)	73.20% (1319)	72.79% (4576)	—
							cU101D3 10528-7598	69.95% (752)	72.16% (1130)	76.29% (428)	60.85% (1413)	75.54% (451)	76.57% (402)	—
						dJ635G19 14246-18064	58.60% (1240)	—	—	—	—	—	—	—
					dJ635G19 26734-18295	70.57% (885)	75.55% (674)	75.36% (1847)	74.49% (1594)	71.98% (2300)	75.76% (1303)	77.08% (817)	71.35% (1843)	73.92% (788)
				dJ635G19 46750-57204	67.47% (701)	56.52% (2080)	—	63.67% (1616)	—	73.81% (834)	—	—	60.53% (1833)	70.55% (206)
			cU250H12 37916-26484	72.34% (761)	71.64% (2059)	54.55% (109)	72.29% (934)	69.45% (4315)	74.69% (1959)	70.29% (3477)	71.99% (2950)	65.31% (1365)	74.16% (2471)	74.36% (1453)
		cU250H12 3169-25009	77.45% (2473)	71.27% (655)	77.23% (831)	—	78.95% (465)	72.34% (2111)	75.87% (1736)	74.92% (2557)	87.09% (8133)	78.28% (1067)	56.53% (1725)	71.73% (1573)
	dJ43H13 9573-14576	80.84% (1974)	67.72% (3443)	68.24% (232)	69.38% (2017)	—	83.55% (122)	72.46% (2713)	63.89% (1575)	70.15% (2080)	69.08% (1834)	76.77% (945)	87.50% (1568)	73.83% (1470)
dJ43h13 14573-22361	80.08% (2130)	77.52% (4731)	74.83% (3167)	—	77.15% (1341)	—	61.76% (1221)	71.51% (2670)	69.50 % (1894)	71.28% (1777)	69.19% (2958)	72.58% (1347)	75.25% (1861)	73.34% (1205)

Table 3.7. Similarity of the P2 (*RAB*-containing) repeats to each other as reported from BLASTz alignments. For table legend, see next page.

Table 3.7. Legend.

The pairwise similarity between sequences is expressed as the percentage of identical bases out of the total number of base pairs included in each alignment. The size ranges given for each repeat are for the maximum extent of the repeat based on comparisons to all of the other repeats in the group, the percentage given is only for the parts of two sequences that are aligned by BLASTz. Not all of an individual repeat unit is necessarily aligned in a particular pairing; the number of base pairs included in each alignment is given underneath the percentage similarity in brackets. Where at least half of either sequence is included in the alignment, the figures are in bold type

3.5.3. P3 repeats

Another group of repeats proximal to *PLP1* was identified by BLASTn searches and BLASTz alignments. This group of sequences, designated as P3 repeats, were found to contain several known and predicted genes, in this instance belonging to the *BEX/NADE* (brain expressed X-linked/p75NTR-associated cell death executor) family of genes.

3.5.3.1. *BEX/NADE* genes

BEX1 is located in human genomic clone dJ198P4 and has found to be expressed at a high level in mouse brain, kidney and testis (Table 3.8. and Figure 3.8.) (Yang *et al.*, 2002). Other members of the *BEX/NADE* family in Xq22 include *BEX2*, *NGFRAP1/BEX3* (Nerve growth factor receptor (TNFRSF16) associated protein 1) and a less well-characterised gene annotated in Ensembl as *Q9NMD9* (Table 3.8. and Figure 3.8.). *BEX/NADE* proteins have been found to play a role in mediating apoptosis in response to nerve growth factor, by interacting with the p75 neurotrophin receptor (Mukai *et al.*, 2000; Mukai *et al.*, 2003). Several genes homologous to *BEX* genes (*Bex1/Rex-3*, *Bex2*, *Bex3*) are present in the syntenic region of the murine X chromosome (Figure 3.9.)

(Brown and Kay, 1999). All the BEX/NADE proteins located in Xq22 contain a CAAX box at the C-terminal end of the protein, which acts as a substrate for post-translational modification by adding either a geranyl-geranyl or farnesyl group to the cysteine residue, resulting in the protein being attached to cellular membranes (Brown and Kay, 1999; Yang *et al.*, 2002).

3.5.3.2. Sequence characteristics of P3 repeats

The DNA sequence similarity between the different P3 repeats is relatively short, compared to some of the other proximal repeats, at 1146bp on average, but it was found to usually include most of the sequence in the repeat units (Table 3.9.). In the cases where there is an annotated gene in the sequence, the similarity does not extend very far outside this gene sequence. The G+C content of the P3 repeats was generally higher than for the other types of proximal repeats, and the G+C content of each P3 copy was always higher than the proximal region average of 40.67% (Table 3.8.). The interspersed repeat content was strikingly low, for a couple of P3 repeats no interspersed repeat units were found by Repeatmasker, which is due to most of these P3 repeated sequences consisting of coding sequence (Table 3.8.).

Clone	Size (bp)	Start	End	G+C (%)	Repeat content (%)	Orientation on chromosome	Gene ID
dJ198P4	1974	25616	23643	55.52	0	-	<i>BEX1</i>
dJ635G19	1978	62956	64933	57.18	0	+	Q9NWD9
dJ79P11	3128	32901	29774	49.58	1.92	-	<i>BEX2</i>
cU105G4	3739	10057	13795	44.80	19.42	+	(NM_152278)
cU104G4	1901	29100	27200	42.61	12.36	-	
bB349O20	2004	14882	16885	55.84	6.59	+	<i>NGFRAP1</i>

Table 3.8. Summary of proximal P3 (*BEX*-like) repeats. The positions of the start and end of each repeat are given from the start of the genomic clone in which it is contained. Interspersed repeat content is as found using Repeatmasker (EBI). (+) orientation is going from centromere to telomere, (-) is the opposite direction.

					dJ198P4 25616-23643
				dJ635G19 62956-64933	70.77% (1087)
			dJ79P11 32901-29774	80.44% (1534)	75.75% (1293)
		cU105G4 10057-13795	62.87% (1717)	—	—
	cU105G4 29100-27200	—	60.59% (1127)	71.90% (755)	62.76% (787)
bB349O20 14882-16885	71.80% (802)	56.16% (1085)	71.66% (1052)	66.54% (1078)	77.18% (1437)

Table 3.9. Similarity of the P3 (*BEX*-like) repeats to each other as reported from BLASTz alignments. The pairwise similarity between sequences is expressed as the percentage of identical bases out of the longer of the two sequences in the alignment. The size ranges given for each repeat are for the maximum extent of the repeat based on comparisons to all of the other repeats in the group, the percentage given is only for the parts of two sequences that are aligned by BLASTz, not all of an individual repeat unit is necessarily aligned in a particular pairing. The number of identical nucleotides included in each alignment is given in brackets underneath the percentage similarity for each comparison. Where there was no significant similarity found between two sequences by the BLASTz algorithm, this is shown by a dash (—).

3.5.4. P4 repeats

The final group of repeated sequences proximal to *PLP1* were named “P4” for the purposes of this study. Members of this group of *PLP1*-region specific repeats were located closest to the *PLP1* gene out of all the proximal repeats and were also found to consist of a family of genes and related sequences, as had been found for the P3 repeats (see section 3.5.3.). Seven short blocks of similar nucleotide sequence were found, between 2-5.4Kb in length, all of which contained known or predicted genes. These genes included *TCEAL1* (Transcription elongation factor A-like 1), which codes for the 21kDa protein p21/SIIR (Pillutla *et al.*, 1999).

3.5.4.1. *TCEAL1*-like genes

p21/SIIR is a nuclear phosphoprotein, found to have 48% similarity to the transcription factor SII, which can modulate transcription (Yeh and Shatkin, 1994; Pillutla *et al.*, 1999). The other genes found within the P4 repeats did show some degree of amino acid sequence similarity to *TCEAL1*, as well as the DNA sequence similarity (Table 3.11.). The P4 repeats, which also contained a large proportion of coding sequence, had higher G+C content than the regional average (Table 3.10.).

3.5.4.2. Sequence characteristics of P4 repeats

Some interspersed repetitive elements were found (by Repeatmasker software) within the P4 repeats (Table 3.10.). A short fragment of an L1ME4a element was found in all seven copies of the P4 (*TCEAL1*-like) repeat, with homology between bases 5846-6117 of the L1ME4a repeat consensus and this repeat element was located within or near the 3' UTR of the genes in the repeat unit (Figure 3.7.). Four of the repeat units also contained partial sequence homologous to the L1MC/D LINE consensus, and some of the repeat units also contained other interspersed repeats (Figure 3.7.). The interspersed repeat content of the repeats (16.25%) was lower than the average for the proximal region (Table 3.10.).

Clone	Size (bp)	Start	End	G+C (%)	Repeat content (%)	Orientation on chromosome	Gene name
cU177E8	5280	23297	28576	45.25	15.08	-	NM_153333
cU177E8	2073	3937	6009	45.49	19.92	-	ENSESTG 00000020633
cU105G4	3464	10341	13804	45.06	20.96	+	NM_152278
cU105G4	3521	36822	40342	44.59	18.74	+	NM_016303
cV857G6	2186	27125	24940	46.29	11.71	+	NM_024863
cV857G6	2388	4724	2337	48.12	13.15	+	NM_032926
dJ1055C14	5334	10539	15872	46.51	14.17	+	<i>TCEAL1</i>

Table 3.10. Summary of proximal P4 repeats. The positions of the start and end of each repeat are given from the start of the genomic clone in which it is contained. Interspersed repeat content is as found using Repeatmasker. (+) orientation is going from centromere to telomere, (-) is the opposite direction.

						cU177E8 23297-28576
					cU177E8 3937-6009	56.16% (770)
				cU105G4 10341-13804	—	64.23% (2219)
		cU105G4 36822-40342	56.35% (1841)	48.09% (792)	53.15% (1867)	
	cV857G6 27125-24940	61.12% (679)	60.60% (706)	70.97% (1327)	62.21% (826)	
	cV857G6 4724-2337	72.27% (1407)	—	64.83% (864)	78.02% (1514)	64.76% (900)
dJ1055C14 10539-15872	—	62.39% (740)	57.52% (1679)	53.31% (1617)	—	61.41% (3149)

Table 3.11. Similarity of the P4 repeats to each other as reported from BLASTz

alignments. The pairwise similarity between sequences is expressed as the percentage of identical bases out of the longer of the two sequences in the alignment. The size ranges given for each repeat are for the maximum extent of the repeat based on comparisons to all of the other repeats in the group, the percentage given is only for the parts of two sequences that are aligned by BLASTz, not all of an individual repeat unit is necessarily aligned in a particular pairing. The number of identical nucleotides included in each alignment is given in brackets underneath the percentage similarity for each comparison. Where there was no significant similarity found between two sequences this is shown by a dash (-). Comparisons highlighted in bold type are those where at least half of the nucleotides in at least one of the sequences are included in the alignment.

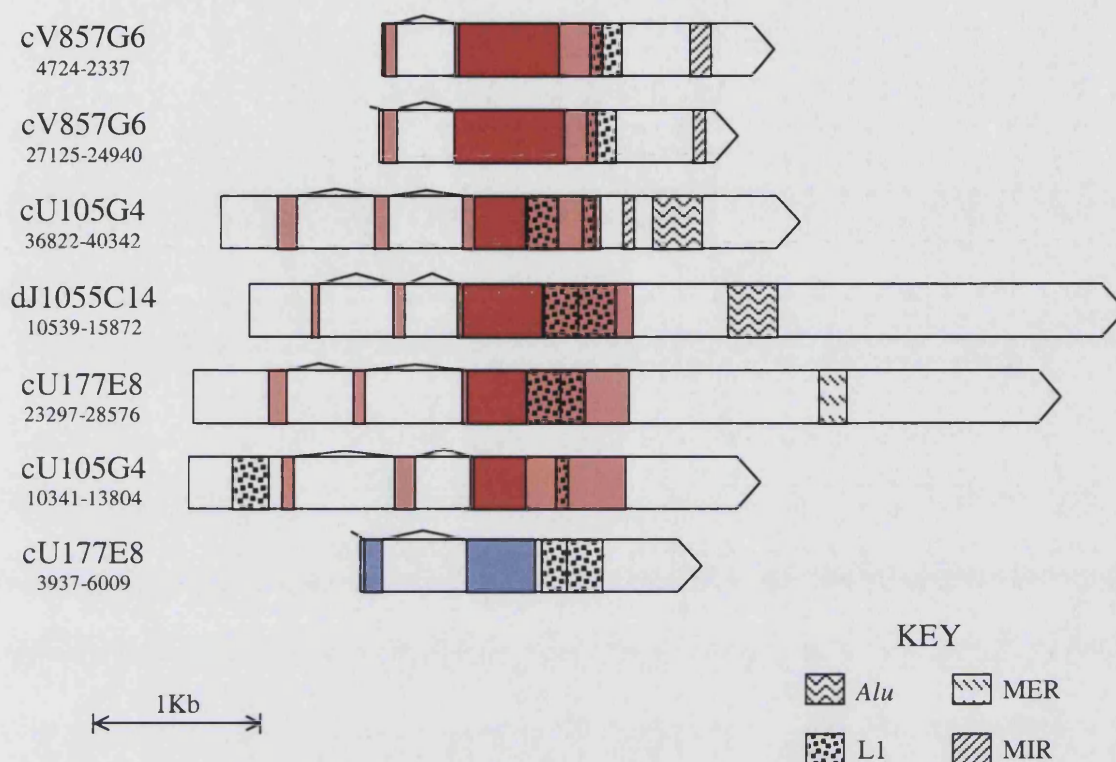


Figure 3.7. Repeat content and other features of the P4 (*TCEAL1*-like) repeats found in Xq22 proximal to *PLP1*. The relative sizes of the repeat units are shown by the large block arrows. Coding regions of genes are shown by red boxes, untranslated regions by pink colouring, and predicted coding regions (Ensembl Genscan prediction) are shown as blue boxes. Splicing patterns are shown above the repeats. Interspersed repeat content, as found by Repeatmasker software is shown by the patterned boxes.

3.5.5. Similarities between P3 (*BEX*-like) repeats and P4 (*TCEAL1*-like) repeats

A fraction of the sequence from genomic clone cU105G4 (3.5-3.75Kb, between bases 10057 and 13804) was picked up by BLASTz alignments as being similar to some of the repeated sequences in both the P3 and P4 repeats (Tables 3.8. and 3.10.). However, this sequence was among the least similar to the other sequences in the two sets of repeats, as it did not show similarity to some of the other sequences in the both groups, which was most apparent for the P3 (*BEX*-like) repeats, where only two of the other sequences showed any similarity (Tables 3.8. and 3.9.). Further investigation of the composition of this part of clone cU105G4 showed that much of the alignments found in the P3 repeats were due to common repetitive elements within this sequence, including a partial L1M4a element (Figure 3.7.). The gene found within this sequence (NM_152278) showed similarity at the amino acid level to *TCEAL1* and other genes in the P4 repeats, but no significant homology was found to the various *BEX*-like genes in the region (Figure 3.7.).

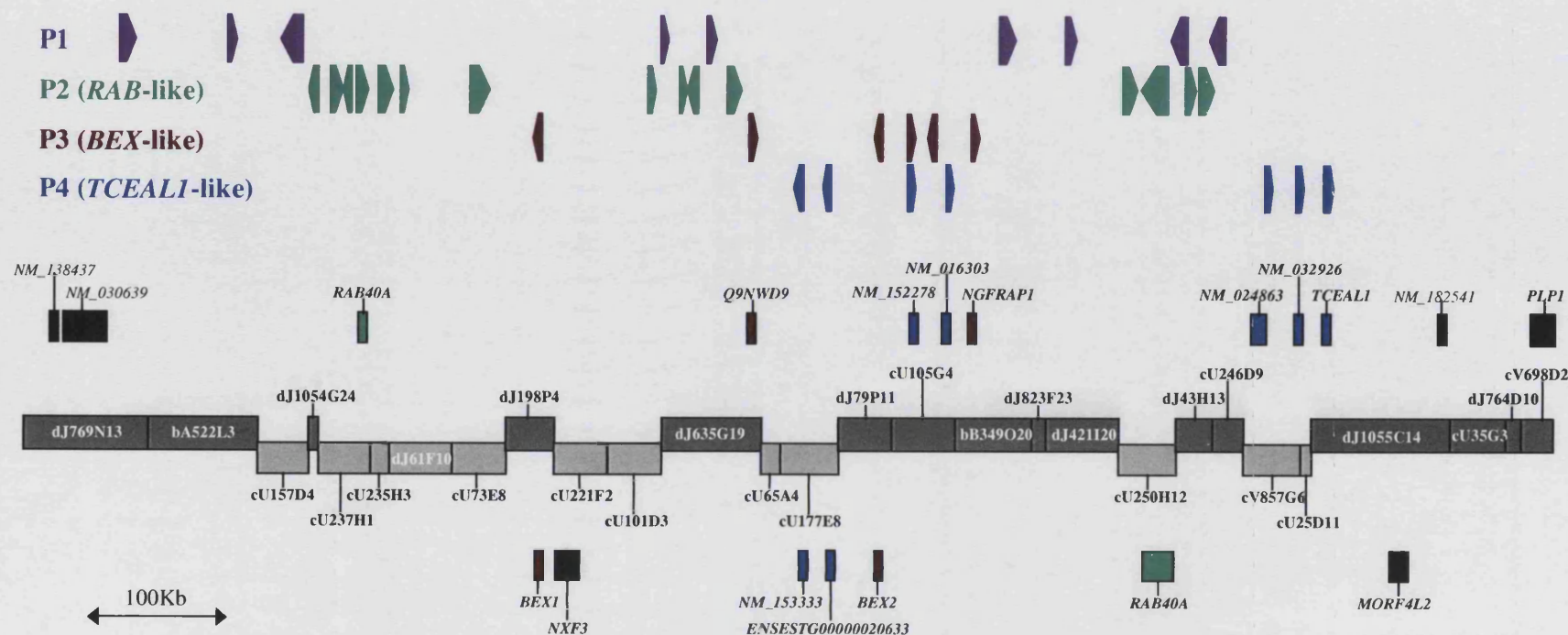


Figure 3.8. Summary of the local repeat structure, genes and clones proximal to *PLP1*. 1.1Mb spanning Xq22.1-Xq22.2 is shown, ending at the end of *PLP1*, with the left hand side of the diagram closest to the X centromere. Genomic clones are shown as grey boxes labelled with the clone name; those in darker grey have been sequenced in the direction centromere → telomere, those shaded in lighter grey have been sequenced in the alternative orientation. The four types of repeat found (P1-P4) are shown as coloured arrows above the contig, with the direction of the arrows showing the orientation of the repeat. Genes in the region are shown as boxes above the genomic clones if transcribed centromere → telomere, below if transcribed telomere → centromere. Genes are coloured according to which repeat unit they fall into, if not located within a repeat genes are shaded black.

3.6. Comparison of local repeat content near *PLP1* with other species

Several region-specific repeats have been identified in the human genomic sequence around *PLP1* as part of this study. It is possible that similar repetitive regions are present in the genomes of other organisms. Many of the repeats identified were found to share only a moderate degree of sequence similarity, indicating that if they had originated from a single ancestral sequence, this was likely to have occurred at a relatively distant point in the evolutionary past. 2Mb regions of the available genomic sequence from three species, the chimpanzee (*Pan troglodytes*), the mouse (*Mus musculus*) and the brown rat (*Rattus norvegicus*) were investigated for the presence of region-specific repeats as had been carried out for the human genomic sequence. For each organism, the 2Mb of genomic sequence was centred on the *PLP1* homologue on the X chromosome, as this region of the X chromosome shows synteny between all four species, according to data from the various genome projects. Each 2Mb of sequence was compared against itself using BLASTz and displayed by Pipmaker as a dotplot, after common interspersed repetitive elements had been masked by the UCSC genome browser (Figure 3.9.). Substantial regions of locally repeated sequence were observed in all three species, but these repeated regions did not necessarily resemble the repeat distribution observed for the syntenic human sequence (Figures 3.1. and 3.9.).

3.6.1. Regional repeats in the chimpanzee syntenic region

Not surprisingly, the chimpanzee genomic sequence had the most similar repeat distribution to the human sequence, which reflects the close evolutionary relationship between the two species. Proximal to the chimpanzee homologue of *PLP1*, a very similar pattern of repeats was apparent on the dotplot, with a similar chequerboard pattern to that seen in the human sequence proximal to *PLP1* (see section 3.5., Figures 3.1. and 3.9a.). Most of the very short similarities between sequences either side of *PLP1* were also still present in the chimpanzee sequence (Figures 3.1. and 3.9a.). Distal to the *PLP1*

homologue, however, no region of repeated sequences as was seen in the human sequence is apparent in the chimpanzee genomic sequence (Figures 3.1., 3.3. and 3.9a.). However, the chimpanzee sequence for this region at the time of writing is incomplete, with several gaps in the assembly (November 2003 assembly - NCBI build 1 version 1) (Figure 3.9a). The largest gap between the contigs in this 2Mb region roughly coincides with the region syntenic to the region containing the distal repeats (LCRs PMDA, B, C and D) in the human sequence (Figures 3.1., 3.4. and 3.9a.). The chimpanzee draft genome is constructed from a 4x shotgun sequence coverage which has then been assembled mostly by alignment to the human genome sequence (Reich *et al.*, 2002). This method of assembly causes difficulties when attempting to align chimpanzee sequence to highly similar duplicated regions in the human genome, as it may not be possible to determine which copy of the repeat the sequence should be aligned to (Reich *et al.*, 2002). A comparison of the human and chimpanzee sequences throughout this region by BLASTz, with a consistent upwards diagonal shown for most of the dotplot indicating directly repeated sequence, apart from where gaps are present in the chimpanzee sequence (Figure 3.10a). Part of the region flanking the large gap in the chimpanzee contig does appear to be similar to part of the distal repeat region mapped in humans, supporting the idea that the chimpanzee genome does contain some sequence with similarity to the human repeats distal to *PLP1* (Figure 3.10a). The proximal repeat regions look very similar between the two genomes, with the same chequerboard pattern also generated as when the human sequence was compared against itself (Figure 3.1.). The two repeat units A1a and A1b are highly similar (>99% sequence identity), and sequences with such high sequence identity are estimated to have originated by duplication within the last 2.5 million years (Table 3.3.) (Samonte and Eichler, 2002). This is much more recent than the estimated date of human-chimpanzee divergence, at approximately 7 million years ago, indicating that the A1a/A1b duplication is very likely to be human-specific (Samonte and Eichler, 2002). Such a difference between the organisation of the repetitive region distal to *PLP1*

between human and chimpanzee could lead to difficulties in the chimpanzee assembly and hence the gap in this region. Future versions of the chimpanzee genome assembly will hopefully give more insight into the evolutionary history of this genomic region.

3.6.2. Repetitive regions close to the murine *Plp1* gene

BLASTz comparisons using Pipmaker were also performed for 2Mb of the murine genomic sequence (October 2003 assembly - NCBI build 32) around *Plp1*, against itself and against the human sequence (Figures 3.9b and 3.10b). Proximal to *Plp1*, a highly repetitive ~300Kb region was picked up by the BLASTz analysis (Figure 3.9b). This repetitive region appeared to consist of two densely arrayed regions of short repeated sequences, flanking a longer pair of inverted repeat sequences (Figure 3.9b). This main repetitive region in the mouse did not bear much resemblance to any of the human repeats from the syntenic region, and there was little similarity between the human and mouse sequences in this particular region once the two sequences were aligned by BLASTz (Figure 3.10b). Much of the regions just proximal to, and all of the 1Mb distal to *PLP1* and *Plp1*, showed a reasonable degree of synteny between the two species, as represented by the roughly diagonal line on the dotplot in Figure 3.10b. The rest of the proximal region did not show such a direct relationship between the two sequences, but there was some similarity between some of the proximal repeat regions (P1-P4) and the mouse sequence, with some vestiges of the chequerboard-like pattern that was seen for the human and chimpanzee sequences being evident in the human-mouse comparison (Figures 3.10a, 3.10b, and 3.6.). There was very little similarity between the human distal repeats region and the corresponding part of the mouse genome, and the diagonal line of syntenic sequence similarity between the two genomes was interrupted at this approximate point, showing that this region had probably expanded by sequence duplication to create this repetitive region in the lineage leading to humans, but not during

the corresponding period of mouse evolution, or alternatively there could have been a deletion of this region in the mouse lineage (Figure 3.10b).

3.6.3. Repetitive regions near the brown rat *PLP1* homologue

Further highly repetitive regions, specific to the rat genome, were found close to the *PLP1* homologue in the available genomic sequence for this species. A BLASTz self-comparison of the 2Mb flanking the gene (June 2003 assembly, Atlas version 3.1.) showed no major region-specific repeats in the distal half of the region, but a large region, approximately 600Kb in length and proximal to the gene, was highly repetitive and consisted of both direct and inverted repeated sequences (Figure 3.9c). Fewer regions of similarity was seen between the rat and human sequences than for the other two interspecies comparisons, with virtually no similarity between the two sequences in the region of the rat proximal repeat (Figure 3.10c). Although sequence in the vicinity of *PLP1*, and also distal to the gene did show synteny in places, a lot of the sequence just distal to *PLP1* in humans was not present in this region in the rat genome (Figure 3.10c).

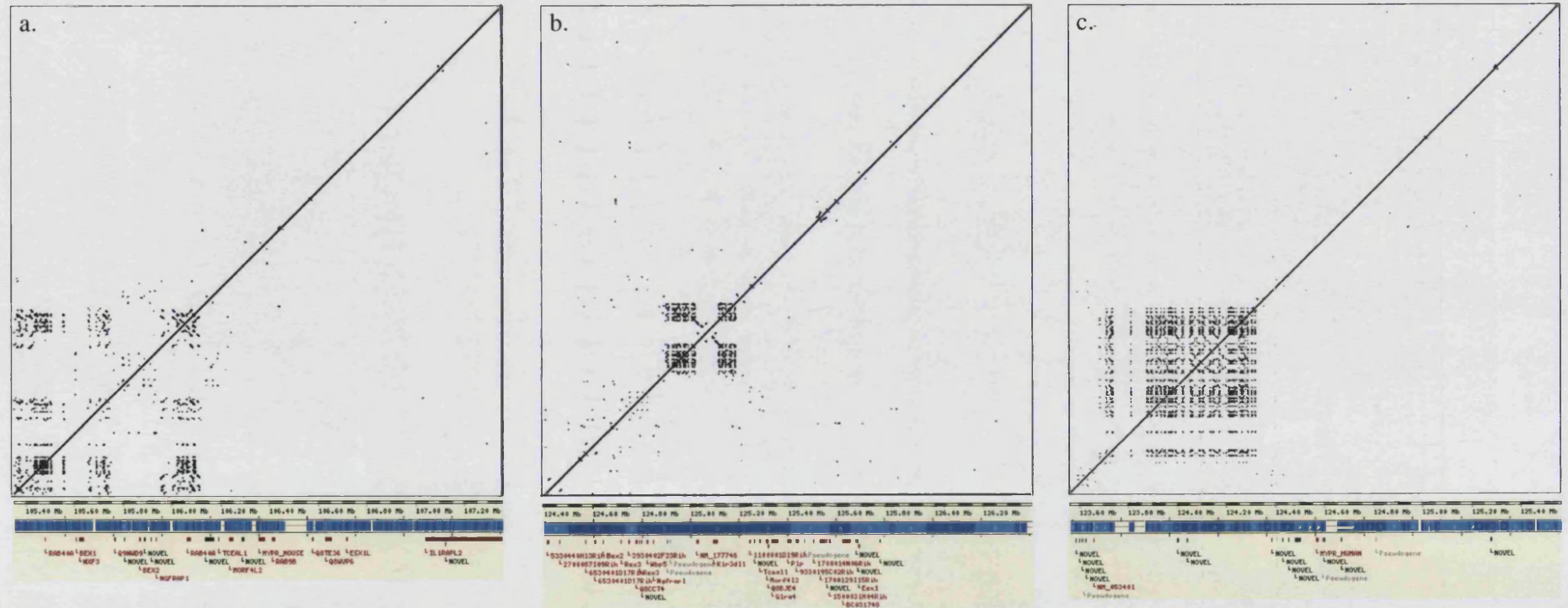


Figure 3.9. Locally repetitive sequences are present in syntenic regions of other genomes. Dotplots created by the Pipmaker program following BLASTz self-comparisons of 2Mb of genomic sequence taken from (a.) chimpanzee, (b.) mouse and (c.) rat, centred on the *PLP1* homologue in each organism. Sequence data, masked for common interspersed repetitive elements, was taken from the UCSC genome browser. Underneath the dotplots are shown the relevant regions of the X chromosome in each species, taken from the Ensembl genome browser. Repeats within this region are shown as dark areas on the dotplot, upwards-facing diagonals (/) indicate directly repeated sequences, and inverted repeated sequences are downwards-sloping diagonals (\).

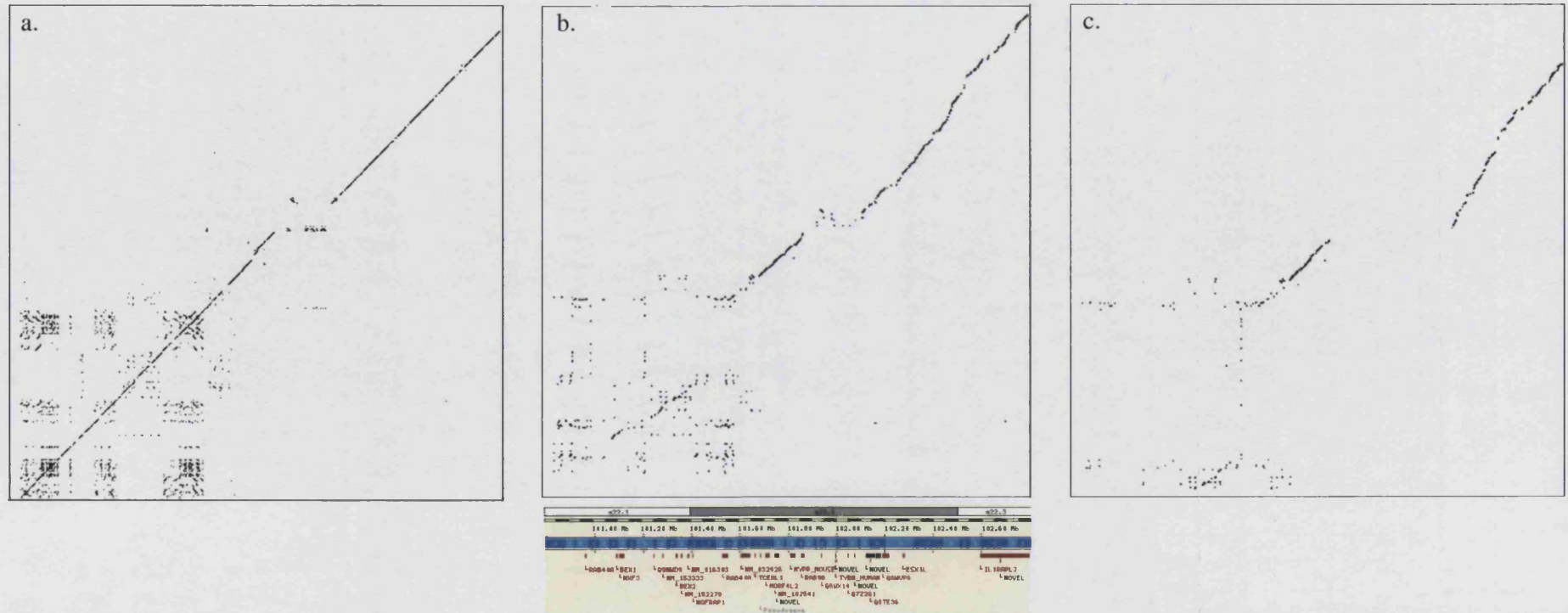


Figure 3.10. Comparisons of 2Mb of sequence around the *PLP1* homologue with human genomic sequence in (a.) chimpanzee, (b.) mouse and (c.) rat. The same regions of the genome of these three species as used for Figure 3.9. is compared against 2Mb of sequence from the human genome centred on *PLP1*, and was also masked for interspersed repeats before downloading from the UCSC genome browser using Pipmaker. In each dotplot the human sequence is aligned on the horizontal axis and the sequence from the other species is aligned on the vertical axis. The position on the human X chromosome and gene content is shown underneath one of the dotplots, and is the same for all three comparisons. Repeats within this region are shown as diagonal lines on the dotplot, upwards-facing diagonals (/) indicate directly repeated sequences, and inverted repeated sequences are downwards-sloping diagonals (\).

3.7. Discussion

3.7.1. LCRs/segmental duplications near *PLP1*

Most low-copy repeats or segmental duplications that are associated with disease-causing rearrangements are relatively large (>10Kb) and share a high degree of sequence identity (>90% for segmental duplications, while LCRs associated with recurrent disease-causing rearrangements usually have a higher degree of identity, >95%) (Bailey *et al.*, 2002; Stankiewicz and Lupski, 2002; Shaw and Lupski, 2004). The only repeats found near *PLP1* that meet these more stringent criteria for LCRs are the two inverted repeat sequences distal to the gene, A1a and A1b, with over 99% similarity to each other for over 20Kb, which had been previously identified by Inoue *et al.* (2002) (Tables 3.2. and 3.3.). The other two long repeats with some homology to A1a and A1b, A2 and A3, nearly meet the criteria for segmental duplications, having similarity to A1a and A1b for more than 10Kb, with the percentage of identical nucleotides ranging between 89-90% (Tables 3.2. and 3.3.). The longest pair of segmental duplications, LCR-PMDC and LCR-PMDD, also distal to *PLP1*, are also nearly similar enough to be classed as bona fide LCRs (92.43% similarity over 25Kb). The various repeated sequences proximal to *PLP1* (P1, P2, P3, P4) are not nearly similar enough or individually large enough to be classified as LCRs. There do not appear to be any LCRs either side of *PLP1*, as is found with many genomic disorders (Lupski, 1998). This is not unexpected, as duplications and deletions involving *PLP1* have been found to vary in size and do not appear to have recurrent breakpoints, making it unlikely that non-allelic homologous recombination is the mechanism behind these rearrangements (Inoue *et al.*, 2002; Hobson *et al.*, 2003; Iwaki *et al.*, 2003).

3.7.2. Interspersed repeats

The 2Mb of genomic sequence surrounding *PLP1* is relatively enriched for interspersed repetitive elements (55.47%), as found by Repeatmasker, relative to the X-chromosome

wide average repeat density of 46% (Venter *et al.*, 2001). The total combined length of all of the region-specific repeats examined in this chapter comes to 379942bp, or 19% of the total sequence, which only amounts to one third of the sequence covered by interspersed repeats. Some of the individual region-specific repeat families occupy a proportion of the sequence which is comparable to that taken up by some families of interspersed repetitive elements: P2 repeats take up 6.26% of the sequence and P1 repeats take up 3.89%, *Alus* make up 8.41% of the sequence and MaLR-type LTR elements account for 4.04% of the 2Mb of sequence (Repeatmasker).

Some reports have suggested that regions with high G+C and *Alu* repeat content are more likely to be involved in segmental duplications (Jurka *et al.*, 2004). Many of the larger distal repeats did have a higher G+C content than the regional average, but all of the distal repeats were found to have lower *Alu* content than was expected for this region (Table 3.2.). It is possible that some aspects of the sequence composition of the distal repeats, reflected in the G+C content, such as the presence of GC rich, recombination-promoting sequence motifs, may have been important for generating duplication events in this region during evolution (Abeyasinghe *et al.*, 2003). There were no *Alu* elements mapped close to the ends of the major distal repeats, so it seems unlikely that *Alu*-mediated recombination was responsible for creating these particular LCRs, although this has been shown to be a significant mechanism in the formation of segmental duplications (Figure 3.4.) (Bailey *et al.*, 2003). Many of the numerous proximal repeated sequences showed a higher than average G+C content, but in many cases this was likely to be due to the high proportion of genes in some of these sequences. A high proportion of coding sequences within the proximal region-specific repeats may also explain their relatively low interspersed repeat content.

3.7.3. Gene families in the *PLP1* region

Many of the genes near to *PLP1* show a degree of homology to other genes in the region and do form parts of the various *PLP1*-region specific repeats (*H2B*-like, *BEX*-like, *RAB40A* and *TCEAL1*-like genes). One addition to this number is the gene *NXF3*, which is a member of a family of nuclear RNA export factors (Jun *et al.*, 2001) (Figure 3.8.). Three other NXF genes (*NXF5* and two copies of *NXF2*) are located within 1.26Mb of *NXF3*, which is the most distally located of this repeated gene family in Xq22.1. The Ensembl genome browser (Release 20.34c.1) has 32 genes annotated within the 2Mb around *PLP1*, including the EST that is part of the *TCEAL1*-related sequences (Table 3.11). 19 out of these 32 genes are part of a locally repeated set of genes (7 *TCEAL1*-like, 5 *H2B*-like, 4 *BEX*-like, 2 copies of *RAB40A*, and *NXF3*), or nearly 60% of the annotated coding sequences in the region. If just the genes proximal to *PLP1* are considered, the numbers become even more striking, with 14/19 genes belonging to a repeat family, leaving *PLP1* as one of a minority of genes with no related sequences in the region (Figure 3.8.). Given that increases in *PLP1* gene dosage are pathogenic in humans and mice, it is unsurprising that there are no genes closely related to *PLP1* located in the region, or even elsewhere in the genome. *PLP1*/DM20 does show homology with M6a and M6b membrane proteins, but it has been shown that the three members of this gene family predate mammalian evolution, so there have not been any more recent duplications within this gene family (see section 1.3.1.) (Kitagawa *et al.*, 1993; Yan *et al.*, 1993). It is interesting to note that most of the genes located closest to the *PLP1* locus, such as *MORF4L2*, *RAB9B* and *TYBN*, also do not have closely related homologues in the region (Figures 1.4 and 3.8.). It is possible that duplications of these genes may not have occurred as for other sequences in the region partly because any duplications involving these genes would have been more likely to include *PLP1*, and would probably be selected against due to the resulting deleterious phenotype.

3.7.4. Repetitive regions in human and other species

Duplicated regions near *PLP1* are found in different species, but the actual sequences involved and the extent of the duplicated sequence differs between species (Figures 3.9. and 3.10.). Comparisons of the genomic sequence around *PLP1* between human and other species has shown that stretches of conserved syntenic regions are broken up by expansions of repeated sequences, especially between the rodent and human lineages (see section 3.6. and Figure 3.10.). Some repetitive sequences in the region are apparently more evolutionarily ancient than others. There is a degree of apparent conservation between parts of some of the human *PLP1*-proximal repeats (P1-4) and some very repetitive sequences proximal to *Plp1* in the mouse and to a lesser extent in the rat (Figure 3.10.). These particular repetitive regions most likely predated the primate-rodent split ~75 million years ago, but the similarities between these repeats has declined more within the rodent lineage than the primate lineage, as a much greater similarity between these proximal repeats is seen in the human and chimp sequence (Figures 3.1., 3.9. and 3.10.). It is possible that the greater similarity between these repeats is a result of concerted evolution, or gene conversion between repeats in the primate lineage (see section 6.4.5.1.). Another explanation is that there has been some selective constraint on the primate *PLP1*-proximal repeats, as many (P2, P3 and P4) are associated with coding sequences (Figure 3.8.).

The human proximal repeats are probably evolutionarily older than the distal repeat regions as the sequence identity between the different copies of each proximal repeat is not as great as the sequence similarity between most of the distal LCRs (Tables 3.3., 3.5., 3.7., 3.9. and 3.11.). In addition the distal repeats are not present in the syntenic regions of the rat or mouse (Figure 3.9.). The expansion of the rodent-specific repeat region proximal to *Plp1* has been noticeably greater in the rat genome as compared to the mouse, which may be due to a generally higher rate of tandem segmental duplications in the rat

genome as compared to the mouse, as has been reported (Figure 3.9.) (Tuzun *et al.*, 2004). Overall, these limited comparisons between the human, chimpanzee, mouse and rat genomes near the *PLP1* gene have indicated that the major difference between these regions is in the composition and distribution of local segmental duplications between the species, particularly the primates and rodents, rather than major changes in the syntenic sequence.

The rate of duplication of individual genes or other sequences in the human genome has been estimated to be between 2.2×10^{-9} – 1×10^{-8} duplicates per gene per year, which is comparable to estimated rates of point mutation per nucleotide per year (2.2×10^{-9} – 2.5×10^{-8}) (Lynch and Conery, 2000; Nachman and Crowell, 2000; Kumar and Subramanian, 2002; Bensasson *et al.*, 2003). The numerous segmental duplications in this region reflects the high rate of sequence duplication during evolution.

3.8. Summary

Analysis of the genomic regions flanking *PLP* has shown that this region contains several repetitive regions, with varying degrees of similarity. Most of the genes close to *PLP1* are duplicated and map within these repetitive regions. It is clear that sequence duplication within this genomic region, as well as elsewhere in the genome, has been an important force in the past. Duplications of this region are still occurring now, which is demonstrated by the PMD-causing duplications including *PLP1* that have been detected in this area.

4.1 DUPLICATION BREAKPOINT MAPPING IN FAMILY 1

Patient 1:9 and his unaffected carrier mother 1:4 were one of the first families to be described with a diagnosis of PMD associated with a duplication of *PLP1*, as determined by quantitative analysis of Southern blots (Figure 2.2.) (Individuals CO and KO in Ellis and Malcolm, 1994). Subsequent published work on this family has included both quantitative PCR and interphase FISH confirmation of *PLP1* duplication in both the boy and mother (Family NO in Woodward *et al.*, 1998). In addition, mapping of the extent of the duplication by interphase FISH has been carried out (see section 4.1.1.) and the duplication has been shown to be tandem in orientation (see section 4.3.) (Woodward *et al.*, 1998; Family PMD9 in Woodward *et al.*, 2000).

4.1.1. Duplication breakpoint mapping – previously published work

Previous interphase FISH data had shown that in this family the proximal end of the duplication was between PAC dJ198P4, which had been found not to be duplicated, and cosmid cU65A4, which was shown to be duplicated (Woodward *et al.*, 1998) (Figure 4.1.). These two genomic clones are spaced approximately 180Kb apart, and cosmid cU65A4 is approximately 540Kb proximal to *PLP1* (Figure 4.1.). The mapping of the distal end of the duplication using interphase FISH had found that cosmid cU240C2, located ~250Kb distal to the transcription start site of *PLP1* was duplicated (Woodward *et al.*, 2000). Another cosmid, cU85B11, which is located approximately 100Kb distal to cU240C2, although not mapped to the current human genome assembly (NCBI Build 34), was scored as not being duplicated (Figure 4.1.) (Woodward *et al.*, 2000).

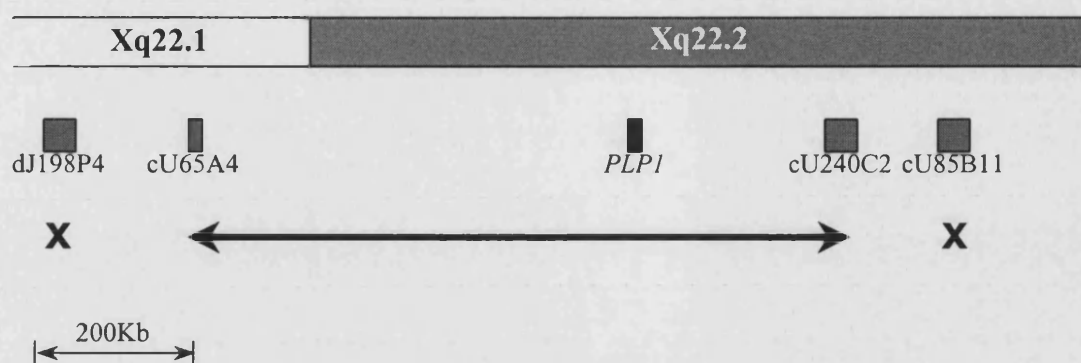


Figure 4.1. Extent of the duplication containing *PLP1* in family 1, as has been previously published. Adapted from (Woodward *et al.*, 1998; Woodward *et al.*, 2000). Location and boundaries of chromosomal bands Xq22.1 and Xq22.2, and the relative positions of *PLP1* (black box) and genomic clones dJ198P4, cU65A4 and cU240C2 (grey boxes) are taken from the Ensembl genome browser, version 21.34d.1. Cosmid cU85B11 is not mapped to the current genome assembly (NCBI Build 34 at time of writing), but is estimated to lie approximately 100Kb distal to cU240C2 (Woodward *et al.*, 2000). The minimum extent of the duplication is shown by the arrowed line and the boundaries are marked by X.

4.2. Fine mapping of duplication breakpoints in family 1 by interphase FISH and UPQFM-PCR

Further mapping of the locations of the duplication breakpoints in family 1 was carried out with the aim of sequencing the breakpoints, to hopefully gain a greater understanding of the mechanism of the rearrangement. Initially a combination of interphase FISH and UPQFM-PCR techniques at the proximal breakpoint and just UPQFM-PCR at the distal end of the duplication were used for finer mapping of the breakpoint (Woodward *et al.*, 1998; Heath *et al.*, 2000).

4.2.1. Proximal breakpoint mapping by interphase FISH

Interphase FISH using human genomic clones close to cosmid cU65A4, which had already been found to be duplicated, revealed that the proximal duplication breakpoint was most probably contained within the adjacent clone, PAC dJ635G19 (Figure 4.2., Table 4.1.). Interphase FISH using this clone showed that just over half of the nuclei were scored as duplicated (Table 4.1.). In many of these duplicated nuclei one signal was smaller than the other, suggesting that only a fraction of the clone was duplicated (Figure 4.2.).

4.2.2. UPQFM-PCR mapping of proximal breakpoint in family 1

Further mapping of the proximal end of the duplication was carried out by UPQFM-PCR (Table 4.2., Figure 4.3.). 2-4 pairs of UPQFM-PCR primers mapping to the region of interest were used alongside two pairs of control UPQFM-PCR primers in each experiment, one from exon 6 of *PLP1* (PLP1) and one from the *CFTR* locus on chromosome 7 (CF). Dosage ratios for each primer pair were calculated by comparing against sex-matched, normal control individuals (see section 2.2.1.3.1.).

UPQFM-PCR results for primer pairs located throughout the clones near the breakpoint further refined the breakpoint-containing region to 12Kb of genomic sequence (Figure 4.3., Table 4.2.). This 12Kb region was contained mostly within cU65A4, but also just included the end of the sequence from dJ635G19 (Figure 4.3., Table 4.2.). One primer pair within cU65A4 (6833-7109) gave mean dosage ratios that were consistent with that sequence being single-copy, despite being flanked by two pairs of primers that were apparently duplicated (Table 4.2.). These data were only based on two experiments in one individual, however, and the raw data from these two experiments was quite discordant. One result was compatible with the target sequence being duplicated (PLP1 ratio 0.9, CF ratio 1.56) and the other was

suggestive of the target sequence only being present once (PLP1 ratio 0.47, CF ratio 1.07). Given that all the other nearby primers appeared duplicated, and the two experiments for this primer gave such different results, it is most likely that this represents a false negative result for this primer pair.

4.2.3. UPQFM-PCR mapping of distal breakpoint.

Further mapping of the location of the distal breakpoint in family 1 was carried out using UPQFM-PCR. Sequence contained within the genomic clone cU240C2 has previously been found to be duplicated by interphase FISH in this family, but a UPQFM-PCR primer pair located in the proximal half of the sequence of this clone was found not to be duplicated in both the boy and carrier mother from family 1 (Figures 4.1. and 4.3., Table 4.2.). cU240C2 is located in a region containing several highly similar LCRs, and in fact contains almost the whole of LCR-PMDB (Figure 3.4.). Due to the highly repetitive nature of this region, it was only possible to design a few primer pairs in unique regions (Figure 3.4.). Preliminary data from these UPQFM-PCR primer pairs indicated that the location of the distal breakpoint in this family might be between the primer pair located in cV362H12 and dJ839M11 (Figure 4.3., Table 4.2.). However, few experiments had been carried out using these distal UPQFM-PCR primers, and many of the results were not particularly consistent. These UPQFM-PCR data were treated with caution, and additional techniques were used to verify the locations of the duplication endpoints (see sections 4.3. and 4.4.)

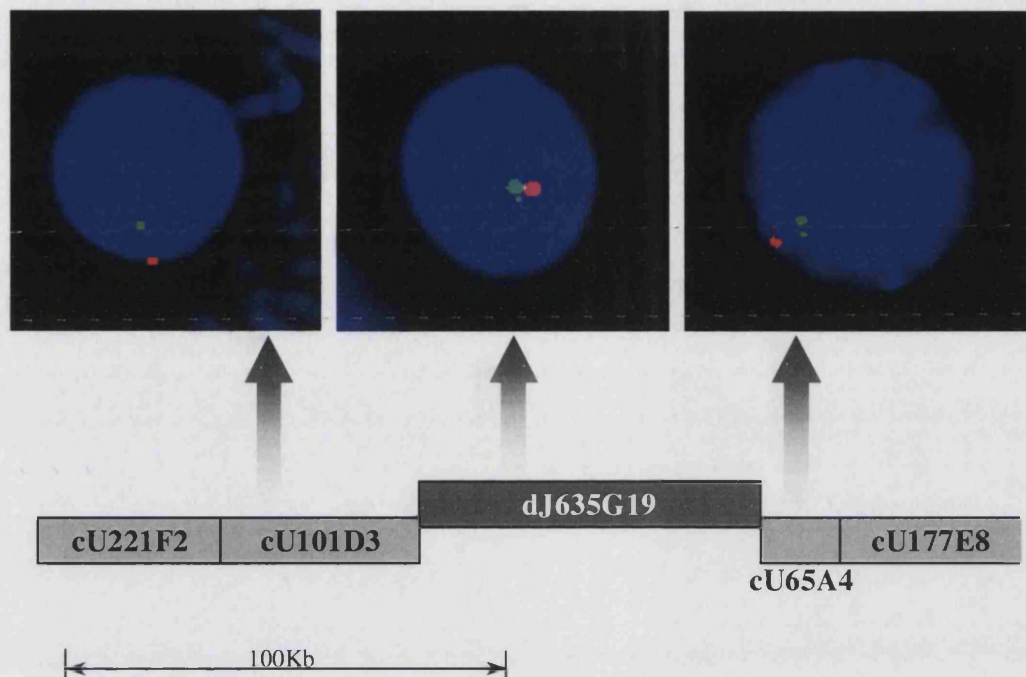


Figure 4.2. Interphase FISH results from the proximal duplication breakpoint using a cell line from the affected boy in family 1. Genomic clones around the proximal and breakpoint region are shown underneath the images of interphase nuclei, with scale bar underneath. In each image, the X centromere is labelled red, and the genomic clone used in the experiment is labelled green.

Clone	0,1	0,2	0,3	0,4	Other	Conclusions
cU101D3	65.75%	10.96%	0%	0%	23.29%	Not duplicated
dJ635G19	38.83%	43.27%	7.23%	2.90%	7.77%	Breakpoint?
cU65A4	23.53%	56.86%	4.90%	0%	14.71%	Duplicated

Table 4.1. Interphase FISH scores for patient 1:9 near the proximal duplication breakpoint. The mean percentage of nuclei falling into each category is shown for each clone (results taken from one slide for cU101D3 and cU65A4 and three slides for dJ635G19). An average of 75 interphase nuclei were scored for each slide. Some of these FISH experiments were carried out by Dr Karen Woodward or Sarah Knight.

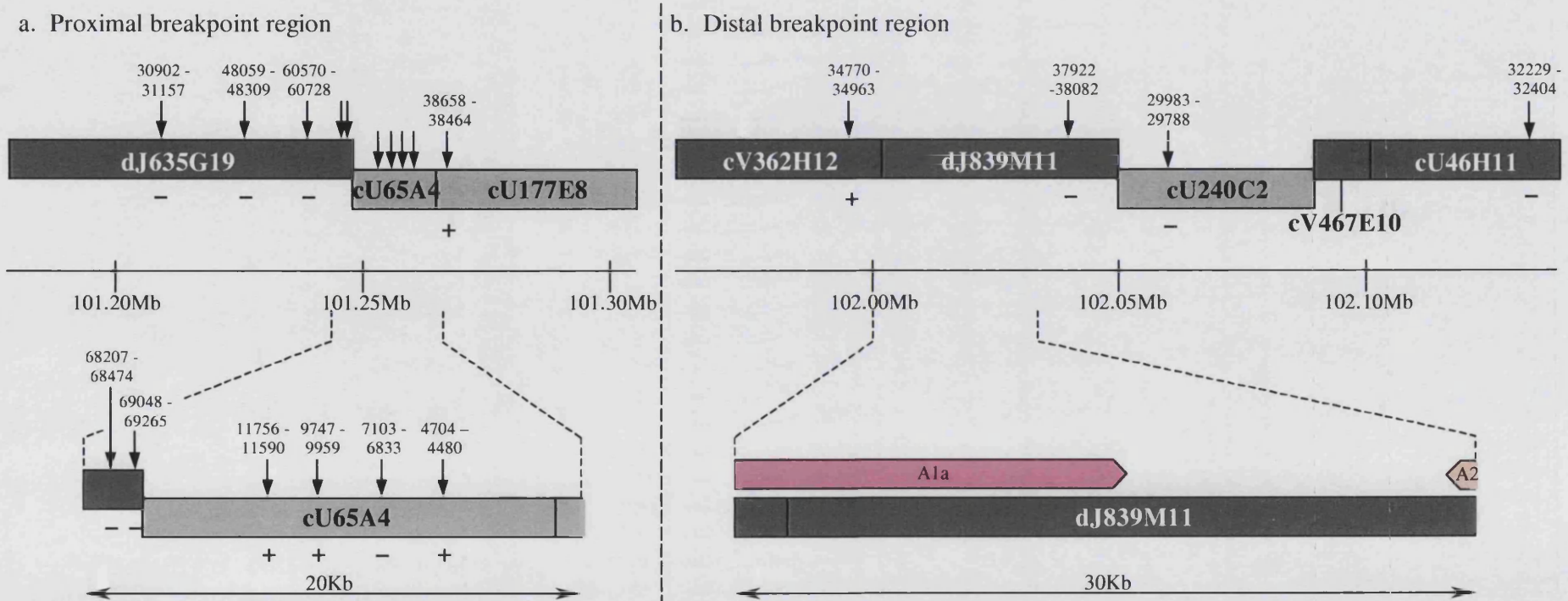


Figure 4.3. Diagrams showing the proximal (a.) and distal (b.) breakpoint regions in family 1. Clones around the proximal and distal breakpoint regions are shown in the upper section of the figure, with scale bar underneath (position on the X chromosome taken from the Ensembl human genome browser, release 21.34d.1). The UPQFM-PCR primers used are shown as vertical black arrows, with the area within the clone amplified by these primers shown above the arrow. (+) underneath a primer pair indicates that this primer pair was duplicated in family 1, (-) is shown underneath those UPQFM-PCR primer pairs that were single-copy. The lower sections of the figure shows a smaller region around the breakpoint, with UPQFM-PCR primers at the proximal end shown as before. The relative position of LCRs (A1a, A2; see Figure 3.4.) around the distal breakpoint, are also shown.

Clone	Position in clone of UPQFM primer pairs	Mean ratio of UPQFM PCR product compared to...		Number of experiments performed
		PLP1	CF	
dJ635G19	30902-31157	0.47 (0.29)	0.96 (0.12)	3 (1)
dJ635G19	48059-48309	0.55 (0.73)	1.42 (1.1)	2 (1)
<i>dJ635G19</i>	<i>60570-60728</i>	<i>0.64 (0.69)</i>	<i>1.31 (1.12)</i>	7 (2)
dJ635G19	68207-68454	0.52	1.08	2
dJ635G19	69048-69265	0.55	1.27	3
cU65A4	11590-11756	1.01	2.87	1
cU65A4	9959-9747	0.89	2.05	3
cU65A4	6833-7109	0.69	1.32	2
cU65A4	4480-4704	1.05 (1.33)	2.04 (2.00)	2 (1)
cU177E8	38464-38658	1.20	2.26	3
<i>cV362H12</i>	<i>34770-34963</i>	<i>1.18 (0.92)</i>	<i>1.13 (1.12)</i>	2 (1)
<i>dJ839M11</i>	<i>37922-38082</i>	– (0.48)	<i>0.59 (0.58)</i>	1 (1)
<i>cU240C2</i>	<i>29788-29983</i>	<i>0.36 (0.47)</i>	<i>0.66 (0.67)</i>	5 (3)
<i>cU46H11</i>	<i>32229-32404</i>	<i>0.56 (1.25)</i>	<i>0.77 (1.55)</i>	4 (1)

Table 4.2. UPQFM ratios for the proximal and distal duplication breakpoint in family 1 (as compared against both *PLP1* exon 6 and CF control primers). The last four pairs of primers in the table are distal to *PLP1*. Average ratios were taken from all experiments carried out that included that primer pair, and the number of experiments is indicated. Values are for dosage measurements from the affected male, where experiments had been additionally carried out using the carrier mother, these results are shown in brackets. Ratios consistent with duplication are highlighted in bold, and the zigzag lines show the assumed location of the breakpoint regions. The position of each target sequence within the relevant genomic clone is shown. Some of the UPQFM-PCR results in this table are from experiments carried out by Dr Karen Woodward, these results are shown in italic type.

4.3. Fine mapping of duplication breakpoints by fibre FISH

Fibre FISH was then carried out on this family to fine map the duplication breakpoints. Previously published work had shown that the duplication was tandem in orientation in the mother by dual colour interphase FISH using two cosmid clones within the duplicated region (Woodward *et al.*, 1998). If the duplication in family 1 is indeed a tandem rearrangement, then duplicated sequence from the distal end should be joined to duplicated sequence at the proximal end of the duplicated region, forming a junction that should not be present in normal genomic DNA. Fibre-FISH using genomic clones spanning the breakpoints should therefore be able to demonstrate the presence of any breakpoint junctions, and provide further positional mapping information for the breakpoint to back up the interphase FISH and UPQFM-PCR data.

4.3.1. Fibre-FISH to determine normal relationships between clones

Initially it was necessary to verify the positions and relationships between genomic clones present at the proximal and distal ends of the duplication. Human genomic clones from either the proximal or distal ends of the duplication were hybridised together on fibres from both the affected male and normal individuals. The clones used in the fibre-FISH experiments were those thought to contain or be very close to the breakpoint location, based on interphase FISH and UPQFM-PCR data (Figures 4.2. and 4.3., Tables 4.1. and 4.2.). At the proximal end of the duplication cosmid cU65A4 and PAC dJ635G19 were used, as the breakpoint had been mapped to a 12Kb region containing the junction of the sequences from these two clones (Figure 4.1.). At the distal end, cosmid cV362H12 and PAC dJ839M11 were used in the fibre FISH hybridisations (Figure 4.3.). The 45Kb breakpoint region provisionally mapped by UPQFM-PCR included sequence from both these genomic clones (Figure 4.3.). Initial hybridisations were carried out using the two proximal or two distal clones together on fibres from a normal cell line, to determine what the normal relationship

was between these clones (Figure 4.4.). The method used to produce the DNA fibres for the hybridisation does not stretch the DNA uniformly, so inferences about signal size can only be made on the basis of comparison to other signals on the same fibre, as different fibres may be stretched to different degrees.

4.3.1.1. Fibre-FISH for the proximal breakpoint region

For the two proximal clones, dJ635G19 and cU65A4, the signals from the two clones appeared to overlap, and the smaller signal, from the cosmid cU65A4, was almost entirely contained within the longer green dJ635G19 signal (Figure 4.4a). According to data from the human genome assembly (NCBI Build 34), sequence from these two clones is adjacent, with just a 100bp overlap. As the sequence submitted to the databases for the human genome is non-redundant (except for short overlaps between adjacent clones) the sequence from each genomic clone is often not the whole insert of a particular clone, so it is not uncommon for such overlaps between supposedly adjacent clones to be present (also see section 6.2.2.1.). Data showing the size and relative positioning of genomic clones based on restriction fragment fingerprint data available from the Sanger Institute showed that dJ635G19 and cU65A4 do overlap substantially, similar to what had been seen with fibre-FISH (Figures 4.4a and 4.5.). Fibre-FISH using both cU65A4 and dJ635G19 was carried out on fibres from the affected male in the family as well as on normal cell lines, and the same relationship between the two clones was observed in both cases.

4.3.1.2. Fibre-FISH for the distal breakpoint region

Fibre-FISH results from experiments co-hybridising genomic clones cV362H12 and dJ839M11 showed two main signals of similar size lying next to each other without any overlap (Figure 4.4b). However, there was an additional smaller signal from cV362H12 (red) present at the other end of the green dJ839M11 signal (Figure 4.4b).

As much of the region covered by the sequence submission for clone dJ839M11, as well as the distal part of cV362H12, has been found to contain inverted repeat sequences (A1a and A3) as part of LCR-PMDA, it is probable that this extra site that cV362H12 hybridises to represents part of the A3 repeat unit (Figure 3.4.). Fibre-FISH using both cV362H12 and dJ839M11 was carried out on fibres from the affected male in the family as well as on normal cell lines, and the same relationship between the two clones was observed for both normal and 1:9 fibres.

4.3.2. Demonstrating tandem duplication breakpoint junctions by fibre-FISH

Fibre-FISH experiments were carried out using all four combinations of one proximal and one distal genomic clone, on stretched DNA fibres from both a cell line from the boy 1:9 and a normal lymphoblastoid cell line. As the proximal and distal breakpoint regions are normally located approximately 800Kb apart in normal sequence, no relationship should be seen between a proximal clone (dJ635G19 or cU65A4) and a distal clone (cV362H12 or dJ839M11) when hybridised together to fibres, as they will be too widely separated. However, with all four combinations of a proximal and a distal clone on fibres from a cell line taken from the affected boy, a consistent relationship was seen between the pair of clones (Figure 4.6b). This provided additional confirmation that the duplication was a tandem rearrangement within this family, and was consistent with the positions of breakpoints as determined using UPQFM-PCR and interphase FISH (Figures 4.3. and 4.6.). Hybridisations using all these four combinations of proximal and distal clones on fibres taken from a normal cell line showed no relationship between any of the proximal and distal clones.

When cosmid cV362H12 was hybridised along with either cU65A4 or dJ635G19, a gap was seen between the signals, and the red cV362H12 signal generally appeared slightly larger than the green cU65A4 signal, and of a similar size to the dJ839M11

signal (Figure 4.6b). When dJ839M11 was used in the fibre-FISH experiments, it was seen adjacent to either of the two proximal clone signals, showing the breakpoint between distal and proximal sequence (Figure 4.6b). This was consistent with the distal duplication breakpoint lying within dJ839M11, most probably within the proximal half, as the junction signals from this clone appeared relatively short (Figure 4.6b-d). Both clones at the proximal end, cU65A4 and dJ635G19, were found adjacent to dJ839M11 on fibres from 1:9 (Figure 4.6b). As there is a large overlap between these two clones, and they both appeared the same approximate size as dJ839M11 at the breakpoint, it seemed most likely that the breakpoint was near to the sequence junction between dJ635G19 and cU65A4 (Figures 4.4., 4.5. and 4.6b,c).

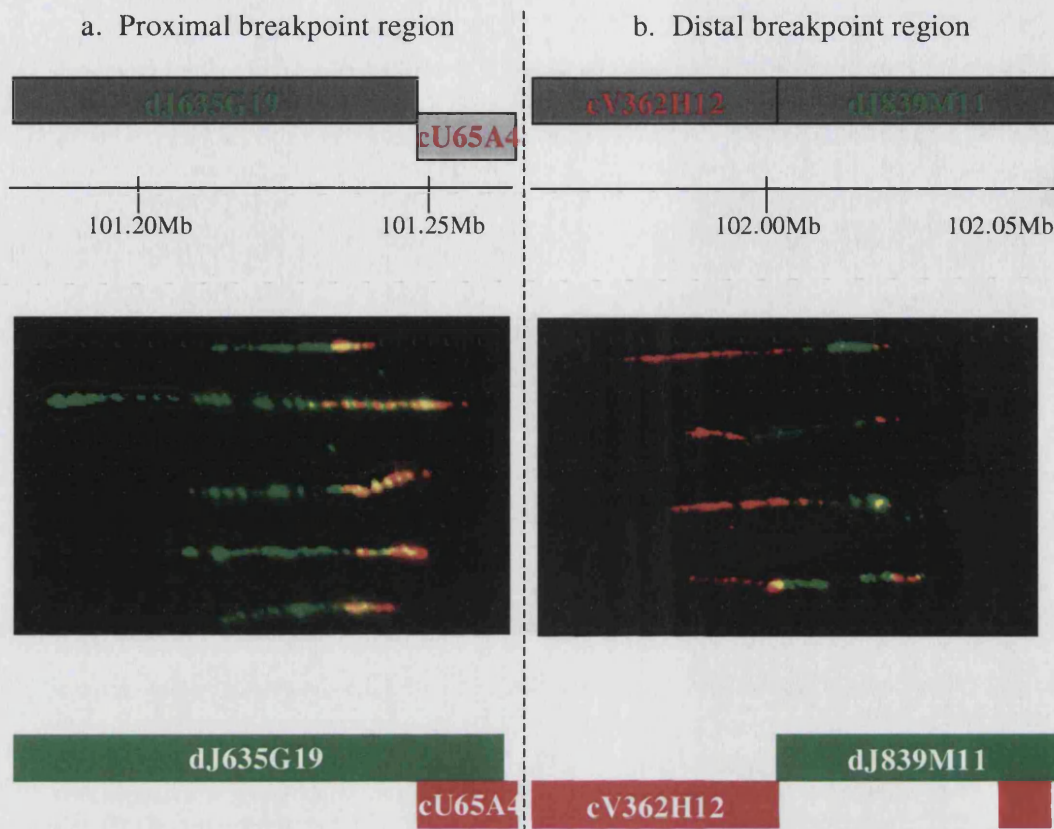


Figure 4.4. Fibre FISH on normal cell lines. The position on the X chromosome is shown as a distance in Mb from the most telomeric Xp sequence (taken from Ensembl release 20.34c.1). The diagram of the clones above the fibres shows the clones as taken from Ensembl. The diagram of the clones below the fibre pictures show the relationship between the clones as based on the fibre-FISH data, the colour of the box representing each clone corresponds to the colour of each labelled clone hybridised to the fibres. (a.) The two clones at the proximal duplication breakpoint. (b.) The two clones closest to the distal duplication breakpoint. The fibre images shown are composites of several different fibres from the same experiment. Where red and green signals overlap, a yellow signal results. Several images of fibres were captured for each experiment, but only 4-5 individual fibres are shown from each hybridisation for simplicity. The fibres shown are representative of the images captured from each slide.

Detailed view: X:98.0Mbp-100.0Mbp

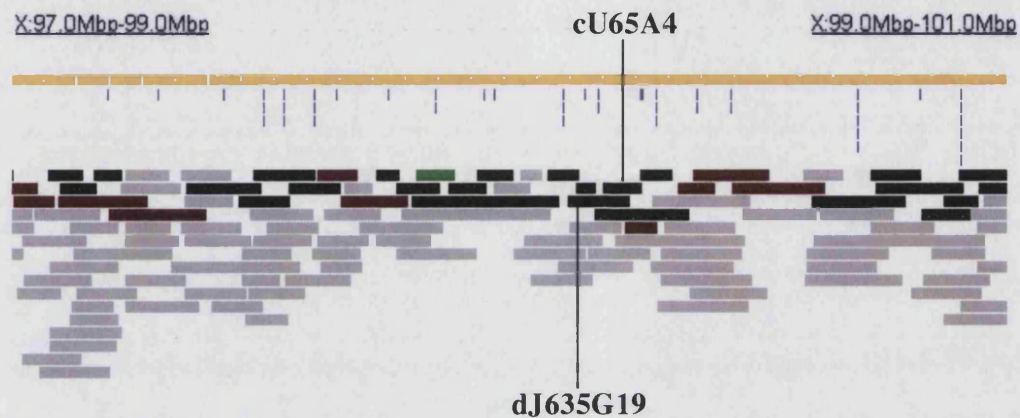


Figure 4.5. Screenshot from Sanger Institute X chromosome FPC (fingerprinted contig) map based on restriction analysis of the whole clones showing relative positions and sizes of genomic clones, including dJ635G19 and cU65A4, in a 2Mb window. Relative sizes of genomic clones are shown by the horizontal boxes, which are colour-coded according to the sequencing status at the time this map was last modified (on 12/12/2002 according to information on the website).

URL: <http://www.sanger.ac.uk/cgi-bin/humace/fpcwebmap.cgi?mode=map&map=bac.X.98.html>

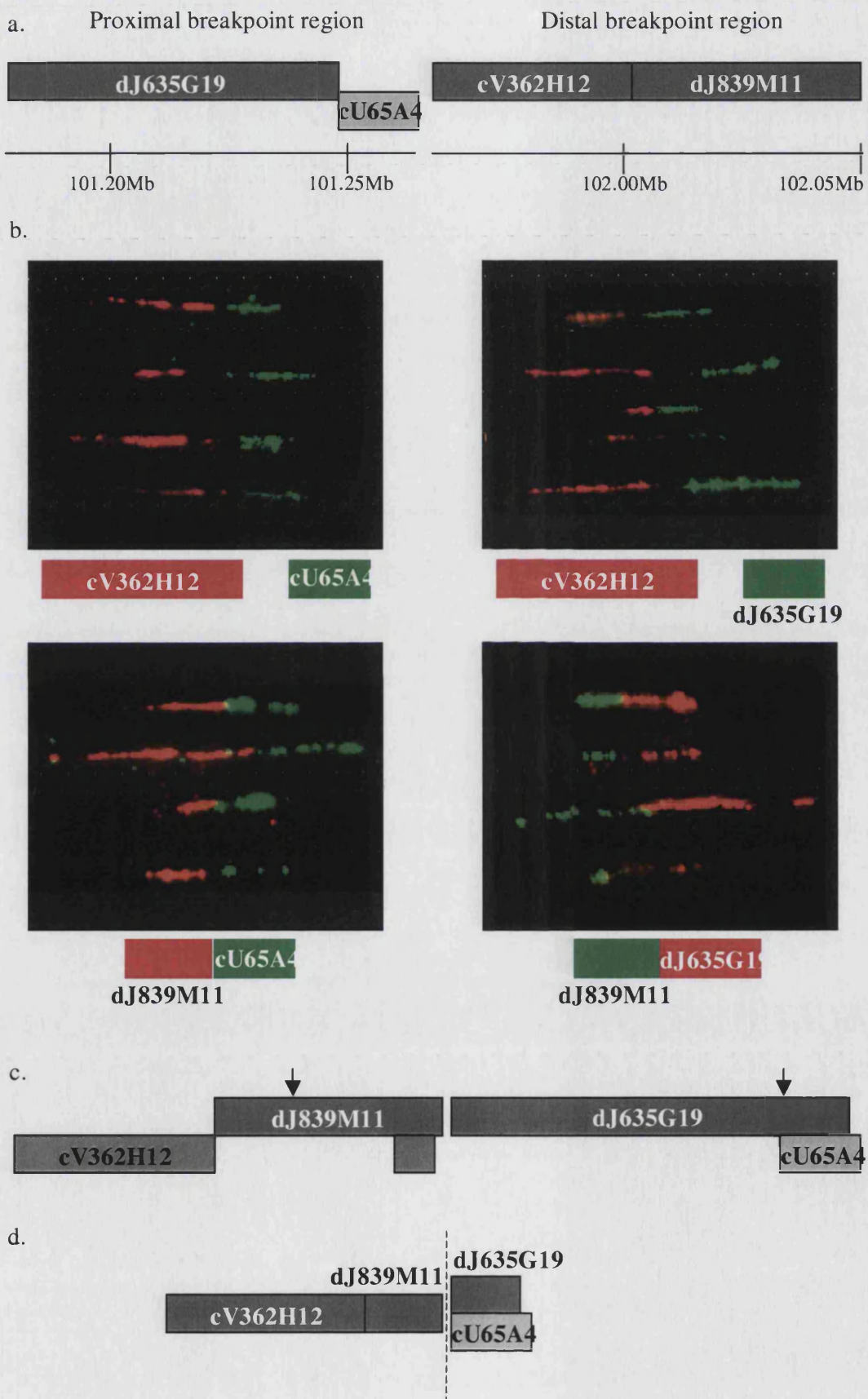


Figure 4.6. Fibre FISH to confirm tandem rearrangement in family 1.

Figure 4.6. Legend

Fibre FISH to confirm tandem rearrangement in family 1. Fibre FISH images are shown on patient 1:9 fibres using clones from both the proximal and distal breakpoint regions. (a.) represents the relative sizes of the clones used in the hybridisations, as based on submitted sequence data, and position on the X chromosome, in Mb from the most centromeric Xp sequence, Ensembl version 20.34c.1. The *PLP1* gene is located between 101.80-101.82Mb in this version of the browser. (b.) Composite images of fibres taken from four separate experiments on 1:9 fibres, each combining different proximal and distal clones. Underneath each fibre-FISH image is a representation of the relationship between these clones, as deduced from the fibre-FISH data. The two clones in each hybridisation are shown as coloured boxes, and the colour of each clone in the diagram is the same as that clone in the corresponding FISH picture. (c.) The likely position of both breakpoints based on fibre-FISH data, indicated by the arrows. Relative position of genomic clones in this diagram is based on previous fibre-FISH experiments shown in Figure 4.2. (d.) The probable orientation of clones relative to the breakpoint, shown by the dashed line. Several images of fibres were captured for each experiment, but only 4-5 individual fibres are shown from each hybridisation for simplicity. The fibres shown are representative of the images captured as a whole from each slide.

4.4. Long-range PCR and direct sequencing to span the breakpoint junction

The UPQFM-PCR and fibre-FISH data were consistent with each other, both placing the proximal duplication breakpoint between the very distal end of dJ635G19 and the proximal part of cU65A4 (Figures 4.3. and 4.6.). The distal breakpoint was similarly mapped using both methods to the proximal part of the genomic clone dJ839M11 (Figures 4.3. and 4.6.). As the results from the fibre-FISH experiments had confirmed that the duplication was tandem in orientation, and shown that the breakpoints did lie in the regions mapped by UPQFM-PCR, long range PCR (LR-PCR) was attempted in an effort to obtain sequence spanning the breakpoint.

4.4.1. Long-range PCR across duplication breakpoint

LR-PCR primers (30mer oligonucleotides) were designed within and close to the breakpoint regions as had been found by UPQFM-PCR (Figure 4.3.). Initial LR-PCR experiments were carried out using all combinations of distal primers from cV362H12 and dJ839M11 with proximal primers mapping to cU65A4 (Figure 4.7.). These initial LR-PCR combinations did produce products in some combinations (Figure 4.7.). A reaction using primers 10069F (from dJ839M11) and 14986F (from cU65A4) resulted in a band approximately 3Kb in size that was not present in normal individuals (Figure 4.8.). There was an additional larger band seen in the lanes for reactions where normal DNA was used, but as the primer in cU65A4, 14986F, consisted of sequence that was entirely contained in an L1 LINE repeat unit, the presence of extra bands was not unexpected (Figures 4.8. and 4.10., Table 4.3.) (Smit *et al.*, 1995). These extra bands were not seen in the lanes using DNA from family 1, which is probably as a result of preferential amplification of the smaller PCR product that is only produced when this breakpoint is present (Figure 4.8.).

4.4.2. Sequencing the duplication breakpoint in family 1

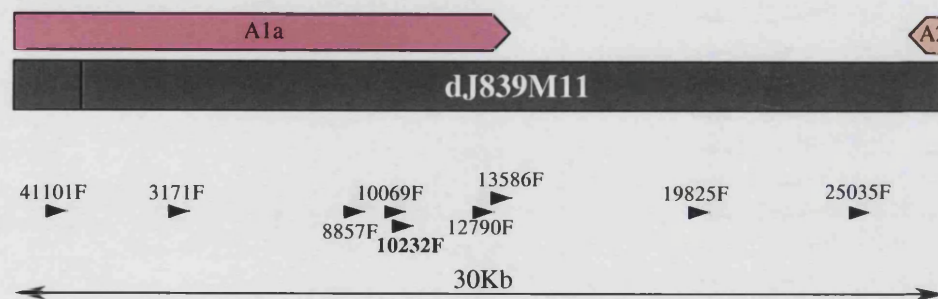
The 3Kb band was gel purified and sequenced (see sections 2.2.1.4.4. and 2.2.4.). Using the nested distal sequencing primer 10232F, sequence across the breakpoint in this family was obtained (Figures 4.7. and 4.9.). BLASTn analysis showed that 543bp of the sequence matched the distal clone dJ839M11, with the similarity ending at position 10809 (Figure 4.9.). This distal sequence was then followed by approximately 100bp of sequence mapping proximal to *PLP1*, starting at position 69210 within genomic clone dJ639G19 (Figure 4.9.). There was a 3 base-pair overlap (CAG) between the two sequences (Figure 4.9.). The reverse sequence across the duplication breakpoint in this family was obtained by using one of the UPQFM-PCR primers used to map the duplication (69265R, one of the pair that amplified the region between positions 69048-69265 from clone dJ635G19) as a sequencing primer in a nested reaction (Figure 4.7.). Assuming that there were no further rearrangements within the duplicated region in family 1, the size of the duplication, based on the NCBI Build 34 human genome assembly, was 765Kb.

4.5. Interspersed repetitive sequences and G+C content at the duplication breakpoints in family 1

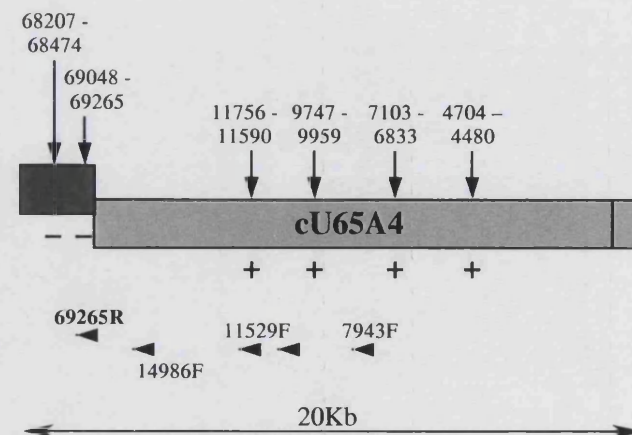
Both the proximal and distal duplication breakpoints were found to map within common human interspersed repetitive sequences (Figure 4.10., Tables 4.3. and 4.4.). The proximal breakpoint was in a L1PA13 element (Table 4.3.) (Smit *et al.*, 1995). The distal breakpoint in dJ839M11 fell within a long terminal repeat (LTR) from the MLT1H1 subfamily (Smit, 1993;Jurka, 2000). The 5Kb sequence immediately around the proximal duplication breakpoint was rich in interspersed repeats (81.1%), mainly due to the presence of L1 repetitive elements near the breakpoint, and had a G+C content of 45.08% (Repeatmasker, RepBase v7.4) (Figure 4.10 and Table 4.3.). The 5Kb sequence immediately around the distal duplication breakpoint consisted of

24.56% known human interspersed repeats and had a G+C content of 47.44% (data obtained from Repeatmasker, RepBase v7.4). The G+C content of both breakpoint regions was higher than the G+C content of either of the 1Mb regions proximal or distal to *PLP1* (40.60% and 40.43% respectively).

a. Distal breakpoint region



b. Proximal breakpoint region



c. LR-PCR products



Figure 4.7. Long-range PCR across the tandem breakpoint in family 1. (a.) and (b.) show the distal and proximal breakpoint regions, as shown in Figure 4.3. The positions and orientation of the various long-range and sequencing primers are shown by the arrows underneath the clones in (a.) and (b.). The sizes of LR-PCR products that were obtained with various combinations of primers are shown in (c.). UPQFM-PCR primers used around the proximal breakpoint are shown as vertical black arrows in (b.), with the area within the clone amplified indicated by these primers shown above the arrow. (+) underneath a primer pair indicates that this primer pair was duplicated in family 1, (-) is shown underneath those primer pairs that were single-copy. The relative position of LCRs (A1a, A2; see Figure 3.4.) around the distal breakpoint, are also shown in (a.).

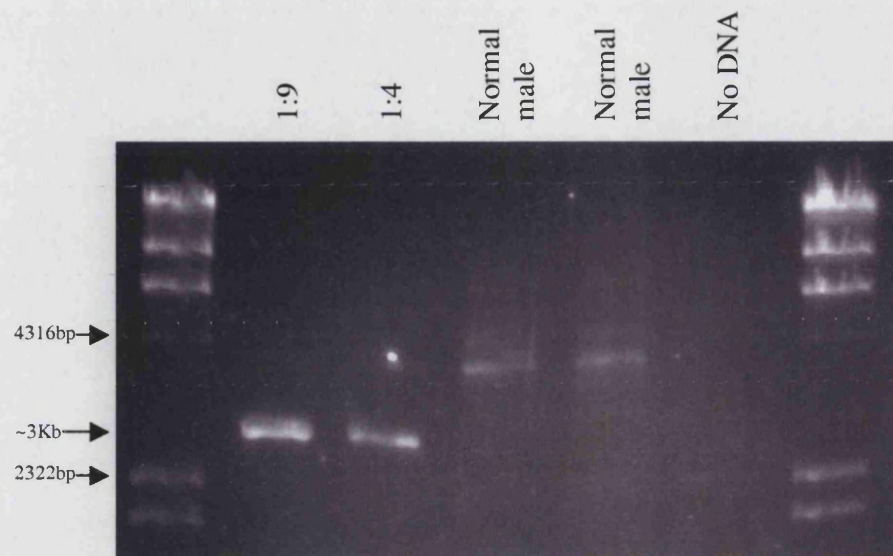


Figure 4.8. Agarose gel stained using ethidium bromide showing LR-PCR product spanning the duplication breakpoint in family 1. LR-PCR was carried out using a primer mapping between bases 14986-15015 in genomic clone dJ839M11 at the distal end, and at the proximal end a primer mapping to between bases 10069-10098 in cosmid cU65A4 was used. Strong bands of about 3Kb in size were seen in both the mother and boy from family 1, and were not seen in the normal controls used in the experiment. λ HindIII digest is used in the two outer lanes as a size standard.

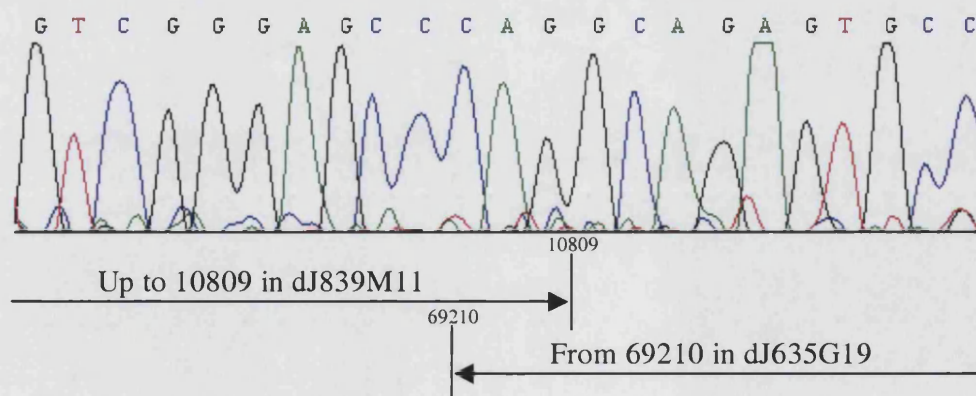
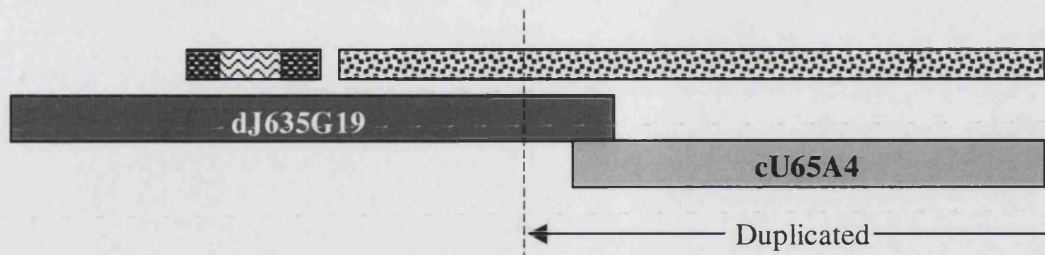


Figure 4.9. Electropherogram showing sequence from the duplication breakpoint in family 1. The regions of homology to the different Xq22 sequences are shown underneath the sequence data, numbers indicate positions within the sequence of each genomic clone.

a. Proximal breakpoint



b. Distal breakpoint



1Kb

KEY

<i>Alu</i>	MER
L1	MIR
LTR	Tandem repeat

Figure 4.10. Interspersed repeat content and other sequence features found in the 5Kb regions centred on the proximal (a.) and distal (b.) duplication breakpoints in family 1. Interspersed repetitive elements are as found by Repeatmasker software, and shown as patterned boxes above the genomic clones in the region. The submitted sequences for dJ635G19 and cU65A4 overlap by 100bp, as shown in (a.). For more information about these repeats, see Tables 4.3 and 4.4. The distal breakpoint is also close to an Xrep enhancer-like sequence, shown as a red box in (b.) (see section 3.4.6. and Figure 3.4.)

Distance from breakpoint (bp)	Repeat type	Position in repeat consensus sequence	Orientation
(-) 1647-1490	MLT1A0 (LTR/MaLR)	1-151 (151/365)	+
(-) 1489-1189	<i>AluSc</i> (SINE/ <i>Alu</i>)	1-299 (299/309)	+
(-) 1189-998	MLT1A0 (LTR/MaLR)	151-365 (206/365)	+
(-) 905-1864 (+)	L1PA13 (LINE/L1)	2-2867 (2866/6797)	+
(+) 1865-5389	L1PA13 (LINE/L1)	3263-6816 (630/6816)	+

Table 4.3. Interspersed repetitive elements from the 5Kb region surrounding the proximal duplication breakpoint in dJ635G19. The first column shows how far in base pairs each repeat element is from the duplication breakpoint, (-) indicates that the repeat element is proximal to the breakpoint and (+) that the repeat is distal relative to the breakpoint. The type of repeat and the class of repeat element to which it belongs are given in the second column. The third column shows which portions within the appropriate repeat consensus sequence each repeat has similarity to, and also how many bases out of the total repeat unit consensus is present in the sequence. In the fourth column, the orientation of each repeat is shown; + indicates a repeat is on the forward strand (*i.e.* running from centromere to telomere), - the repeat is on the reverse strand. A partial L1PA13 element was found to coincide with the end of the 5kb segment, so the adjacent sequence was also analysed with Repeatmasker, until the true end of the repeat was found. The L1PA13 repeat element that the duplicated breakpoint is contained within (at position 933 within the consensus) is in bold type.

Distance from breakpoint (bp)	Repeat type	Position in repeat consensus sequence	Orientation
(-) 2433-2331	MER94 (DNA/AcHobo)	35-134 (100/134)	-
(-) 2063-1844	MIR (SINE/MIR)	13-258 (246/262)	+
(-) 1628-1105	<i>Xrep-like enhancer</i>	1-522 (522/529)	
(-) 1467-1276	MIR (SINE/MIR)	46-230 (185/262)	+
(-) 266-42 (+)	MLT1H1 (LTR/MaLR)	37-365 (329/555)	-
(+) 925-994	(TA) _n (Simple repeat)	1-73	-
(+) 995-1164	<i>AluJb</i> (SINE/ <i>Alu</i>)	124-301 (178/302)	-
(+) 1166-1195	(TG) _n (Simple repeat)	1-32	+
(+) 1196-1321	<i>AluJb</i> (SINE/ <i>Alu</i>)	1-124 (124/302)	-
(+) 2401-2498	L1M4 (LINE/L1)	2694-2793 (100/6146)	+

Table 4.4. Interspersed repetitive elements from the 5Kb region surrounding the distal duplication breakpoint in dJ839M11. The first column shows how far in base pairs each repeat element is from the duplication breakpoint, (-) indicates that the repeat element is proximal to the breakpoint and (+) that the repeat is distal relative to the breakpoint. The type of repeat and the class of repeat element to which it belongs are given in the second column. The third column shows which portions within the appropriate repeat consensus sequence each repeat has similarity to, and also how many bases out of the total repeat unit consensus is present in the sequence. In the fourth column, the orientation of each repeat is shown; + indicates a repeat is on the forward strand (*i.e.* running from centromere to telomere), - the repeat is on the reverse strand. The repeat element that the duplicated breakpoint segment is contained within is highlighted in bold type. The distal breakpoint in family 1 is at position 88 within the consensus sequence for MLT1H1. The position of the Xrep enhancer-like sequence relative to the breakpoint is also shown (Riley *et al.*, 1986).

4.6. Searching for similarities between the proximal and distal duplication breakpoints

The two 5Kb regions surrounding the proximal and distal breakpoints were compared against each other to search for any large regions of homology near the breakpoints that could be involved in mediating the duplication. Comparisons were made using both BLAST2 and BLASTz, and only one short region of similarity between the two sequences was found, which mapped to *Alu* repeat elements which were approximately 1Kb distant from either breakpoint (Figure 4.11, Tables 4.3. and 4.4.). Additionally, 100bp regions flanking the duplication breakpoint were also aligned together, using ClustalW. There was no extensive homology found close to the 3bp breakpoint region and no regions of similarity extending for more than 4bp (Figure 4.12.). Although there was 50% similarity seen between dJ635G19 and dJ839M11 proximal to the breakpoint, and 43% similarity was seen between the two sequences after the breakpoint, the nucleotides included in the alignment were only in small clusters (Figure 4.12.).

4.7. Location of proximal duplication breakpoint relative to *PLP1*-proximal repeats

The location of the proximal end of the duplication, at position 69210 within genomic clone dJ635G19, did not fall into any of the short repetitive regions that had been mapped proximal to *PLP1* (see Figure 3.8.). The closest copy of one of these repeats (one of the “P3” grouping), was 4.3Kb proximal to the breakpoint location, and consisted mostly of the *Q9NWD9* gene, one of the *BEX/NADE* family of genes in this region, (see Table 3.9.). The closest gene to the breakpoint that was contained within the duplicated region was a transcript 32Kb away annotated on the Ensembl genome browser as *NM_153333*, which has similarities to the *TCEAL1* gene and is also contained within one of the distal “P4” repeats (see Table 3.11.).

4.8. Location of distal breakpoint within low-copy repeats.

The distal duplication breakpoint was located within one of the LCRs located distal to *PLP1*, namely repeat unit A1a that forms part of LCR-PMDA (Figure 3.4.). It is located within the distal third of this repeat unit, at position 17327 (distance from the start of the repeat). As well as a few interspersed repetitive elements in the 5Kb surrounding the breakpoint, there is also a copy of the Xrep enhancer-like sequence 1105bp proximal to the breakpoint (Table 4.4., Figures 4.10. and 3.4.). As the distal breakpoint region is contained within a large repeated sequence (A1a), similar sequence is also present in other copies of this repeat. The 5Kb surrounding the distal breakpoint at 10809bp within dJ839M11 was compared to the other three large repeats that share sequence identity with A1a (A1b, A2, A3) (Table 4.5). Sequence similarity was found between some of the 5Kb breakpoint region sequence and the other three copies of the distal repeat, but the actual breakpoint sequence was only present in two of the other repeats, A1b and A2 (Table 4.4.). Only the first 1757bp of the 5Kb around the breakpoint were aligned to repeat unit A3, which does not include sequence homologous to the actual breakpoint as it ends before this point in repeat A1a (Table 4.5., Figure 3.4., Table 3.2.). The nearest genes to the distal duplication breakpoint were the various copies of the H2b-like genes, which mapped within the distal LCRs (see section 3.4.5. and Figure 3.4.).

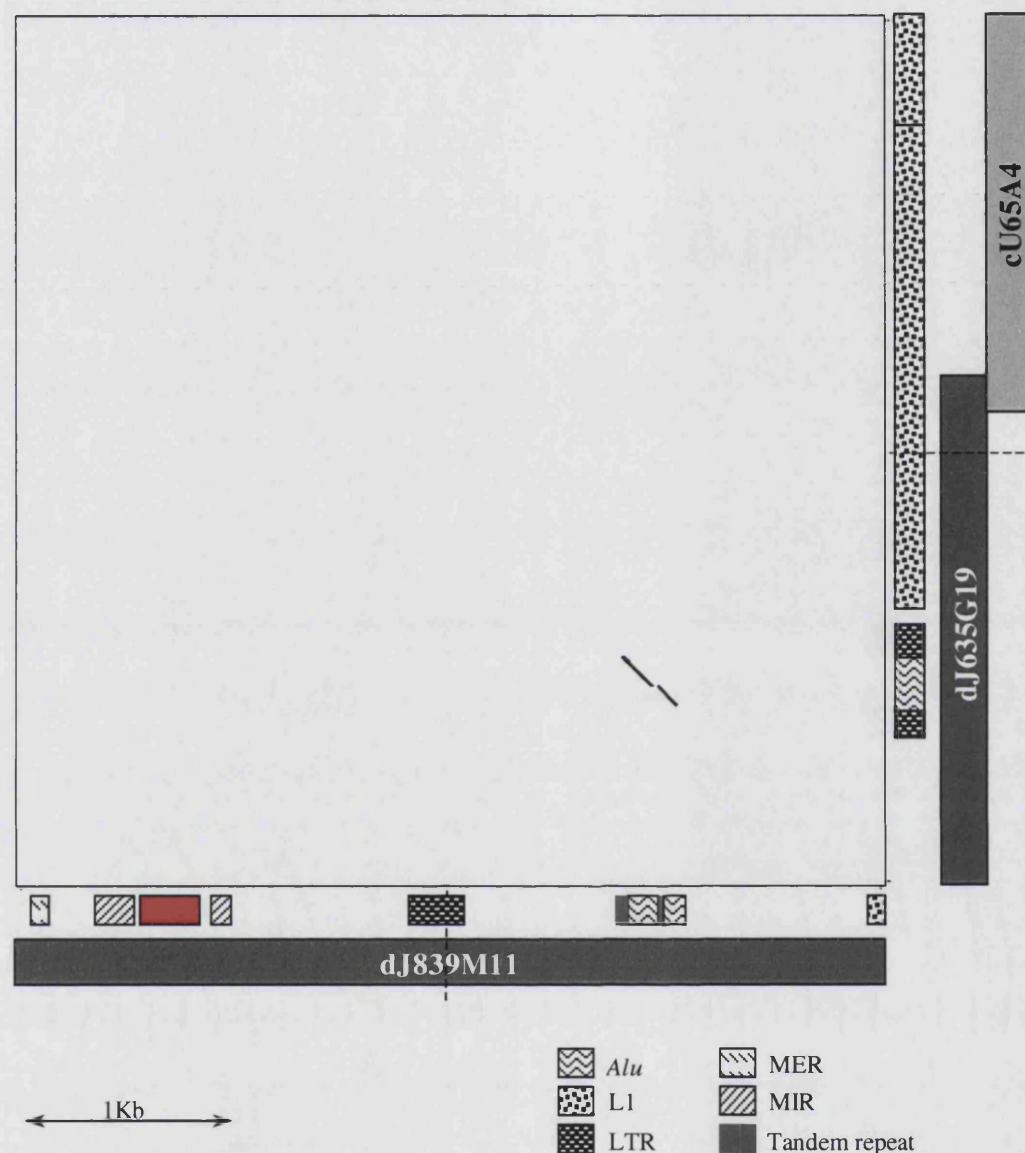


Figure 4.11. Comparison of 5Kb around the proximal and distal breakpoints against each other by BLASTz using the Pipmaker program. The results of the comparison are shown as a dotplot, where any similarities are shown as black dots on the white background. The human genomic clones and interspersed repeat content are shown in the relevant positions next to the dotplot. The position of each breakpoint is indicated by the dashed line. For more expansion of the repeat content, see Figure 4.10. and Tables 4.3. and 4.4. Upwards-facing diagonal lines on the dotplot (/) are indicative of directly repeated sequences; downwards-sloping lines on the dotplot (\) represent areas present in both sequences in an inverted orientation. The position of Xrep is shown by the red box.

dJ635G19 69160 TG**AGTGCA**CCCGCC**CCCTT**CCCC**CAA--CCACCA**TTGTTG**TTG**GAGAGTTG
Junction **GTCCACCATTCTCCATGTGTCAAAGCCATCACAGGC--TTGTCGGGAGC**
dJ839M11 10759 **GTCCACCATTCTCCATGTGTCAAAGCCATCACAGGC--TTGTCGGGAGC**

dJ635G19 GC**CAGGCAGAGTGCCCCACCCGTGTGCGTGGAAATGGCCTCCAGGTGCGAAA** 69259
Junction **CCAGGCAGAGTGCCCCACCCGTGTGCGTGGAAATGGCCTCCAGGTGCGAAA**
dJ839M11 **CCAGAGGGAGGGGCAAAC--AGATGCGACCAATGGGCCAACGGGAAGAGCATTCA** 10859

Figure 4.12. Alignment of duplication breakpoint junction from family 1 and 100bp flanking the two duplication breakpoints. Each breakpoint region was individually aligned by ClustalW to the sequence from the junction breakpoint and the two alignments were then combined together manually. Proximal junction sequence that originated from dJ635G19 is highlighted in red, and nucleotides aligned with this sequence from the genomic sequence around the distal breakpoint are also highlighted in red. Similarly, junction sequence originating from dJ839M11 is highlighted in blue, and any bases from the proximal genomic sequence aligned with this sequence are also shown in blue. The 3bp overlap at the junction between the two sequences is shaded purple.

Repeat unit	Percentage identity	Percentage gaps	Position within repeat unit	Range of 5Kb breakpoint region in alignment
A1b	98.26%	0%	14829-19825	1-5000
A2	93.24%	0%	6866-11457	3-4535
A3	89.70%	0.11%	13103-14856	1-1757

Table 4.5. Comparison of 5Kb surrounding the distal duplication breakpoint in family 1 against the related distal repeat units A1b, A2 and A3. Comparisons and alignments were carried out using the BLASTz algorithm. Positions within the repeat unit are given in base pairs from the beginning of each repeat unit. Not all the 5Kb region around the breakpoint was included in every alignment, the amount that is contained in the alignment is shown in the table.

4.9. Analysis of sequence from breakpoint regions for recombination and rearrangement-associated motifs

4.9.1. 5Kb regions around breakpoints

Various sequence motifs have been reported to be associated with rearrangement breakpoints and 5kb centred on each breakpoint was searched for the presence of various motifs (Table 2.4.).

4.9.1.1. Calculating relative occurrence of motifs in the 5Kb regions around breakpoints

5Kb either side of each duplication breakpoint was examined for the presence of such motifs on both strands using the DNA Pattern Find program (see Table 2.4. for details of each sequence motif). When motifs were found within the two 5Kb regions, the number of times each sequence was found was compared against the number of times it was expected to occur given the nucleotide content of the region (Day and Blake, 1982). The difference between the number of times that a motif was observed and the number of times it was expected to be seen was then expressed as a proportion of the expected occurrence. Those motifs that were found at a frequency more than five times the expected prevalence were considered to be moderately enriched and investigated further. Any motifs that were present at more than 20 times the expected occurrence were classified as being highly enriched in the region. However, just because a motif is not enriched near a breakpoint does not necessarily imply that it is not involved in the breakpoint mechanism, so any motifs that occurred near a breakpoint, whether “enriched” or not, were noted.

4.9.1.2. Matrix attachment regions

It is possible that features of the DNA that have a role in chromosome and chromatin structure may be important in the mechanisms leading to DNA sequence rearrangements. One such structural feature that does play an important part in chromosome and chromatin dynamics are matrix attachment regions (MARs), which anchor chromatin fibres to the nuclear matrix, with the intervening DNA forming 50-100Kb looped domains (Vogelstein *et al.*, 1980). Although the ability of a sequence to function as a MAR can only be confirmed by *in vitro* studies, regions with the potential to act as MARs can be predicted to some degree by *in silico* methods (Singh *et al.*, 1997). One program that can be used to predict MARs is MAR-Wiz, which searches for various motifs and sequence patterns associated with MARs, returning a matrix-binding potential across the region examined (see section 2.2.10.4.). MAR-Wiz was used to determine the matrix-binding potential throughout a 5Kb region surrounding each breakpoint.

4.9.2. Sequence features found in 100bp regions around breakpoints

It is likely that any features of the DNA sequence that may have contributed to the duplication rearrangement will be found near to the breakpoints. For this reason, a short stretch of sequence (100bp) surrounding each breakpoint was searched for local properties of the sequence that could conceivably be involved in a sequence rearrangement at this locus.

4.9.2.1. Alternating purine/pyrimidines, polypurine and polypyrimidine tracts

It has been found that alternating tracts of purines and pyrimidines may stimulate both homologous and illegitimate recombination (Bullock *et al.*, 1986; Boehm *et al.*, 1989; Stary and Sarasin, 1992). Purine/pyrimidine tracts are targets for topoisomerase II and have also been shown to be significantly over-represented close to deletion

breakpoints (Spitzner *et al.*, 1990; Abeysinghe *et al.*, 2003). Purine/pyrimidine tracts can form stretches of Z-DNA, an alternative DNA conformation that may be recombinogenic (Wang *et al.*, 1979; Weinreb *et al.*, 1988; Boehm *et al.*, 1989; Majewski and Ott, 2000).

Homopurine and homopyrimidine stretches of DNA can adopt unusual conformations and may be important in stimulating DNA breakage or recombination (Arnott *et al.*, 1983; Moser and Dervan, 1987; Konopka, 1988). Purine tracts of between 25-39 nucleotides in length have been found to be significantly over-represented at deletion breakpoints, and at translocation breakpoints shorter polypurine tracts (2-23bp) and polypyrimidine tracts (2-44bp) have been found to be significantly over-represented (Abeysinghe *et al.*, 2003). In this study, only homopurine, homopyrimidine or alternating purine/pyrimidine tracts of length 10bp or greater were searched for close to the breakpoints, as this length of sequence equates to roughly one helical turn of the DNA (Ussery *et al.*, 2002).

4.9.2.2. Inverted repeats and secondary structures

It has been suggested that various types of small repeats near breakpoints could contribute to rearrangements by creating secondary structures that stabilise loops between strands of DNA (Chuzhanova *et al.*, 2003). Secondary structure present in the DNA may also lead to rearrangements by causing breakages in the DNA molecule. DNA hairpin structures, which are formed by adjacent inverted repeated sequences, can be specifically cleaved by topoisomerase II (Froelich-Ammon *et al.*, 1994). Several different types of repeats were searched for in the 100bp surrounding each breakpoint, as well as in the rearranged junction sequence, using the Oligorep algorithm (see section 2.2.10.5. and Table 2.3.). These types of repeat were classified as direct (e.g. 5'ACGTA^{3'}....5'ACGTA^{3'}), inverted (e.g. 5'ACGTA^{3'}....5'TACGT^{3'}),

symmetric (e.g. 5'ACGTA^{3'}....5'ATGCA^{3'}) and complementary repeats (e.g. 5'ACGTA^{3'}....5'TGCAT^{3'}) (Chuzhanova *et al.*, 2003).

4.10. *In silico* analysis of sequence in 5Kb regions around breakpoints

4.10.1. Recombination/rearrangement-associated motifs

Many short DNA sequence motifs thought to be involved in various rearrangements and recombination events were found in the 5Kb regions centred on both the duplication breakpoints in family 1 (see section 4.9.1.1., Table 2.3. and Appendix C). Some of the motifs searched for were found at a higher rate than expected within the 5Kb regions flanking the breakpoints. These included scaffold attachment sequences and *Saccharomyces cerevisiae* autonomously replicating sequence (ARS) consensus (WTTTATRTTTW) in the 5Kb region around the proximal duplication breakpoint (Broach *et al.*, 1983; Dobbs *et al.*, 1994). More sequence motifs present within the 5Kb of genomic sequence surrounding the distal duplication breakpoint in family 1 were found to be over-represented, based on the nucleotide frequency in this region (Appendix C). Most of these motif sequences (Yeast ARS, scaffold attachment regions) were not located particularly close to the duplication breakpoint, so it was considered unlikely that these sequences could have been involved in the duplication mechanism (Appendix C). Some of the sequence motifs were found relatively close to the breakpoint including a copy of the heptamer V(D)J joining motif (CACTGTG), 160bp 5' to the breakpoint and a chi-like sequence known to occur within the repeat units of human minisatellites (GCWGGWGG), 224bp 5' from the breakpoint (Max *et al.*, 1979; Akira *et al.*, 1987; Krowczynska *et al.*, 1990). A few 3bp sequence motifs, such as topoisomerase I cleavage sites (CTY, GTY) and DNA polymerase α pause (GAG, GCS) were found close to the breakpoint, as well as occurring frequently throughout the surrounding sequence (Appendix C) (Been *et al.*, 1984; Kunkel, 1985). Although these short sequences are found very frequently in genomic sequence, it is

still possible that these motifs could be important in generating DSBs and rearrangements in this region. One copy of the murine MHC recombination hotspot sequence (CAGRCAGR) was found to overlap with the proximal duplication breakpoint (Appendix C) (Steinmetz *et al.*, 1986;Shiroishi *et al.*, 1995). One half of this motif (CAGR) has also been associated with deletions in the mouse genome, and a copy of this sequence also overlaps with the distal breakpoint (Steinmetz *et al.*, 1986).

4.10.2. Potential MARs near duplication breakpoints in family 1

The MAR-Wiz program was used to search for potential matrix attachment sites close to both breakpoints. No regions that reached the threshold MAR potential set by MAR-Wiz were found near the proximal breakpoint, but a potential MAR was located 1.2Kb distal from the distal duplication breakpoint (Figure 4.13.). This MAR-potential value can be quite dependent on the local sequence context, so a larger (20Kb) region around the distal breakpoint was also tested for the presence of MARs (data not shown), and the same region was still predicted to be a MAR (Namciu *et al.*, 2004).

4.11. *In silico* analysis of sequence from 100bp regions flanking breakpoints

4.11.1. Alternating purine/pyrimidines, polypurine and polypyrimidine tracts

100bp regions surrounding the proximal and distal duplication breakpoint were examined for the presence of purine/pyrimidine, homopurine and homopyrimidine tracts of 10bp or greater (see 4.9.2.1.). Around the proximal breakpoint, there was one polypyrimidine tract 11bp in length located 26bp 5' to the breakpoint and one 10bp alternating purine/pyrimidine tract found 18bp distal to the breakpoint within the duplicated region (Figure 4.14.). Only one polypurine tract was found near the distal duplication breakpoint, an 11bp region that started within the 3bp of overlapping

sequence at the breakpoint and extended into the non-duplicated sequence distal to the breakpoint (Figure 4.14.).

4.11.2. Repeats and secondary structures

Few repeats with the potential to form secondary structures were found in the 100bp surrounding the two duplication breakpoints from family 1 using the Oligorep program (see section 4.9.2.2.). Near the proximal breakpoint, two short symmetric repeats, each 7bp in length, were found, but no repeats were found near the distal breakpoint (Figure 4.15.). When the sequence spanning the duplication breakpoint was considered, as well as one of the short symmetric repeats already described, a longer direct repeat was also present, with one copy either side of the breakpoint, 14 nucleotides in length, with 11 of the bases having identity between the two repeats (Figure 4.15.). Overall, slightly more nucleotides were within repeated regions in the rearranged sequence than the original genomic sequence (Figure 4.15.). It is possible that the presence of some of these repeats, which have the potential to form secondary structures, may have helped to stabilise the rearrangement (Chuzhanova *et al.*, 2003).

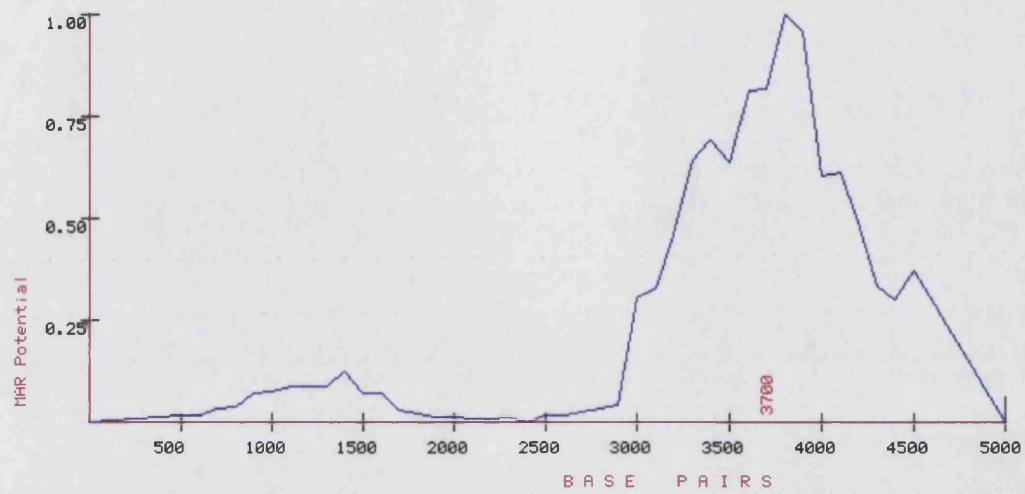


Figure 4.13. MAR potential in the 5Kb surrounding the distal duplication breakpoint in family 1 (at position 2500 in this sequence) as found by MAR-Wiz.

Proximal duplication breakpoint 69160-69259bp in dJ635G19

TGAGTGCACCCCGCCCCCTTCCCCAACCACCATTGTTGTTGGAGAGTTGGCAGGCAGAGTGCCCCACCCGTGTGCGTGGAATGGCCTCCAGGTGCGAAA
YRRRYRYRYYYYR**YYYYYYYYYY**RRYYRYRYRYRYRYRRRRRRYYRRYRRRYRRRRYRYYYYRYYRYRYRYRRRRYRRYYYYYRRRYRYRRR

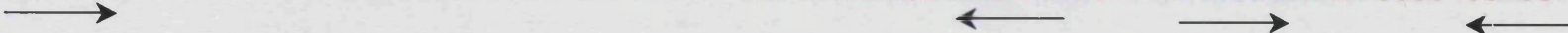
Distal duplication breakpoint 10759-10858bp in dJ839M11

GTCCACCATTCTCCATGTGTCAAAGCCATCACAGGCTTGTCGGGAGCCAGAGGGAGGGGCAAACAGATGCGACCAATGGGCCAACGGGAAGAGCATTCA
RYYRYRYRYYYYRYRYRYRRRRYYRYRYRRRRYYRYRRRRYY**RRRRRRRRRR**YRRRYRRRYRYRYRYRRRYRRRYRRRRRRRRRYRYYYR

Figure 4.14. Purine and pyrimidine content of 100bp regions surrounding the proximal and distal duplication breakpoints in family 1. The 100bp of sequence flanking the two duplication breakpoints is shown on the top of the pairs, and the classification of each nucleotide (R – purine, Y – pyrimidine) is shown on the lower line of the pairs. Tracts of alternating purines/pyrimidines (R/Y), or pyrimidines/purines (Y/R), of 4 nucleotides or greater, are underlined. Homopurine and homopyrimidine runs of 10 nucleotides or greater are in bold type. Sequence from dJ635G19 that is duplicated at the proximal duplication breakpoint is shaded red, while duplicated sequence from the distal breakpoint region is shaded blue. The 3 nucleotides that overlap at the breakpoint are shaded purple, and sequence that is not duplicated in family 1 is shaded grey.

Proximal duplication breakpoint 69160-69259 in dJ635G19

TGAGTGC**CACCCCG**CCCCCTTCCCCAACCACCATTGTTGTTGGAGAGTTGGCAGGCAGAGT**GCCCCACCCGTGTGCGTGGAAATGGCCTCCAGGTGCGAAA**



Distal duplication breakpoint 10759-10858 in dJ839M11

GTCCACCATTCTCCATGTGTCAAAGCCATCACAGGCTTGTCTGGGAGCCCAGAGGGAGGGGCAAACAGATGCGACCAATGGGCCAACGGGAAGAGCATTCA

Duplication junction

GTCCACCATT**CTCCATGTGTCAA**AGCCATCACAGGCTTGTCTGGGAGCCCAGGCAGAGTGCCCCACCCGTGT**GCGTGGAAATGGCCTCCAGGTGCGAAA**




Figure 4.15. Inverted repeats around the duplication breakpoints in family 1, the duplication junction fragment and the reciprocal rearrangement. The normal sequence is shown, from the proximal duplication breakpoint (dJ635G19) and also from the distal duplication breakpoint (dJ839M11). The recombinant deletion breakpoint is shown with the proximal half in red (originating from bA346E8) and the distal half shown in blue (originating from dJ203P18). The 3 nucleotides that form the duplication breakpoint are shaded purple. Nucleotides that form part of repeat sequences are shown in bold. The type of repeat found is shown by the arrows above and below the sequence: direct (→) and symmetric (↔). Repeats within the sequence were searched for using Oligorep (see section 4.9.2.2.).

4.12. Discussion

From analysis of the duplication breakpoint in family 1, it seems most likely that the rearrangement in this family was mediated by non-homologous end joining, as has been found to be the case in most other PMD rearrangements where the breakpoints have been sequenced (Inoue *et al.*, 2002;Iwaki *et al.*, 2003;Woodward *et al.*, in preparation).

4.12.1. Interphase FISH and duplication mapping

The distal duplication breakpoint in family 1 has been shown in this study to be in the proximal third of dJ839M11 (Figures 4.3. and 4.9.), but this contradicts the previous work done with interphase FISH that had shown that the next most distal clone, cU240C2, was duplicated in this family (Figure 4.1.) (Woodward *et al.*, 1998;Woodward *et al.*, 2000).

One possible explanation for the apparent duplicated signals seen for cU240C2 is DNA replication. Scoring replicated signals as duplicated signals is a common problem associated with interphase FISH (see section 4.2.1.). Additionally, if there is a replication origin in this region, this may lead to a higher frequency of duplicated signals originating from DNA replication, especially if this region tends to replicate before the X centromere region, because if extra copies of the X centromere are visible, the nucleus is not scored (see section 4.2.1.). cU240C2 contains two copies of the Xrep motif, within the distal LCR PMDB as part of repeat units A3 and A1b, which has been suggested to function as an origin of DNA replication (see section 4.12.1.2. and Figure 3.4.) (Riley *et al.*, 1986).

Possibly the most likely explanation for the discrepancy between the FISH and breakpoint sequence data, however, is the existence of the various distal LCRs.

Additional hybridisation signals during interphase FISH could result from hybridisation of cU240C2 (containing most of LCR-PMDB) to genomic sequence from LCR-PMDA (Figure 3.4.). The distal duplication breakpoint is located within one of the repeats that make up LCR-PMDA, A1a, so a substantial part of this repeat unit (17.3Kb) is duplicated in family 1 (Figure 4.7., Table 4.5.). Much of repeat unit A1b maps to within cU240C2, and A1a and A1b are over 99% identical at the sequence level (Figure 3.4., Table 3.3.). It therefore seems probable that the duplicated signals seen when using cU240C2 in interphase FISH on this family were due to the cosmid hybridising to the partially duplicated copy of the A1a repeat unit, which is probably large enough to result in a good signal. In normal circumstances, although cU240C2 might also be hybridising to sequences from the more proximal LCR-PMD, this would be too close to the actual cU240C2 signal for the additional signals to be resolved in interphase nuclei, but in family 1, where the duplicated region is relatively distant from the original copy, the extra signal is visible. If the genomic DNA were examined with a higher resolution using fibre-FISH, as has been carried out in this study using different genomic clones, then the additional hybridisations to LCR-PMDA that have been postulated may be detectable. Fibre-FISH using the cosmid cV362H12 (which contains part of repeat unit A1a) on normal individuals did show an additional smaller signal separated from the main signal for this probe, which could correspond to sequence from the A1b repeat unit (see section 4.3.1.4., Figures 3.4. and 4.4.). It is likely that the reciprocal situation does occur when cU240C2, which contains the repeat unit A1b, is hybridised to DNA, with additional ectopic signals corresponding to cU240C2 sequence additionally hybridising to A1a (Figure 3.4.).

4.12.2. NHEJ

NHEJ, as opposed to homologous recombination, is the predominant pathway for the repair of double-strand breaks (DSBs) in humans (Lieber *et al.*, 2003). The breakpoint in family 1 bears all the hallmarks of the NHEJ repair process, as the two sequences joined together at the duplication junction share no appreciable homology, apart from a short (3bp) overlap at the actual junction (Figures 4.9. and 4.11.). Such microhomologies between sequences may help to stabilise the two ends during the process of NHEJ (Roth and Wilson, 1986). Microhomologies are not necessary for NHEJ to occur, but the NHEJ process favours joining at short regions of homology (Lieber *et al.*, 2003). DSBs are usually processed before repair by NHEJ, as the ends of the DSBs may be incompatible. This end-processing step may involve removal of nucleotides by nucleases and also the filling in of gaps on one or both strands by polymerases (reviewed in Lieber *et al.*, 2003).

4.12.2.1 Mechanisms for DSB formation

As it appears that NHEJ has occurred in this instance to join the two ends of the duplication breakpoint together in this family, it follows that when the rearrangement occurred, it will probably have been triggered by DSBs, either at one end of the duplication or at both. When the two ends of a DSB remain in close proximity following breakage, the NHEJ pathway is able to join the correct ends together, thereby maintaining genomic integrity, albeit with the likely loss of a few nucleotides (reviewed in Lieber *et al.*, 2003). However, if DNA ends from two different DSBs are aberrantly joined together, then a large-scale rearrangement such as described in family 1 in this chapter may result. Due to the mechanisms involved in processing ends during NHEJ, it is likely that any DSBs triggering the duplication event will have occurred 5' of the proximal duplication breakpoint and 3' of the distal breakpoint. Some motifs and other features of the sequence have been found that

could conceivably be involved in the formation of DSBs, such as a sequence present in a murine recombination hotspot (CAGRCAGR), and an 11bp polypurine tract (see sections 4.10.1. and 4.11.1.). However, it is also possible that DSBs could be caused by another factor, such as ionising radiation, or intracellular compounds like reactive oxygen species (reviewed in Lieber *et al.*, 2003).

4.12.2.2. Xrep, replication origins and DSBs

However, there is one feature of the sequence around the distal breakpoint that may be implicated in the mechanisms behind the rearrangement. The distal duplication breakpoint lies within one of the distal LCRs (A1a), and over 50% of the other duplication breakpoints that have been mapped so far in PMD families are also within the distal LCRs (Woodward *et al.*, in preparation) (Figure 3.4.). It is possible that these LCRs could be involved in the initiation of DSBs in this region and thereby stimulate rearrangements (Inoue *et al.*, 2002). The proximity of the Xrep enhancer-like sequence to the distal duplication breakpoint may not be coincidental, as this short stretch of 973bp, could be involved in stimulating the rearrangement (Table 4.4.) (Riley *et al.*, 1986). It has been suggested that Xrep could function as an origin of replication (see section 3.4.6.) (Riley *et al.*, 1986). One potential source of DSBs during the cell cycle is from the collapse of replication forks, which can occur if the DNA polymerase encounters a single strand nick, or a damaged nucleotide (Kuzminov, 2001). However, Xrep has not been shown as yet to act as an origin of replication in human or even mammalian cells, although the collapse or stalling of a replication fork does not necessarily have to be close to an origin of replication (Riley *et al.*, 1986). Some mammalian origins of replication have been shown to overlap with MARs, and there is a region with the potential to form a MAR 800bp 3' to the distal duplication breakpoint in family 1, so it is possible that collapse of replication forks proceeding from this point could also lead to a DSB (Girard-Reydet *et al.*,

2004). However, this potential MAR has only been predicted according to the sequence data from the region, and predicted MARs do not necessarily correspond to genuine MARs (Namciu *et al.*, 2004)

4.12.3. Origin of the duplication in family 1

The event that led to the duplication event in family 1 must have occurred within cells leading to or part of the germ line to be transmitted to 1:4 and to her son 1:9. Most *PLP1* duplications originate from male germ cells, with an estimated bias towards the male germ line of between 9-fold and 11-fold (Mimault *et al.*, 1999). Previous analysis of polymorphic markers within the duplicated region has shown that 1:9 was homozygous at the microsatellite marker DXS1106, which is contained within the duplicated region, while his mother was heterozygous at this locus (Woodward *et al.*, 1998). Both 1:4 and 1:9 were homozygous for the other two markers which were genotyped in this family that do map to within the duplicated region (Woodward *et al.*, 1998). Homozygosity of alleles at a polymorphic marker within the duplicated region suggests that the rearrangement may have been an intrachromosomal event, which is compatible with a mutational event taking place during spermatogenesis, as is most common for *PLP1* duplications (Woodward *et al.*, 1998;Mimault *et al.*, 1999). The production of germ cells and the process of meiosis differs between spermatogenesis and oogenesis, and this is reflected in a parental origin bias for different types of mutations. Chromosome non-disjunction, leading to aneuploidies, is much more likely to occur during oogenesis, during the first meiotic division (Chandley, 1991). Point mutations and structural rearrangements are events that more frequently originate in males during spermatogenesis (Chandley, 1991;Uematsu *et al.*, 2002).

4.12.4. DSBs, recombination and meiosis

The production of DSBs is a process inherent to meiosis, as DSBs are necessary for synapsis of homologous chromosomes and meiotic recombination (Mahadevaiah *et al.*, 2001). Recombination between chromosomes does not occur randomly throughout a particular chromosome; instead recombination events have been found to occur in small “hotspot” regions (1-2Kb), which are flanked by large regions of recombination-suppressed DNA that form linkage disequilibrium blocks (reviewed in (Kauppi *et al.*, 2004)). The distribution and numbers of recombination hotspots in yeast is similar to the number and distribution of DSBs in yeast meiosis, and it is likely that recombination hotspots are equivalent to DSB hotspots (Baudat and Nicolas, 1997;Kauppi *et al.*, 2004). It is probable that recombination hotspots in mammals, such as humans, are also correlated with meiotic DSB formation hotspots (Kauppi *et al.*, 2004). Although most of the genetic material in the sex chromosomes does not recombine during male meiosis, as pairing can only occur within the homologous psuedoautosomal region, DSB breaks are still initiated prior to synapsis throughout the sex chromosomes, and appear to persist until mid pachytene (Ashley *et al.*, 1995;Plug *et al.*, 1996;Mahadevaiah *et al.*, 2001). Most DSBs initiated prior to synapsis and meiotic recombination, especially within the sex chromosomes, do not lead to a crossing over event and are subsequently repaired (Kauppi *et al.*, 2004). It is tempting to suggest that duplications and some other rearrangements involving *PLP1* could result from aberrant joining of meiotic DSBs within the X chromosome by NHEJ. The documented higher frequency of *PLP1* duplications in spermatogenesis could be explained by this proposed mechanism, as the unpaired regions of the X chromosome may be more susceptible to rearrangements of this type, which could occur between sister chromatids (Mimault *et al.*, 1999). However, it has been shown that proteins essential for the major pathway of NHEJ are downregulated in early meiosis in male mice, suppressing the error-prone NHEJ pathway and ensuring

accurate repair of meiotic DSBs by homologous recombination (Goedecke *et al.*, 1999). Recombination rates throughout the genome can be assessed by creating genetic maps based on genotyping large numbers of polymorphic markers in numerous families. A recent high resolution recombination map of the human genome based on over genotyping 5000 polymorphic markers in 1257 human meioses is available as an annotated track on the UCSC genome browser (Kong *et al.*, 2002). The average recombination rate on the X chromosome calculated from this genetic map is 1.19 cM/Mb (excluding the centromere), but the female recombination rate close to *PLP1* is substantially less than this (the male recombination rate on the X chromosome, except for within pseudoautosomal regions, is zero) (Kong *et al.*, 2002). In a 1Mb region including *PLP1* (between 101-102Mb on the X chromosome, NCBI Build 34), the average recombination rate was 0.3 cM/Mb (UCSC genome browser) (Kong *et al.*, 2002). The two flanking 1Mb regions had even lower recombination rates, the proximal 1Mb region had a recombination rate of 0.2cM/Mb and the next most distal 1Mb region had a recombination rate of zero (UCSC genome browser) (Kong *et al.*, 2002). Although still a theoretical possibility, it seems unlikely that DSBs during meiosis I are responsible for the rearrangements seen in PMD duplication families, as the homologous recombination pathway appears not to be involved in most PMD duplications, and additionally there is no evidence for a meiotic recombination hotspot in the region, which could be a source of DSBs.

4.12.5. DSBs, strand invasion and replication

There are other alternative mechanisms for the generation of *PLP1* duplications. DNA rearrangements that are passed on through the germ line do not necessarily have to take place during meiosis, as cells in the germ line undergo numerous meiotic divisions. The higher frequency of duplications originating in the male germ line may just be a function of the greater number of cell divisions in spermatogenesis as

compared to oogenesis. Probably the least complicated mechanism for producing a tandem duplication such as seen in family 1, and several other PMD families, just involves misrepair, by NHEJ, of two DSBs (Woodward *et al*, in preparation). If two DSBs occur on sister chromatids, one located proximal to *PLP1*, and the other distal to *PLP1*, these could be mis-repaired, distal break to proximal break by aberrant NHEJ, resulting in a duplication of the sequence between the two breaks. The DSBs would probably have to occur almost simultaneously during S or G₂ phases of the cell cycle for two free ends from each DSB to be ligated together by NHEJ. This is not an unlikely prospect, as it has been shown that DSBs do occur frequently in mammalian cells (Karanjawala *et al.*, 1999). The simplicity of this model is appealing, although it does have some drawbacks. One problem is how two of the free ends created by DSBs are brought together prior to repair. It could be the case that the two ends are randomly associated closely together in the nucleus, or that the two genomic regions near the breakpoints were located in a particular domain within the nucleus prior to the production of the DSBs. However, the more accurate repair pathway of homologous recombination (HR), using the intact sister chromatid to repair damage, is a more frequently used pathway for DNA repair during late S/G₂ phases (Takata *et al.*, 1998). To generate a tandem duplication, typically two copies of the duplicated sequence need to be present, which could either originate from a homologous chromosome pair, or from a pair of sister chromatids. PMD duplications are generally intrachromosomal in origin and frequently originate in males, which strongly suggests that the rearrangement occurs during S/G₂ phases when two sister chromatids from one X chromosome are present. The high likelihood of the rearrangement occurring during these later phases of the cell cycle, while HR is commonly utilised for DSB repair, suggests that HR could also have a role in *PLP1* duplications.

One possible duplication mechanism invokes HR pathways, but also still uses NHEJ in the repair process, which is important for *PLP1* duplications as most breakpoints do show characteristics of NHEJ repair (Woodward *et al.*, in preparation). Break-induced replication is a phenomenon mostly studied in yeast models which is initiated by a single DSB (Morrow *et al.*, 1997; Kraus *et al.*, 2001). Following the generation of a DSB in a DNA molecule, just one of the free ends created by the break then invades into homologous sequence on another chromosome or chromatid and then replication is primed from the invading free end (Kraus *et al.*, 2001). A similar process based on recombination has been shown to be involved in restarting stalled and damaged replication forks in *E.coli* (reviewed in Michel, 2000). The single replication fork initiated from an invading end can then continue for a considerable distance, even up to the end of the recipient chromosome (Morrow *et al.*, 1997). Alternatively, the break-induced replication may only continue for a limited distance before being terminated by a non-homologous rejoining of the end of the newly replicated region to the other side of the initial DSB (Kraus *et al.*, 2001). This process has also been reported to occur in mammalian cells where it may lead to non-reciprocal gene conversion or duplication of sequences (Richardson and Jasin, 2000; Johnson and Jasin, 2000). *PLP1* duplications could be produced by the following mechanism: A DSB occurs in the *PLP1* region, and a free DNA end 5' to the DSB then invades the homologous sequence on the sister chromatid (Figure 4.16.). This invading end primes DNA synthesis, which then proceeds along the chromosome, past the *PLP1* gene (Figure 4.16.). At some point at the other side of *PLP1*, the replication fork is stalled, or encounters an obstacle that stops replication and causes the newly synthesised strand to dissociate from the template sister chromatid. As replication factories are generally believed to be stationary in the nucleus, it is likely that the progressing replication fork initially induced by one invading end of the DSB will still be in the vicinity of the other end of the DSB, which will then be available as a

substrate for NHEJ with the end of the replicated region (Figure 4.16.) (Cook, 1999; Richardson and Jasin, 2000; Johnson and Jasin, 2000). There is some clustering of distal breakpoints in PMD duplications, with over 50% of duplication breakpoints, including family 1, mapping to the LCRs distal to *PLP1* (Figure 3.4.) (Woodward *et al*, in preparation). The presence of the distal LCRs may cause secondary structure in the DNA, which could lead to replication termination (if the initial DSB had been proximal to *PLP1*). Alternatively sequences within these LCRs could be prone to DSB formation, which initiates a strand invasion event proximal to *PLP1*, with subsequent replication through *PLP1* until termination of replication proximal to the gene.

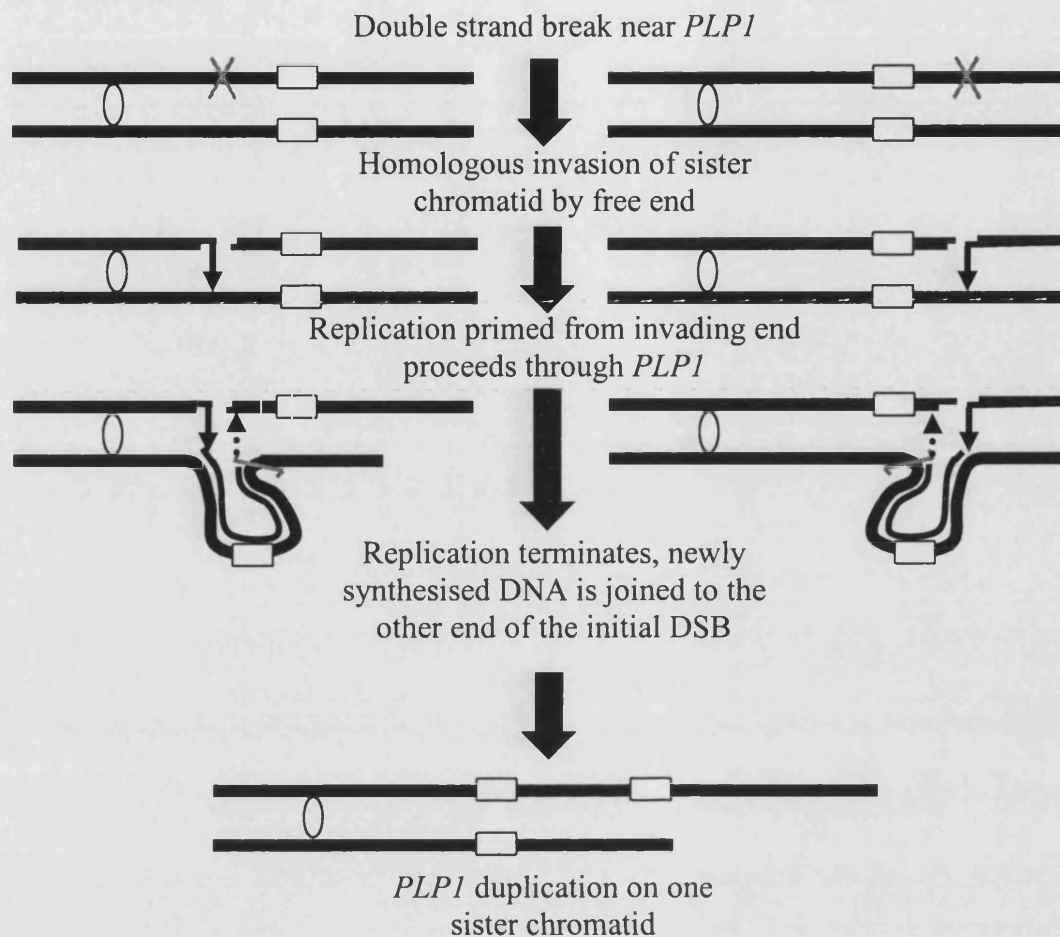


Figure 4.16. Diagram showing possible mechanism for *PLP1* tandem duplications. X chromosome sister chromatids are shown schematically by the black horizontal lines, the X centromere is a white oval, and *PLP1* is the white rectangle. Homologous strand invasion events are shown by solid arrows and non-homologous repair events are indicated by the dotted arrows. Newly synthesised DNA is shown as a thinner line. A DSB-causing event is shown by the grey cross and replication termination is shown by the grey zigzag.

4.12.6. Interspersed repetitive elements

Both the proximal and distal duplication breakpoints in family 1 were within interspersed repetitive sequence elements (see section 4.5.). Several other sequenced *PLP1* tandem duplication breakpoints have been found to map within interspersed repetitive elements, more than would be expected by chance, given the repeat content of the region (Woodward *et al.*, in preparation). It is possible that these common repeat elements may stimulate rearrangements, perhaps by being prone to DSBs. The proximal duplication breakpoint in family 1 is within an L1 retrotransposon sequence (L1PA13) and the distal duplication breakpoint is within a solo LTR element (MLT1H1) (Tables 4.3. and 4.4.). Interspersed repeat elements, including L1 and LTR sequences have been found to be over-represented at sites of DNA rearrangement in other studies, suggesting that these sequences could be contributing to the mechanism behind these rearrangements (Toriello *et al.*, 1996;Graw *et al.*, 2000;Rockwood *et al.*, 2004). Solo LTR elements have been found in some documented mammalian recombination hotspots, and it has been suggested that instead of creating DSBs, solo LTRs may be captured during the process of DSB repair, as has been found in some model systems (Shiroishi *et al.*, 1995;Lin and Waldman, 2001a). A similar mechanism of L1 sequences being captured by or inserting into DNA lesions has also been proposed as a repair mechanism for DNA DSBs (Morrish *et al.*, 2002;Rockwood *et al.*, 2004). It is possible that the presence of some interspersed repeats may be a marker of past repair of a DSB by insertion of a retrotransposed sequence, so the repeat elements may not in themselves cause further rearrangements, but could be indicative of a predisposition towards DSBs, rearrangement or recombination in the region (Lin and Waldman, 2001a).

4.12.7. Nucleotide content of breakpoints

Both duplication breakpoints had an elevated G+C content compared to the surrounding genomic region (see section 4.5.). Regions of high G+C content are more likely to be involved in translocations and segmental duplications, and have been found to have higher recombination rates, which may be due to recombination-promoting sequences such as *Alus* clustering in G+C rich regions (Eisenbarth *et al.*, 2000; Yu *et al.*, 2001; Abeysinghe *et al.*, 2003; Jurka *et al.*, 2004). The sequences around the breakpoints were examined for repeated sequences, polypurine, polypyrimidine and alternating purine/pyrimidine tracts, but only a limited number were found, so it was considered that these sequence features were unlikely to be important in the rearrangement mechanism (see Figures 4.14. and 4.15.).

4.13. Summary

During the course of this study a tandem duplication including the *PLP1* gene has been characterised in a family with PMD. The duplication breakpoint has been sequenced with little homology evident between the proximal and distal ends of the duplication, implicating NHEJ in the rearrangement mechanism. Various different mechanisms for the rearrangement have been proposed, and it is also possible that a form of homologous recombination is also involved in the generation of duplications such as this one characterised in family 1.

5.1. DUPLICATION BREAKPOINT MAPPING IN FAMILY 2

5.1.1. Previously published data

Patient 2:9 and his mother 2:5 have previously been shown to carry a duplication of Xq22.2 including *PLP1*, and the approximate extent of the duplication has also been mapped by interphase FISH (Figure 2.2.) (Family PMD3 in Woodward *et al.* (2000)). The size of the duplication in family 2 had been found to be approximately the same as in family 1, with cosmid cU65A4 duplicated at the proximal end and cosmid cU240C2 the most distal duplicated clone (see section 4.1.1. and Figure 4.1.) (Woodward *et al.*, 2000).

5.2. FISH mapping proximal duplication breakpoint in family 2

5.2.1. Further mapping of proximal duplication breakpoint by interphase FISH

Additional interphase FISH experiments were carried out to further localise the proximal duplication endpoint. Adjacent genomic clones including cU65A4 (previously found to be duplicated) from near the end of the duplicated region were hybridised to interphase nuclei from individual 2:9 (Figure 5.1. and Table 5.1.). These data indicated that the duplication breakpoint was likely to be contained within the cosmid cU177E8, just distal to cU65A4 (Figure 5.1. and Table 5.1.). In contrast to the previously published results, the interphase FISH score for cU65A4 was consistent with this clone not being contained within the duplicated region (Table 5.1.). However, in the nuclei where there was a double cU65A4 signal, one signal was frequently smaller than the other, which could be an indication of a very small part of cU65A4 being included in the duplicated region. The clone just centromeric to cU65A4, dJ635G19, also was scored as not being duplicated, whereas the cosmid mapping just distal to cU65A4, cU177E8, could not be classified as either duplicated or single-copy (Table 5.1. and Figure 5.1.). In addition, when a double signal was

seen, one of the dots was smaller than the other (Figure 5.1.). These data from cosmid cU177E8 were consistent with a duplication breakpoint being present within the genomic sequence in this clone.

5.3. Mapping of proximal duplication breakpoint by UPQFM-PCR

As some of the interphase FISH results were inconsistent, another method was used to assay dosage in the presumed proximal duplication breakpoint region. Finer mapping of the extent of the duplication was carried out using UPQFM primer pairs in the region (Figure 5.2.). UPQFM-PCR data narrowed down the region to just over 2Kb, between primer pairs 38464-38658 and 36491-36712 in human genomic clone cU177E8 (Figure 5.2. and Table 5.2.). However, there was one apparently discordant result found, but this was only for one individual compared against one control probe, and it was considered likely to be a false positive result (Table 5.2.).

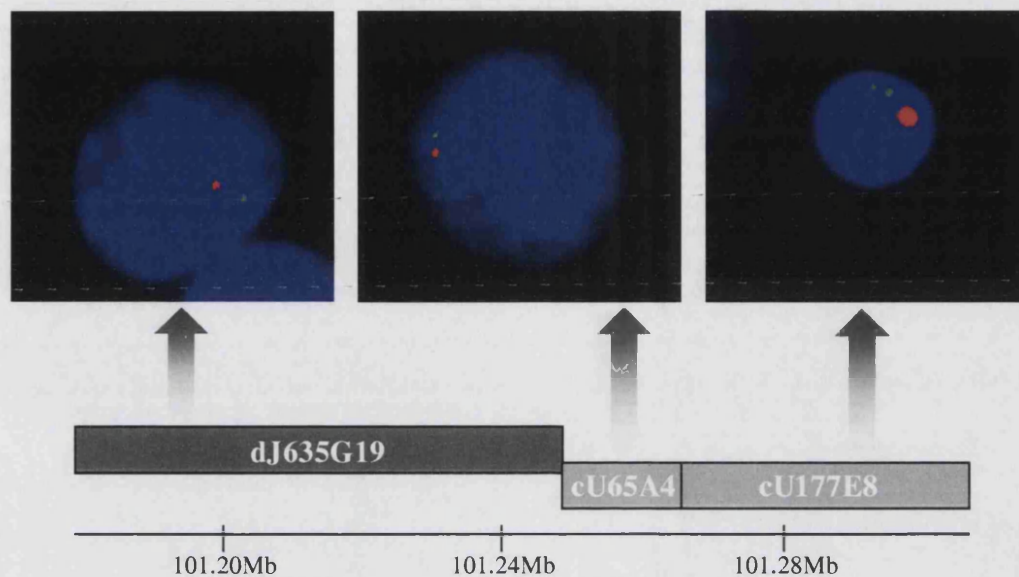
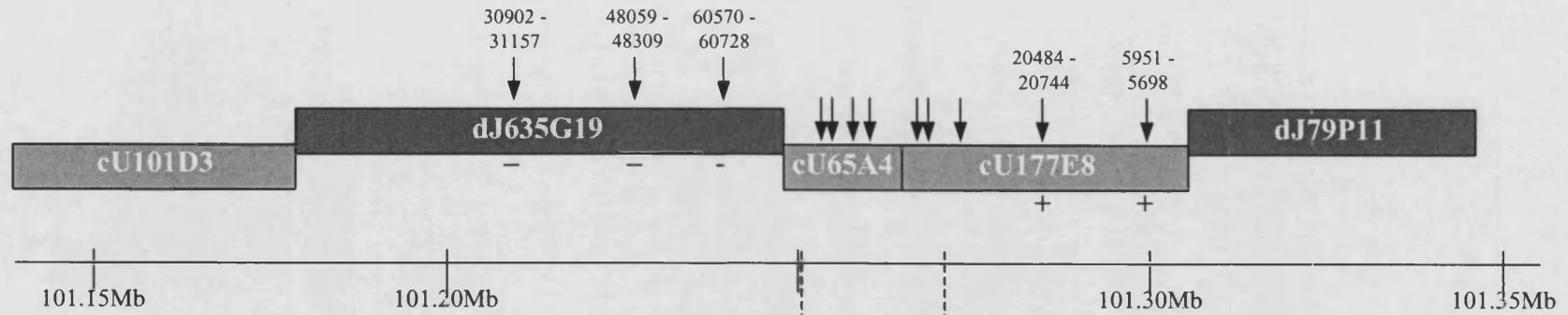


Figure 5.1. Interphase FISH results from the proximal duplication breakpoint using a cell line from the affected boy in family 2. Genomic clones around the proximal breakpoint region and their relative size are shown underneath the images of interphase nuclei. The position of the clones on the X chromosome is shown, taken from the NCBI Build 34 version of the human genome sequence. In each image, the X centromere is labelled red, and the genomic clone used in the experiment is labelled green.

Clone	0,1	0,2	0,3	0,4	Other	Conclusions
dJ635G19						Not duplicated
cU65A4	74.76%	20.39%	4.85%	0	0	Not duplicated
cU177E8	68.47%	26.13%	0.91%	0	4.50%	Not duplicated
cU177E8	45.28%	44.34%	6.60%	0	3.77%	Breakpoint?

Table 5.1. Interphase FISH scores for 2:9 near the proximal duplication breakpoint. The percentage of nuclei falling into each category is shown for each clone. Results are taken from a single slide for each probe, and an average of 107 interphase nuclei were scored for each slide.

a.



b.

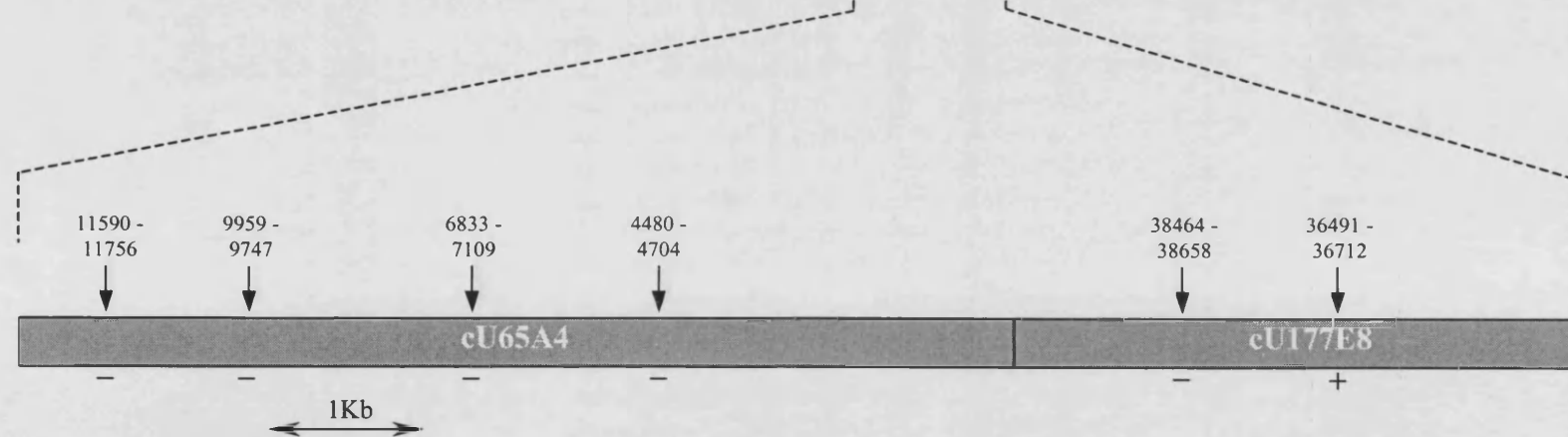


Figure 5.2. UPQFM-PCR mapping of proximal breakpoint in family 2. Genomic clones from the region are shown as shaded boxes labelled with the clone name, those shaded dark grey are orientated centromere → telomere, those in light grey have been sequenced on the opposite strand. Locations of UPQFM-PCR primers in the region are indicated by arrows, and the position within each clone for each pair of primers is shown. (+) indicates a sequence that appeared to be duplicated, (-) a sequence with normal copy number. Position on the X chromosome is shown by the scale bar in (a.) and is taken from the Ensembl human genome browser, release 21.34d.1. A more detailed view of 20Kb around the duplication breakpoint is shown in (b.).

Clone	Position in clone of UPQFM primer pairs	Mean ratio of UPQFM PCR product compared to...		Number of experiments performed
		PLP1	CF	
dJ635G19	30902-31157	0.67	0.95	1
dJ635G19	48059-48309	0.56 (0.77)	0.89 (0.86)	2 (1)
<i>dJ635G19</i>	<i>60570-60728</i>	<i>0.54 (0.28)</i>	<i>1.21 (1.05)</i>	<i>4 (1)</i>
cU65A4	11590-11756	0.47	1.18	1
cU65A4	9959-9747	0.51	0.84	2
cU65A4	6833-7109	0.44	0.98	2
cU65A4	4480-4704	0.54	0.84	2
cU177E8	38464-38658	0.43 (0.75)	1.09 (1.47)	3 (3)
cU177E8	36491-36712	0.94 (0.98)	2.31 (2.07)	3 (3)
cU177E8	32435-32693	0.92 (1.46)	2.74 (1.67)	2 (1)
cU177E8	20484-20744	1.11	2.47	2
cU177E8	5951-5698	1.03	1.63	2

Table 5.2. UPQFM ratios for the proximal duplication breakpoint in family 2 (as compared against both *PLP1* exon 6 and CF control primers). Average ratios were taken from all experiments carried out that included that primer pair, and the number of experiments is indicated. Values are for dosage measurements from the affected male, where experiments had been carried out on the carrier mother as well, these results are shown in brackets. Ratios consistent with duplication are highlighted in bold, and the zigzag lines show the assumed location of the breakpoint regions. The position of each target sequence within the relevant genomic clone is shown. Data obtained from UPQFM-PCR experiments carried out by Dr Karen Woodward are shown in italics.

5.4. Inverse PCR to obtain breakpoint sequences

Initial attempts to clone the breakpoint in family 2 using long-range PCR were unsuccessful (data not shown). A tandemly duplicated rearrangement (similar to that found in family 1) was assumed but not demonstrated. Therefore an alternative approach was taken, in case there was a more complicated rearrangement present. A strategy using inverse PCR was pursued to obtain unknown sequence from the duplication breakpoint in this family, utilising DNA sequence known at the well mapped proximal breakpoint region (see sections 5.2. and 5.3.). Inverse PCR is a technique that can be used to generate sequence from an unknown region using primers specific for adjacent known sequence orientated in divergent directions in genomic DNA (see section 2.2.1.4.) (Ochman *et al.*, 1988).

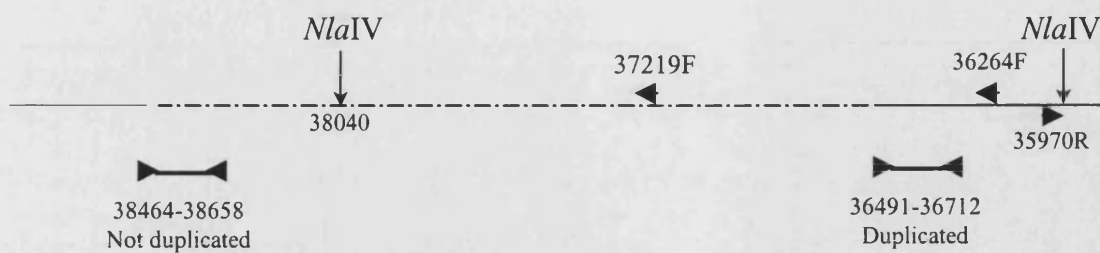
5.4.1. Inverse PCR strategy

Two 30mer oligonucleotide primers were used close to the end of the duplicated sequence within clone cU177E8 (Figure 5.3.). A restriction enzyme was chosen, in this case *Nla*IV, that would cut close to the reverse primer sequence, but as far away from the other primer as possible, to maximise the chances of amplifying the breakpoint sequence following circularisation. The *Nla*IV site that was in the known duplicated sequence for this experiment was just a couple of nucleotides after the end of the reverse primer, at position 35938 within cU177E8 (Figure 5.3a.). Genomic DNA that had been fully digested by *Nla*IV was ligated at a low concentration to promote self-ligation and circularisation of individual DNA fragments. This ligated DNA was used as the substrate for PCR using the specially designed oligonucleotide primers (Figure 5.3a.).

5.4.2. Inverse PCR results

Following the inverse PCR procedure, the size of the PCR product that was expected from normal genomic DNA was 1785bp, and a band of approximately this size was seen in both normal male controls and members of family 2 carrying the duplication (Figure 5.3b.). An additional band, slightly shorter than the expected band, was seen for just the affected boy and his carrier mother, and not in normal genomic DNA (Figure 5.3b.). This was assumed to be amplification of a novel junction fragment formed by the duplication event. This PCR product was gel purified and sequenced in both directions using the inverse PCR primers (see sections 2.2.1.4.4. and 2.2.4.).

a.



b.

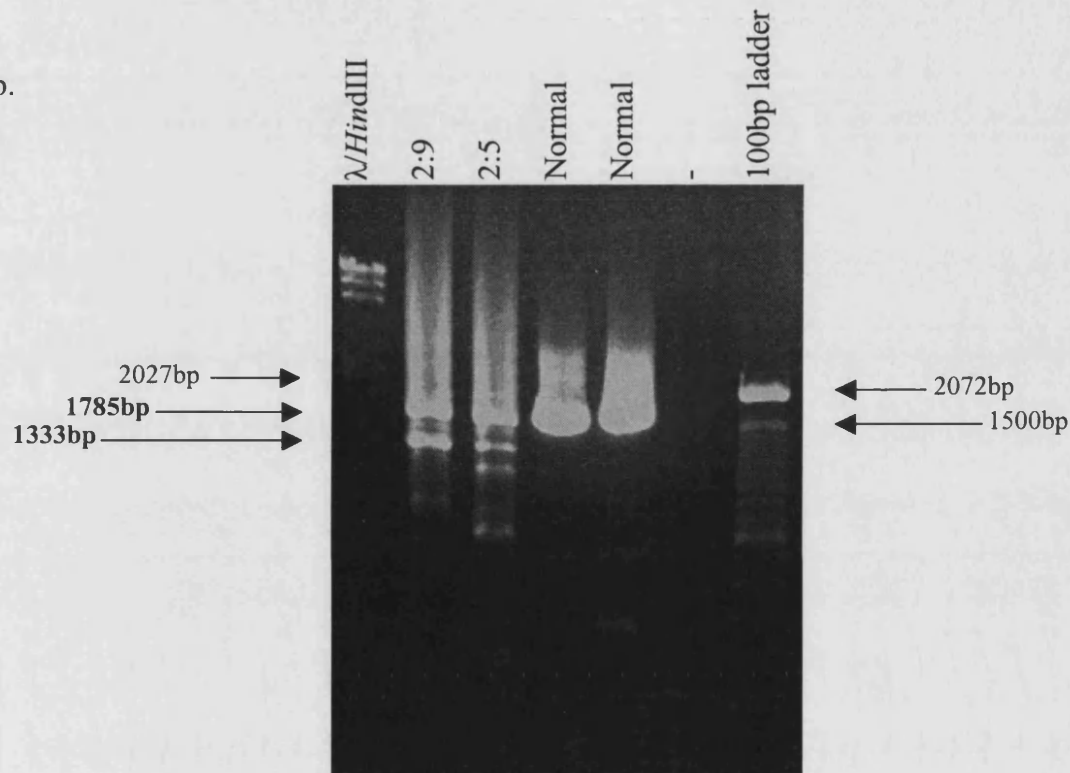


Figure 5.3. Inverse PCR strategy to clone the proximal duplication breakpoint for family 2. (a.) shows the approximate position of primers and *Nla*IV cut sites used for inverse PCR (not to scale). Positions within cU177E8 sequence are shown, and the region which the proximal duplication breakpoint has been mapped by UPQFM-PCR is shown by the dashed line. (b.) shows inverse PCR products run on an agarose gel along with λ *Hind*III and 100bp ladder size standards. The sizes of the expected band (1785bp) and the altered band (1333bp) are shown in bold, whereas the sizes of nearby bands in the two size standards are shown in normal type.

5.4.3. Sequencing of inverse PCR products.

Sequencing of the additional inverse PCR products in the patient and his mother revealed a breakpoint junction, that surprisingly only involved sequences proximal to *PLP1* (Figures 5.3., 5.4. and 5.6.). Sequencing from the additional inverse PCR band using the 35970R primer and a nested primer, 36264F, showed that the breakpoint within cU177E8 was at position 37339 (Figures 5.4. and 5.5.). There was a 4 base pair overlap (AGAG) at the breakpoint and then the adjacent sequence was a 49bp stretch of DNA, also originating from cU177E8, between positions 36006 and 36054 (Figure 5.4.). The following sequence in the junction originated from a region more proximal to *PLP1*, within human genomic clone dJ1055C14 (Figures 5.4. and 5.5.). The sequence began at position 59809 in dJ1055C14, where there was a two base pair (TT) overlap with sequence from cU177E8 at position 36053 (Figure 5.4.). 22 nucleotides into the dJ1055C14 sequence there was an additional small rearrangement in the sequence, a 13bp deletion within dJ1055C14 sequence (Figure 5.4.). There was a 3bp overlap (ATG) flanking this short deletion (Figure 5.4.). To verify that these rearrangements were not just an artefact generated by the inverse PCR procedure, a junction fragment was amplified from DNA from patient 2:9, using primers flanking the sequenced junction (cU177E8 F37219 and dJ1055C14 F60006). DNA extracted from lymphoblastoid cell lines was used, as a DNA sample from blood was not available. Sequencing of this PCR product gave the same sequence in both directions as that obtained from the inverse PCR product (Figure 5.4.).

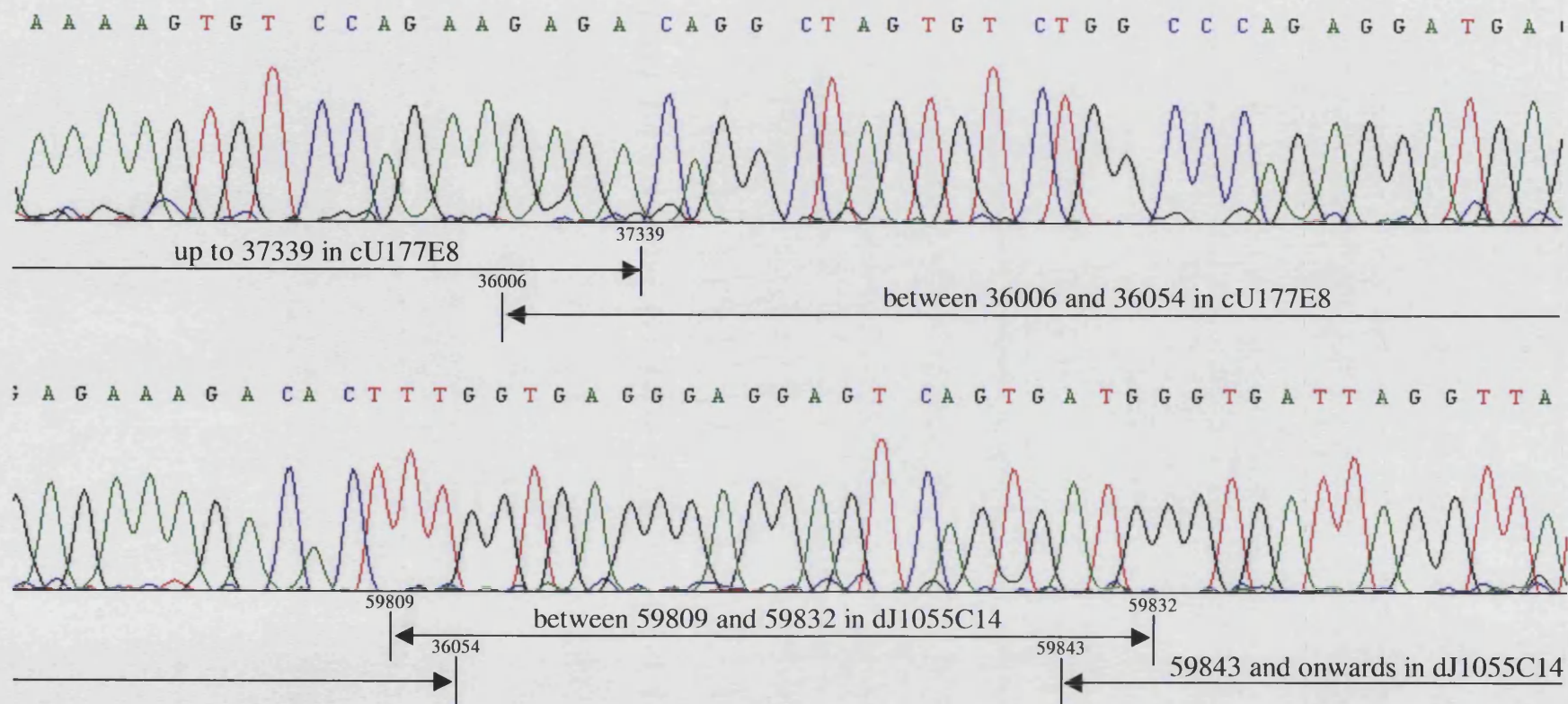


Figure 5.4. Electropherogram showing sequence from the cU177E8-dJ1055C14 breakpoint junction in family 2. The regions of homology to the different Xq22 sequences are shown underneath the sequence data, numbers indicate positions within the sequence of each genomic clone.

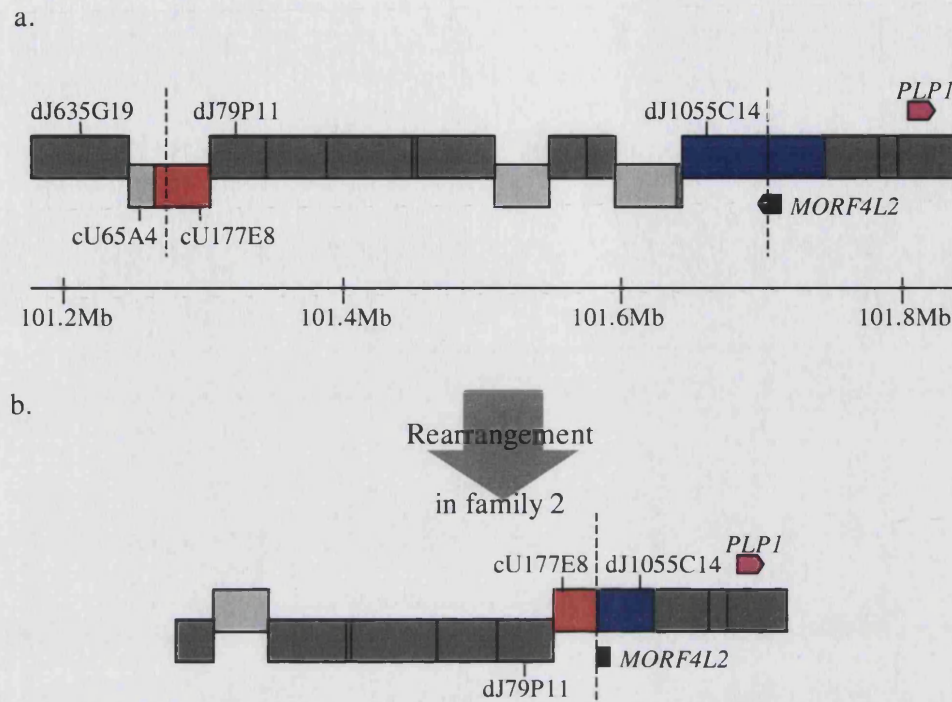


Figure 5.5. Diagram illustrating the location of the breakpoints sequenced from the family 2 inverse PCR reaction and a large-scale view of the rearrangement found. (a.) shows the sequenced contig (adapted from the Ensembl genome browser) proximal to and including the *PLP1* gene and the locations of the sequence found at the breakpoint. The two genomic clones involved in the rearrangement, cU177E8 and dJ1055C14, are shaded in red and blue respectively. The other genomic clones in the region are shaded in dark grey (for clones sequenced on the forward strand) or light grey (clones sequenced on the reverse strand). The *PLP1* gene (shaded pink), and a gene that spans the breakpoint within dJ1055C14, *MORF4L2* (shaded black), are shown as arrows, with the orientation of the arrow indicating the direction of transcription. The scale bar shows the position along the X chromosome in megabase pairs (taken from the Ensembl genome browser, release 22.34d.1). The dashed lines show the locations of the breakpoints. (b.) illustrates the rearrangement that has occurred at this breakpoint in this family, with sequences from cU177E8 and dJ1055C14 brought together, resulting in inversion of one of the sequences relative to the original (just one orientation is shown in (b.), but the opposite orientation is also possible).

5.5. Verification of cU177E8-1055C14 breakpoint by long-range PCR

The junction found using inverse PCR, which involved sequences from human genomic clones cU177E8 and dJ1055C14, was also checked by PCR. Long-range PCR was carried out using a single primer close to the breakpoint in cU177E8 (37219F, see Figure 5.3a) and different oligonucleotide primers mapping between 3.2 and 9Kb from the start of the sequence homologous to dJ1055C14 (Figure 5.6.). There was no indication of further large-scale rearrangements at this breakpoint for at least 9Kb into dJ1055C14 sequence from the breakpoint.

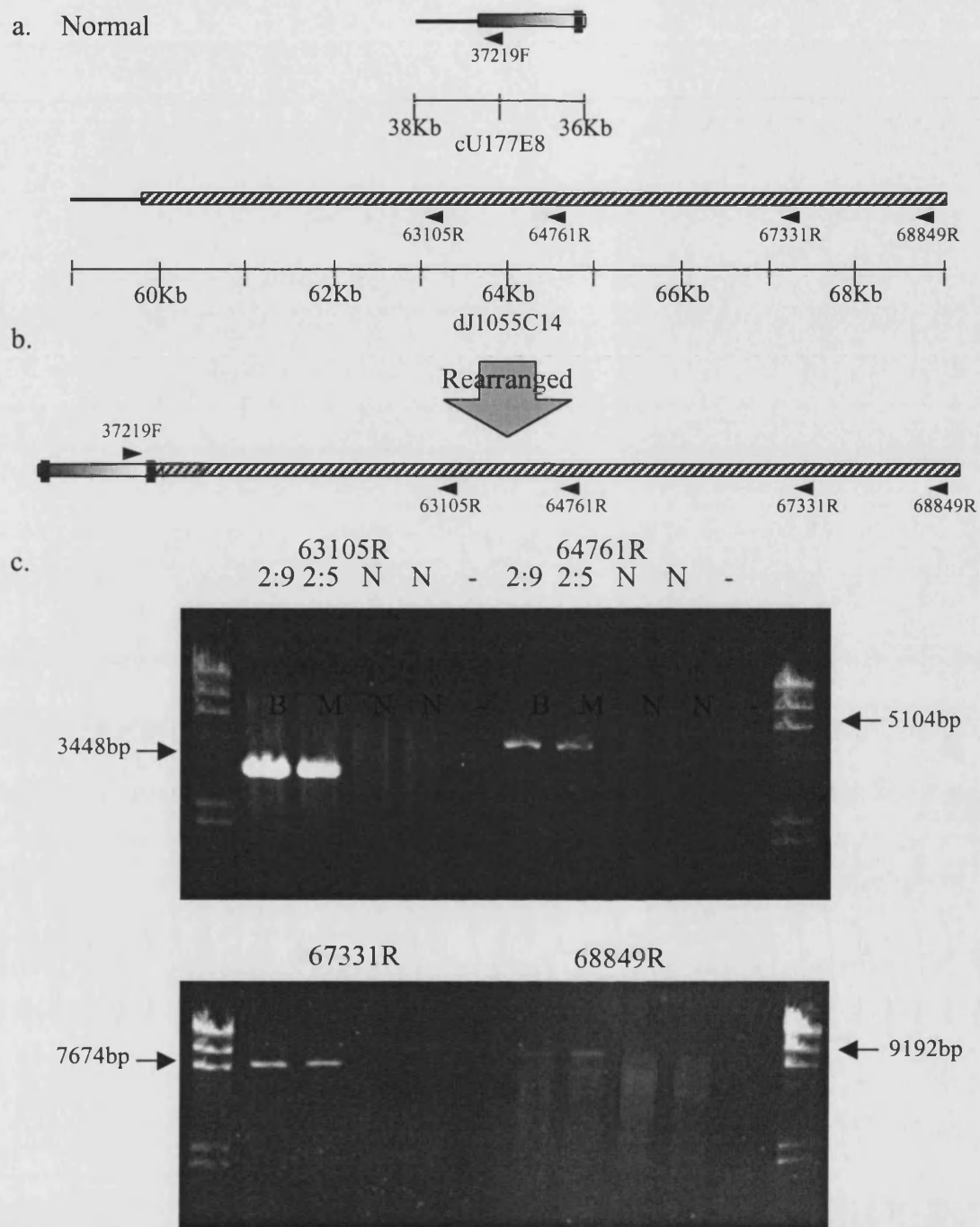


Figure 5.6. Confirmation of cU177E8-dJ1055C14 breakpoint by long-range PCR.

Legend on next page

Figure 5.6. Legend

(a.) Diagrams of normal sequence within both cU177E8 (2Kb shown) and dJ1055C14 (10Kb shown). DNA not involved in the breakpoint is shown as a black line, the duplicated sequence from cU177E8 is represented by a shaded box, and the 49bp sequence from cU177E8 inserted at the breakpoint is shaded black. Sequence duplicated originating from dJ1055C14 is shown as a striped box. The orientation of primers used is shown by the arrowheads, and their respective positions within each genomic clone is shown. (b.) Diagram of the rearrangements around the breakpoint, as found by direct sequencing. (c.) Agarose gels showing long-range PCR products, each reaction using the same primer from cU177E8 (37219F) and a different reverse primer from dJ1055C14, which is indicated above the gel picture. The first and last lane in each gel contains a λ /HindIII ladder, and DNA used in the reactions was either cell line DNA from the affected boy (2:9), his mother (2:5), genomic DNA from normal males (N), and a no DNA control (-).

5.6. Determination of sequence copy number around dJ1055C14 in family 2

5.6.1. UPQFM-PCR

Following the unexpected discovery of a breakpoint junction between two relatively distant areas of sequence, but both proximal to *PLP1*, the copy number of sequences near to the breakpoint within dJ1055C14 was examined. This had not been carried out previously as the proximal breakpoint had initially been mapped centromeric to this region, and the rearrangement in this family had been assumed to be a simple tandem duplication including *PLP1* (Woodward *et al.*, 2000). UPQFM-PCR primer pairs throughout genomic clone dJ1055C14 were used to assay copy number. This revealed that sequences distal to the breakpoint within dJ1055C14 at position 59809

were duplicated in both the boy and mother, whereas target sequences proximal to this breakpoint, further away from *PLP1*, were just present in normal copy number in both the affected boy and carrier mother (Figure 5.7., Table 5.4.). More UPQFM-PCR primer pairs, mapping to genomic clones located proximal to dJ1055C14, were then used to ascertain copy number (Figure 5.7., Table 5.4.). This showed that the sequence copy number was normal for approximately 100Kb, between two non-contiguous sections of duplicated sequence, one from cU177E8 down to the most proximal part of cV857G6 sequence, the other starting within dJ1055C14 and possibly continuing to *PLP1* and beyond (Figure 5.7.).





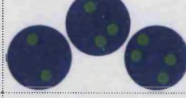
5.6.2. Interphase FISH around dJ1055C14

Interphase FISH results for the genomic clones mapping to this region were supportive of the UPQFM-PCR data (Figure 5.7., Tables 5.3a,b). The FISH results for the clone containing the sequenced breakpoint, dJ1055C14, were consistent with the breakpoint being within this clone (Tables 5.3a,b). Although more nuclei were scored as duplicated than not, the difference between the two classifications was not great, and in many cases where a duplicated signal was seen, the additional signal was smaller than the other, which is also an indication that the clone is only partially duplicated (Figure 5.7.). The breakpoint is positioned within the sequenced portion of dJ1055C14 so that less than half of the clone is duplicated, but this equates to about 40Kb of sequence, which is enough to give a good signal in interphase nuclei.

UPQFM-PCR data indicated that only a small part of cV857G6 sequence was duplicated, and all of the sequence from the next clone upstream, cU246D9, was probably duplicated (Figure 5.7., Table 5.4.). Over half of nuclei in both 2:9 and 2:5 only contained signals that were consistent with one copy of cV857G6 being present

on each X chromosome and about a quarter of nuclei had duplicated signals (Tables 5.3a,b). In these nuclei where duplicated signals were seen, the second signal was generally much smaller than the other signal in the pair (Figure 5.7.). Interphase FISH data from cU246D9 had fairly similar numbers of single-copy and duplicated signals, although the scores from the affected boy and carrier mother for this clone were tending towards different outcomes (2:9 – more nuclei scored as not duplicated, 2:5 – more nuclei scored as duplicated) (Tables 5.3a,b). This did not agree with UPQFM-PCR data, which had indicated that the whole of the sequence from cU246D9 was duplicated (Table 5.4.). However, the sequence submitted for each genomic clone is often not the whole cloned insert, and data based on restriction digestion analysis of genomic clones shows that cV857G6 and cU246D9 do overlap, and about half of the length of cU246D9 overlaps with cV857G6 (Figure 5.8.). It is quite likely that only part of the genomic insert cloned in cU246D9 is actually duplicated in family 2, which would explain why the interphase FISH results did not show that this clone was duplicated. This discrepancy between FISH results may be due to poor hybridisation in the female sample, with a higher proportion of nuclei interpreted as having a signal from just one X chromosome ((0,1), (0,2), and (0,3)), which could be attributed to poor hybridisation to one of the X chromosomes on these slides (Table 5.3.).

a.

Clone	0,1	0,2	0,3	0,4	Other	Conclusions
						
cU246D9	51.73%	42.95%	3.87%	0	1.45%	Breakpoint?
cV857G6	74.26%	24.04%	0.99%	0	0	Not duplicated
dJ1055C14	45.00%	45.00%	10.00%	0	0	Breakpoint?

b.








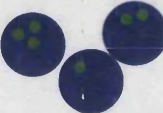
Clone	1,1	1,2	1,3	1,4	2,2	2,3	2,4	Other	Conclusions
									
cU246D9	25.96%	41.35%	7.69%	0.96%	6.73%	2.88%	0	13.46%	Duplicated?
cV857G6	49.11%	21.43%	0.89%	0	2.68%	0	0	25.89%	Not duplicated
dJ1055C14	27.78%	50.00%	2.78%	0	6.48%	2.78%	0	10.19%	Breakpoint?

Table 5.3. Interphase FISH scores, expressed as percentages, from family 2 near to the second proximal breakpoint. (a.) shows the interphase FISH scores from the male, 2:9, and (b.) gives the interphase FISH scores from the carrier female (2:5). All scores are from a single slide, except for clone cU246D9 on individual 2:9 (a.), which is the mean percentage score from two experiments. An average of 105 nuclei were scored on each slide.

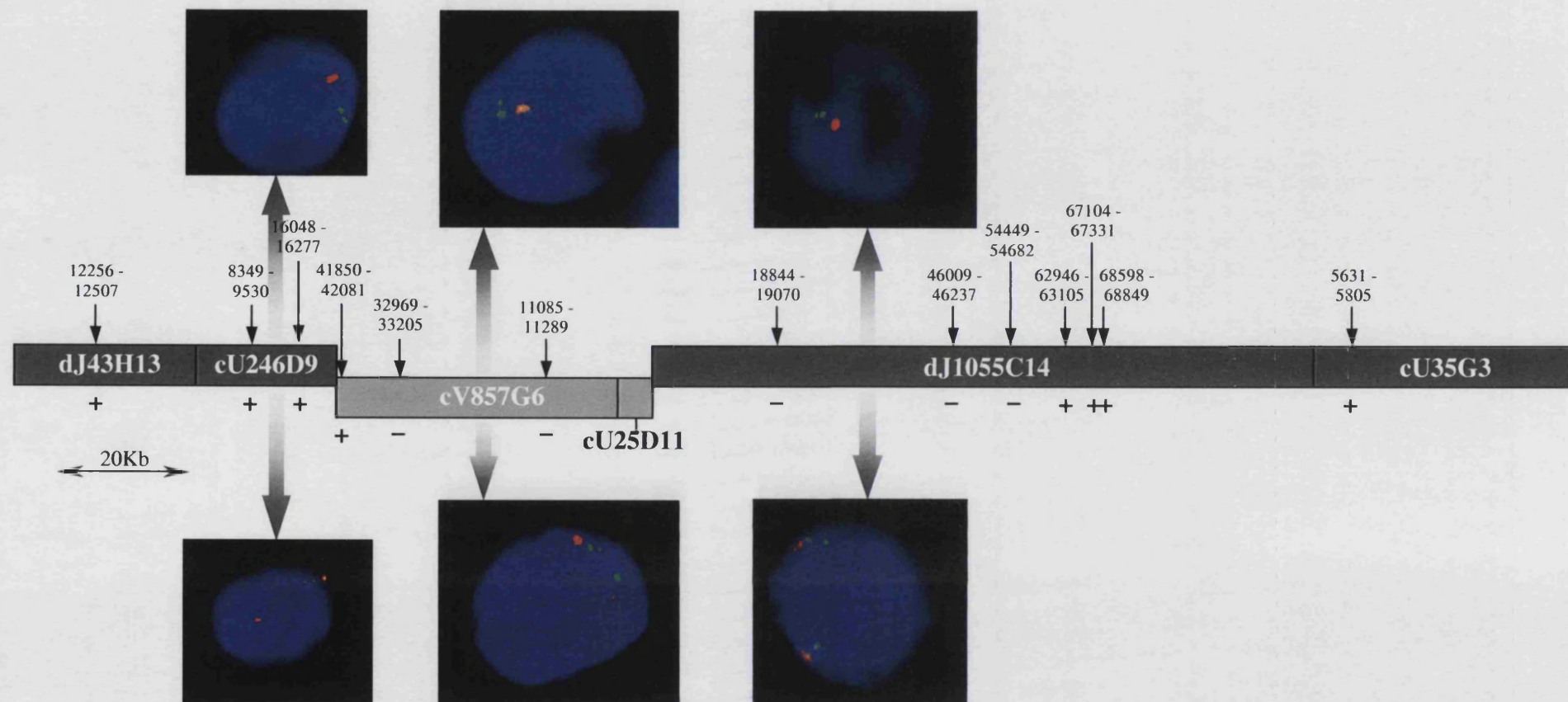


Figure 5.7. UPQFM-PCR and interphase FISH data mapping two breakpoints within dJ1055C14 and cV857G6. Legend on next page.

Figure 5.7. Legend

The locations of the human genomic clones in the region are shown by the grey boxes in the centre of the diagram; clones on the forward strand (oriented left to right) are coloured in dark grey, those oriented in the opposite direction are shaded light grey. The black arrows show locations of UPQFM-PCR primer pairs within the genomic clones, (+) underneath the contig indicates that sequence was duplicated, (-) that the sequence is single-copy. Interphase FISH nuclei are shown for three of the clones in the region. In all pictures the X centromere probe is red, and the Xq22.2 genomic clone is green. Nuclei from the affected male in family 2 are shown above the contig and UPQFM-PCR data, the pictures underneath are from the mother in family 2.

Clone	Position in clone of UPQFM primer pairs	Mean ratio of UPQFM PCR product compared to...		Number of experiments performed
		PLP1	CF	
dJ43H13	12256-12507	1.01 (0.92)	2.54 (1.51)	2 (2)
cU246D9	8349-9530	0.89 (1.04)	2.41 (1.53)	2 (2)
cU246D9	16048-16277	0.96 (0.94)	-	1 (1)
cV857G6	41850-42081	0.85 (1.12)	2.01 (2.01)	2 (3)
cV857G6	32969-33205	0.32 (0.47)	0.83 (0.85)	1 (2)
cV857G6	11085-11289	0.52 (0.54)	1.05 (0.95)	4 (2)
dJ1055C14	18844-19070	0.51 (0.54)	1.16 (0.95)	3 (2)
dJ1055C14	46009-46237	0.56 (0.56)	1.37 (1.02)	3 (3)
dJ1055C14	54449-54682	0.53 (0.63)	1.21 (0.97)	1 (1)
dJ1055C14	62946-63105	0.78 (0.95)	1.78 (1.47)	1 (1)
dJ1055C14	67104-67331	1.04 (1.06)	2.37 (1.64)	1 (1)
dJ1055C14	68598-68849	0.95 (0.90)	2.17 (1.40)	1 (1)
cU35G3	5631-5805	1.55	2.80	1

Table 5.4. UPQFM ratios for the proximal duplication breakpoint in family 2 (as compared against both *PLP1* exon 6 and *CF* control primers). Average ratios were taken from all experiments carried out that included that primer pair, and the number of experiments is indicated. Values are for dosage measurements from the affected male, where experiments had been carried out on the carrier mother as well, these results are shown in brackets. Ratios consistent with duplication are highlighted in bold, and the zigzag lines show the assumed location of the breakpoint regions. The position of each target sequence within the relevant genomic clone is shown.

Detailed view: X:99.0Mbp-101.0Mbp

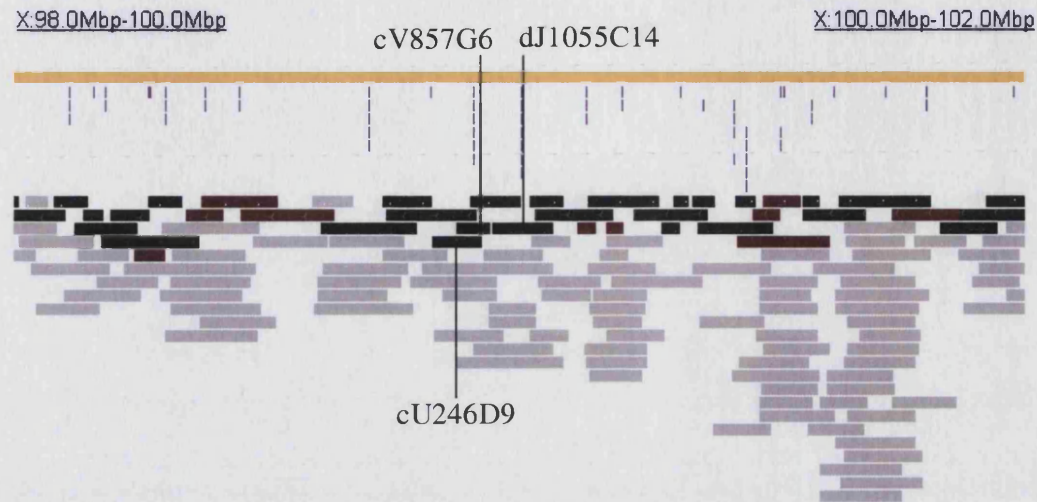


Figure 5.8. Screenshot from Sanger Institute X chromosome FPC map based on restriction analysis of the whole clones showing relative positions and sizes of the genomic clones, cU246D9, cV857G6 and dJ1055C14, in a 2Mb window. Relative sizes of genomic clones are shown by the horizontal boxes, which are colour-coded according to the sequencing status at the time this map was last modified (on 12/12/2002 according to information on the website).

URL: <http://www.sanger.ac.uk/cgi-bin/humace/fpcwebmap.cgi?mode=map&map=bac.X.98.html>

5.7. Analysis of genomic sequence around the two proximal breakpoints in family 2

5.7.1. Genes near breakpoint closer to *PLP1*

The distal breakpoint, at position 59809 within the sequence originating from dJ1055C14 falls within the third exon of a known gene, *MORF4L2* (Mortality factor 4-like 2) (Figure 5.9.). This gene has also been known as *MRGX* (MORF-related gene X), and is part of a gene family of transcription factors that seem to have roles in transcriptional activation and repression (Bertram and Pereira-Smith, 2001; Tominaga *et al.*, 2003). *MORF4L2* contains various motifs including a leucine zipper, helix-loop-helix domain, ATP/GTP binding domain and nuclear localisation signal (Tominaga *et al.*, 2003). *MORF4L2* has four exons, but only the longer fourth exon contains the coding sequence; the first three short exons together make up part of the 5' UTR, and it is transcribed from the reverse strand of dJ1055C14. The breakpoint falls before the coding sequence of *MORF4L2*, which remains single copy, therefore it is unlikely that this protein would be overexpressed as a result of the duplication in this family.

5.7.2. Genes near proximal breakpoint

The breakpoint within cU177E8 did not interrupt any known genes. The closest annotated gene to this breakpoint is also within cU177E8, 8.7Kb distal to the breakpoint and therefore expected to be within the duplication. This nearby gene is uncharacterised and annotated into the Ensembl genome browser under the identifier NM_153333. The gene shows sequence similarities to *TCEAL1*, which is a nuclear phosphoprotein that may modulate transcription (see section 3.5.4.1.) (Pillutla *et al.*, 1999). Interestingly, *TCEAL1* is mapped within genomic clone dJ1055C14, but is nearly 50Kb 5' to the breakpoint in family 2 (Table 3.11.).

5.7.3. Proximal region-specific repeats

Neither of the genomic regions brought together in the breakpoint sequenced in family 2 are particularly close to any of the repeated sequences that have been described in the region proximal to *PLP1* (see Figure 3.8.). The nearest repeat units are part of the “P4” grouping, a family of repeated sequences which all include coding sequences related to *TCEAL1* (see sections 3.5.4. and 5.7.2. and Table 3.10.).

5.7.4. Interspersed repetitive elements and G+C content

5Kb of sequence centred on the major sequenced breakpoints in family 2 (position 37339 in cU177E8 and position 59809 in dJ1055C14) was analysed by Repeatmasker to determine the numbers and types of common interspersed repeats. Both breakpoint regions contained a low proportion of interspersed repeats compared to the average proximal to *PLP1* (see section 3.5.1.). The 5Kb around the cU177E8 breakpoint contained 24.24% interspersed repeats (Figure 5.9. and Table 5.5.). The 5Kb centred on the dJ1055C14 breakpoint contained a similar percentage of interspersed repeats, 26.76% (Figure 5.9. and Table 5.6.). Neither of the two breakpoints fell within a known interspersed repetitive element (Tables 5.5. and 5.6.). The G+C content near both breakpoints was slightly below the average for the region (40.60% G+C in the 1Mb region proximal to *PLP1*), at 40.24% near the cU177E8 breakpoints and 39.36% near the dJ1055C14 breakpoint.

Distance from breakpoint (bp)	Repeat type	Position in repeat consensus sequence	Orientation
(-) 2492-2280	L1ME (LINE/L1)	5319-5545 (227/6121)	–
(-) 1977-1813	MIR (SINE/MIR)	10-204 (195/262)	+
(-) 1821-1764	L2 (LINE/L2)	3254-3313 (60/3313)	–
(-) 1739-1695	(CA) _n (Simple repeat)	2-46	+
(-) 819-527	<i>AluY</i> (SINE/ <i>Alu</i>)	1-296 (297/311)	–
(-) 526-342	L1ME4a (LINE/L1)	5901-6103 (203/6121)	–
(-) 246-65	L1MC/D (LINE/L1)	2270-2452 (183/6215)	+
(+) 18-93	CT-rich (Low complexity)	1-76	+
(+) 2288-2328	(CA) _n (Simple repeat)	2-42	+
(+) 2328-2357	(GA) _n (Simple repeat)	2-31	+
(+) 2376-2693	<i>AluSx</i> (SINE/ <i>Alu</i>)	1-317(318/318)	–

Table 5.5. Interspersed repetitive elements from the 5Kb region surrounding the breakpoint in cU177E8 (at position 37339bp in this genomic clone). The first column shows how far in base pairs each repeat element is from the breakpoint, (-) indicates that the repeat element is proximal to the breakpoint and (+) that the repeat is distal relative to the breakpoint. The type of repeat and the class of repeat element to which it belongs are given in the second column. The third column shows which portions within the appropriate repeat consensus sequence each repeat has similarity to, and also how many bases out of the total repeat unit consensus is present in the sequence. In the fourth column, the orientation of each repeat is shown; + indicates a repeat is on the forward strand (*i.e.* running from centromere to telomere), - the repeat is on the reverse strand. One interspersed repeat in the region, an *AluSx* copy, overlapped with the boundary of the 5Kb region, and the adjacent sequence was also searched for repeated sequences until the end of this *Alu* element was found. Interspersed repeat sequences were searched for using Repeatmasker software (Repbased update v7.4). The other short section from cU177E8 that is inserted in the breakpoint in family 2 is between 1285-1333bp distal to the centre of the 5Kb region.

Distance from breakpoint (bp)	Repeat type	Position in repeat consensus sequence	Orientation
(-) 205-135	L2 (LINE/L2)	3201-3272 (72/3272)	–
(+) 229-543	L2 (LINE/L2)	1939-2275 (337/3314)	–
(+) 910-1060	MIR (SINE/MIR)	63-230 (168/262)	+
(+) 1082-1270	L1MC4a (LINE/L1)	5679-5878 (200/7808)	+
(+) 1331-1639	<i>AluSg</i> (SINE/ <i>Alu</i>)	1-292 (293/310)	+
(+) 1968-2007	(A) _n (Simple repeat)	1-40	+
(+) 2011-2313	<i>AluSx</i> (SINE/ <i>Alu</i>)	1-305 (306/312)	+

Table 5.6. Interspersed repetitive elements from the 5Kb region surrounding the breakpoint in dJ1055C14 (at position 59809bp in this genomic clone). The first column shows how far in base pairs each repeat element is from the breakpoint, (-) indicates that the repeat element is proximal to the breakpoint and (+) that the repeat is distal relative to the breakpoint. The type of repeat and the class of repeat element to which it belongs are given in the second column. The third column shows which portions within the appropriate repeat consensus sequence each repeat has similarity to, and also how many bases out of the total repeat unit consensus is present in the sequence. In the fourth column, the orientation of each repeat is shown; + indicates a repeat is on the forward strand (*i.e.* running from centromere to telomere), - the repeat is on the reverse strand. Interspersed repeat sequences were searched for using Repeatmasker software (Repbased update v7.4).

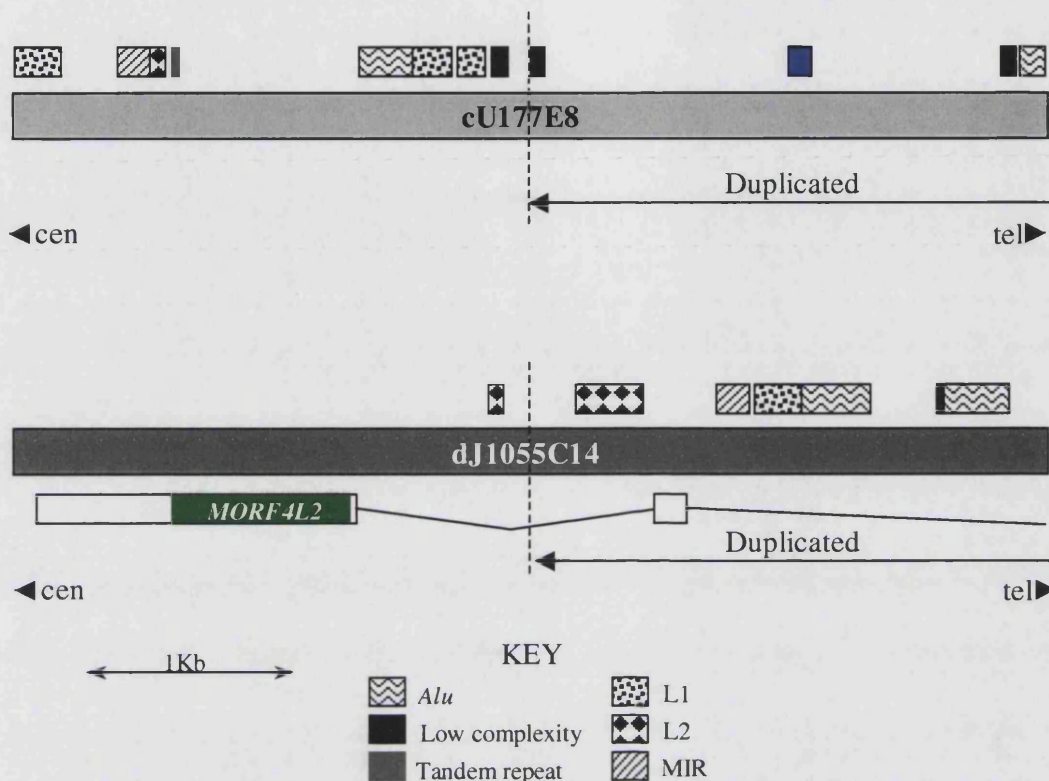


Figure 5.9. Interspersed repeat content around breakpoints in family 2. 5Kb around each breakpoint (dashed vertical line) is shown. The upper part of the figure shown the repeat content centred on position 37339bp in cU177E8, and the lower half of the diagram shows the region centred on position 59809bp in dJ1055C14. The different types of human interspersed repetitive elements are represented by the patterned and shaded boxes above each genomic clone. The original location of the short stretch of additional sequence from cU177E8 (36054-36006bp) present at the breakpoint is shown by the blue box. The locations of exons from a gene spanning the breakpoint (*MORF4L2*) are shown underneath dJ1055C14, unfilled boxes are non-coding sequence and the green box shows the coding sequence of *MORF4L2*. This gene is transcribed from the reverse strand, *i.e.* from right to left on this diagram.

5.7.5. Comparisons of sequence around cU177E8 and dJ1055C14 breakpoints

The 5Kb of sequence around both ends of the sequenced breakpoint in family 2 were compared against each other using the BLASTz algorithm, to determine if there were any similarities between the breakpoint regions that may have mediated the rearrangements (see section 2.2.10.1.3.) (Schwartz *et al.*, 2000). The only parts of the sequence that were found to be similar near the two breakpoints corresponded to copies of *Alu* elements (Figure 5.10.). In addition, 100bp regions flanking the breakpoints were also aligned together, using ClustalW. There was no extensive homology found between these sequences outside of the short (4bp and 2bp) sequence overlaps at the junctions (Figure 5.11.).

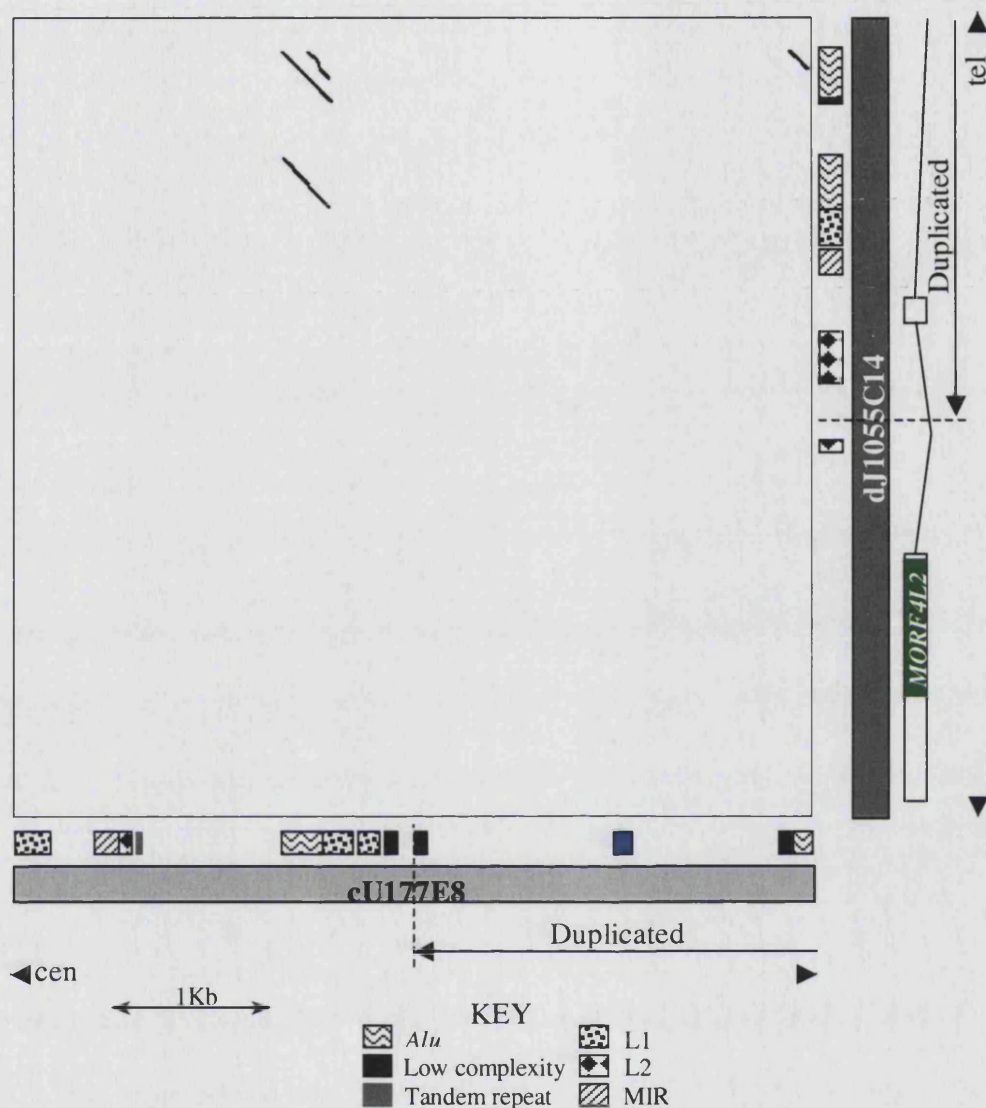


Figure 5.10. Comparison of 5Kb around breakpoints by BLASTz (Pipmaker dotplot output). The results of the comparison are shown as a dotplot, where any similarities are shown as black dots on the white background. The human genomic clones and interspersed repeat content are shown in the relevant positions next to the dotplot. The position of each breakpoint is indicated by the dashed lines. For more explanation of the repeat content, see Figure 5.9., Tables 5.5. and 5.6. Upwards-facing diagonal lines on the dotplot (/) are indicative of directly repeated sequences; downwards-sloping lines on the dotplot (\) represent areas present in both sequences in an inverted orientation.

Junction sequence GAGGAGGAGAAAGAGGAGGAGAAAGAGGAGGAGAAGAAAAGTGTCCAGAAGAGACAGGCTAGTGTCTGGCCCAGAGGATGAGGGGGAGAAAG
cU177E8 37289 GAGGAGAAAGAGGAGGAGAAAGAGGAGGAGAAGAAAAGTGTCCAGAAGAGATTTTTAAAAATT-----AAGAAATAAATTTAGAGTA
cU177E8 36104 TTCCTCTTATTCAGGTCTTTTTTTATTTTTGCAGCACTCCCCTTTCTGTTGAGACAGGCTAGTGTCTGGCCCAGAGGATGAGGGGGAGAAAG
dJ1055C14 59759 ATGGTACATGACACCACTGTCAGTGATTGTTTAAAGAGGTAAAAAAGA---

Junction sequence ACACTTT---GGTGAGGGAGGAGTCAGTGATG-----GGTGATTAGGTTATTAGTATTATCACCAAATGATAATACTCCTCCAT
cU177E8 ATGCATC---GGATGAAAAGA 37388
cU177E8 ACACTTTCCCAGAATGGAGCAGGTCGCG-TG-----AGCAACTGGACTTGGAGCTGTG-CATC 35956
dJ1055C14 -----TT---GGTGAGGGAGGAGTCAGTGATGACTCCCAGGTATGGGTGATTAGGTTATTAGTATTATCACCAAATGATAATACTCCTCCAT 59892

Figure 5.11. Alignment of breakpoint junction and 100bp flanking the sequenced breakpoint in family 2. Each breakpoint region was individually aligned by ClustalW to the sequence from the junction breakpoint and the alignments were then combined together manually. Junction sequence that originated from the more centromeric part of cU177E8 (37289-37339) is highlighted in red, and nucleotides aligned with this sequence from the genomic sequence around the distal breakpoint are also highlighted in red. Junction sequence from the slightly more distal sequence within cU177E8 (36006-36054) is shaded blue, and bases aligned with this sequence are also shaded blue. The 4bp overlap between these two sequences is shaded purple. Junction sequence originating from dJ1055C14 (59809-59832; 59843-59892) is light green, and any bases from the genomic sequences matching this part of the junction sequence in the alignment with this are also shown in green. The 2bp overlap at the junction between dJ1055C14 and cU177E8 is coloured pink. The 3bp overlap in sequence between the ends of the short duplication in dJ1055C14 sequence is coloured a darker green

5.7.6. Recombination/rearrangement-associated motifs

5.7.6.1. 5Kb around breakpoint in cU177E8

Only a few sequence motifs (*S. pombe* ARS, scaffold attachment regions) were found to be over-represented in the 5kb around the major breakpoint in cU177E8, which also included the short sequence from cU177E8 inserted at the junction, but none were very close to the various breakpoints (Appendix C). One motif, (TTTAAA) that can induce curvature in the double helix, was present just 2bp distal to the duplication breakpoint (see Appendix C, Figure 5.14.) (Crothers *et al.*, 1990; Toffolatti *et al.*, 2002). Some sequence motifs were found near the 34bp stretch of sequence that was inserted at the breakpoint (Appendix C). Two copies of the murine parvovirus genome deletion hotspot (CTWTTY) were located upstream of the breakpoint, the nearest ending 4bp before the junction sequence (Appendix C) (Hogan and Faust, 1986). One short sequence motif, (CAGR), which has been associated with deletion breakpoints in mice, was found near both ends of the inserted sequence (Appendix C, Figure 5.14.) (Steinmetz *et al.*, 1986).

5.7.6.2. 5kb around breakpoint in dJ1055C14

Several sequence motifs (the chi sequence GCTGGTGG, the V(D)J nonamer motif ACAAAAACC, the *S.cerevisiae* ARS, scaffold attachment consensus sequences and the murine MHC recombination hotspot sequence CAGRCAGR) were also over-represented in the region around the breakpoint in dJ1055C14, but were not located close to the breakpoint (Appendix C) (Smith *et al.*, 1981; Steinmetz *et al.*, 1986; Akira *et al.*, 1987; Krowczynska *et al.*, 1990; Shiroishi *et al.*, 1995). A single copy of the budding yeast consensus replication origin sequence (WTTTATRTTTW), which was over-represented relative to the expected frequency, was just 161bp distal to the breakpoint (Appendix C).

Some sequence motifs were very close to the major breakpoint in dJ1055C14 in family 2. These included a DNA polymerase β frameshift hotspot (TTTT) and the murine parvovirus recombination hotspot (CTWTTY), all of which were either just proximal to or overlapping with the cU177E8-dJ1055C14 breakpoint in family 2 (Figure 5.14. and Appendix C) (Been *et al.*, 1984;Hogan and Faust, 1986).

5.7.7. Matrix attachment regions

Both 5Kb regions around the breakpoints were searched for potential MARs by the MAR-Wiz program (see section 4.9.1.2.) (Singh *et al.*, 1997). Regions of high MAR potential were located close to the breakpoints in both cU177E8 and dJ1055C14 (Figure 5.13.). It is possible that normally distant sequences could be brought into closer association if both regions were associated with the chromosome scaffold with intervening sequences looped out, which could facilitate rearrangement of these sequences. As the MAR potential predicted using MAR-Wiz can be dependent on the sequence context, a larger (20Kb) region surrounding the breakpoint was also tested for the presence of MARs (Namciu *et al.*, 2004). When this was attempted, neither of the two potential MARs from the 5Kb regions were predicted to have the highest MAR potential in the region, but there was still a peak present at the relevant location in both regions, although eclipsed by other nearby regions of greater MAR potential. Although the MAR-Wiz program is only a predictor of MAR regions, and these would need to be experimentally verified for confirmation, it is still possible that the two breakpoints do fall within genuine scaffold associated regions.

5.7.8. *In silico* analysis of sequence from 100bp regions flanking breakpoints

Three separate short regions of sequence immediately flanking the breakpoints and inserted sequence were analysed in family 2, the main cU177E8 breakpoint (37289-37388bp), the additional sequence from cU177E8 (36104-35956bp) and the dJ1055C14 breakpoint (59759-59892bp) (Figure 5.4.).

5.7.8.1. Alternating purine/pyrimidines, polypurine and polypyrimidine tracts

The most striking feature of the purine/pyrimidine content close to the breakpoints in family 2 is a long homopurine tract just before the first breakpoint in cU177E8 (Figure 5.14.). This 85bp run of purines ends just 9bp before the 4bp breakpoint region and contains 97.6% purine residues (Figure 5.14.). This region of low sequence complexity was identified by Repeatmasker as a CT rich sequence (Repeatmasker was using sequence from the other strand in these analyses) (Table 5.5.). No other homopurine, homopyrimidine or alternating purine/pyrimidine tracts of 10bp or greater were present in this 100bp region from cU177E8 (Figure 5.14.).

A single polypurine tract (14bp) and one polypyrimidine tract (12bp) were noted in the 134bp that included the short sequence inserted at the breakpoint from cU177E8, with each type accounting for about half this percentage amount (Figure 5.14.). The polypurine tract was within the sequence inserted at the breakpoint (Figure 5.14.).

The sequence from around the more distal breakpoint in dJ1055C14 just contained two 10bp polypurine tracts, each just either side of the cU177E8-dJ1055C14 breakpoint (Figure 5.14.).

To determine whether the occurrence of a polypurine tract near the breakpoint in cU177E8 was unusual, a 100Kb region (centred on the breakpoint at 37339bp in cU177E8) was searched for polypurine tracts of between 40-140bp in length on both DNA strands using the DNA Pattern Find program (see section 2.2.10.3.). 3 regions with at least 40 contiguous purine residues were found, including the region nearest to the breakpoint. One of the other tracts consisted of a 59bp homopurine run (mainly adenines), associated with the 3' end of an *AluSg/x* repeat located in dJ635G19, 18Kb proximal to the breakpoint. This A-rich sequence was probably derived from the poly(A) tail during retrotransposition. The other long homopurine.homopyrimidine tract was positioned more proximal to the breakpoint (43Kb), also within dJ635G19. This region consisted of 116bp of pure pyrimidines (or purines on the other strand), which comprised several short tandem repeats of sequence, including 39bp of (TTTTC)_n tandem repeats and 56bp matching a (TTCC)_n tandem repeat array, as found by the Tandem Repeats Finder programme (see section 2.2.10.8.) (Benson, 1999). This homopurine.homopyrimidine tract was also located close to the 3' end of a partial *AluJb* element, and could have been derived from a poly(A) sequence associated with this repeat, with a subsequent expansion of the homopurine tract by either slipped strand mispairing during DNA replication, or by misaligned intra-allelic recombination (Arcot *et al.*, 1995).

Polypurine.polypyrimidine tracts of 40bp or longer were also searched for in a 100Kb region around the breakpoint in dJ1055C14. Four regions were found, one of which was 2Kb distal to the breakpoint, a 40bp A-rich sequence that had also been identified by Repeatmasker analysis of the region (Table 5.6.). One of the polypurine tracts, which was 18Kb proximal to the breakpoint, was 138bp in length, and was located within an MSTA LTR element sequence.

This analysis indicates that homopurine.homopyrimidine tracts of greater than 40bp in length, such as the one found close to the cU177E8 breakpoint, are not particularly common within these regions of genomic sequence, occurring 3-4 times per 100Kb, so the association of one of these regions of low complexity sequence with one of the breakpoints in family 2 may be important.

5.7.8.2. Repeats and secondary structures

Both the original genomic sequences from the breakpoint regions and junction sequences were searched for the presence of various types of repeats (Figure 5.15.). Repeated sequences in the DNA may predispose to the formation of secondary structures, which could be involved in stimulating the rearrangements, or alternatively the presence of repetitive sequences could help to stabilise the DNA during the rearrangement processes (see section 4.9.2.2.) (Chuzhanova *et al.*, 2003). A few different types of repeats were found in the original genomic sequence (Figure 5.15.). In the junction sequence, however, many different repeats, of all the different types (direct, inverted, symmetric, complementary) were present (Figure 5.15b.). Many of these repeated sequences were associated with the polypurine tracts in the junction sequence (Figures 5.14. and 5.15.).

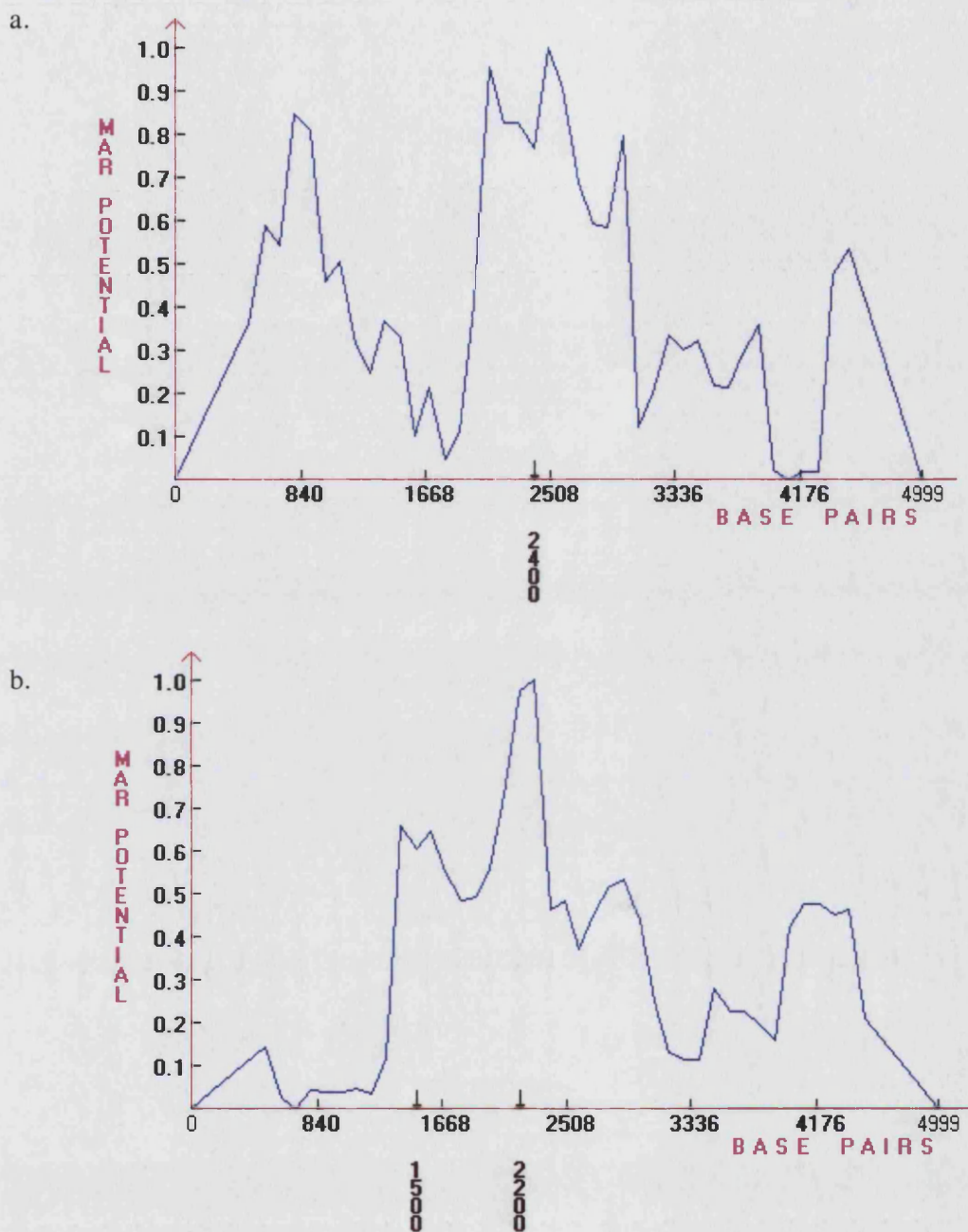


Figure 5.12. Position of potential MARs near the proximal breakpoints in family 2. (a.) is centred on the breakpoint in cU177E8 and (b.) shows the 5Kb region around the breakpoint in dJ1055C14. MAR potential is as found by the MAR-Wiz program (Singh *et al.*, 1997).

Main breakpoint in cU177E8 (37289-37388bp)

GAGGAGAAAGAGGAGGAGAAAAGAGGAGGAGAAGAAAAAGTGTCCAGAAAGAGATTTTATAAAATTAAAGAAATAAATTTAGAGTAATGCATCGGATGAAAAGA
RRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRRYYYYYRRRRRRYYYYYRRRRRRYYRRRRRYRRRRYYRRRYRRYYRRRYRRRRRR

49bp insertion from cU177E8 into the breakpoint (36104-35956bp)

[illegible]

CCAGAATGGAGCAGGTCTGCGTGAGCAACTGGACTTGAGCTGTGCATC
YYRRRRYRRRRYRRRRYYRYRYRRRRYRRYYRRRRYYRRRRYYRYRYRY

Main breakpoint and short deletion from dJ1055C14 (59759-59892bp)

[illegible]

TTAGTATTATCACCAAATGATAATACTCCTCCAT
YYRRYRYRRYRYRRYRRYRRYRRYRYRRYRRYRRY

Figure 5.13. Nucleotide composition of short regions surrounding the breakpoints sequenced from family 2. 50bp either side of the various breakpoints are shown on the top of the pairs, and the classification of each nucleotide (R – purine, Y – pyrimidine) is shown on the lower line of the pairs. Tracts of alternating purines/pyrimidines (R/Y), or pyrimidines/purines (Y/R), of 10 nucleotides or greater, are underlined. Homopurine and homopyrimidine runs of 10 nucleotides or greater are in bold type. Sequences not found at the breakpoint are shaded grey, sequences present in the rearrangement are shaded in various colours, as in Figure 5.13.

Main breakpoint in cU177E8 (37289-37388bp)

GAGGAGAAAGAGGAGGAGAAAGAGGAGAGAAGAAAGTGTCCAGAAGAGATTTTAAATAAGAAATAAATTTAGAGTAATGCATCGGATGAAAAGA

49bp insertion from cU177E8 into the breakpoint (36104-35956bp)

TTCTCTTAT**TT**CAGGTCTTTTTTATTTTGCAGCACTCCCCTTTTCTGTT**GAGACAGGCTAGTGTCTGGCCCAGAGGATGAGGGGGAGAAAGACACT**TTTC

CCAGAATGGAG**CAGGTCT**GCCTGAGCAACTGGACTTGGAGCTGTGCATC

Main breakpoint and short deletion from dJ1055C14 (59759-59892bp)

ATGGTACATGACACCACT**GTCAGTGATT**GTTTAAGAAGGTAAAAAAGAT**TTGGTGAGGGAGGAGTCAGTGATG**ACTCCAGGTATGGGTGATTAGGT**TA**

TTAGTATTATCACCAAATGATAATACTCCTCCAT

Figure 5.14a. Location of internal sequence repeats around the breakpoint regions in family 2.

Junction sequence



Figure 5.14b. Location of internal sequence repeats within the junction sequence from family 2. Legend on next page.

Figure 5.14. Legend. Internal repeats around the breakpoints and junction fragment in family 2. Figure 5.14a. shows the normal sequence, with 50bp on either side from around the breakpoints in cU177E8 and dJ1055C14, and Figure 5.15b. shows the junction sequence. Sequences found at the breakpoint are coloured as before in Figure 5.11., and the sequences not found in the final rearrangement are shaded grey. Nucleotides that form part of repeat sequences are shown in bold. The type of repeat found is shown by the arrows above and below the sequence: direct (→), inverted (←), symmetric (→) and complementary (→). Repeats were searched for using Oligorep (see sections 4.9.2.2. and 2.2.10.5.).

5.8. Discussion

5.8.1. Sequencing of a breakpoint in family 2 has revealed a complex rearrangement

This work has demonstrated that the affected boy and carrier mother from family 2 carry a complicated rearrangement of the *PLP1* genomic region that has resulted in the duplication of *PLP1*, and a phenotype of PMD in the affected boy [2:9]. Previous studies had mapped the most proximal and distal extent of the duplicated regions in this family without detailed investigation of the dosage of sequences within the apparently duplicated region (apart from *PLP1*), and had assumed that this family carried a simple tandem duplication, as already described in family 1 (Figure 4.1.) (Woodward *et al.*, 2000). It has now been shown that there are at least two non-contiguous regions of duplicated sequence within Xq22.2 in this family. The breakpoint sequenced from this family so far only includes sequence proximal to *PLP1* on the X chromosome, and as *PLP1* is duplicated there must also be at least one breakpoint distal to *PLP1*. The probable location of another breakpoint proximal to *PLP1* in family 2 has been mapped by UPQFM-PCR in this study to the centromeric portion of genomic clone cV857G6 (Figure 5.7. and Table 5.4.). It is tempting to speculate that the breakpoint junction in this region may be joined to sequence distal to *PLP1*, which would result in an inversion of the cU177E8-cV857G6 duplicated region, with the duplicated sequence including *PLP1* remaining in the original orientation (Figure 5.15.). Until further sequence from the other breakpoint(s) in this family is obtained, the true nature of the rearrangement cannot be fully determined. The putative rearrangement suggested in Figure 5.15. is probably the simplest scenario based on the sequence and dosage data from this family so far, but it is just as likely that there are several more undetermined breakpoints involved in this rearrangement. The X chromosomes from this family appear to have a normal

morphology, so large-scale rearrangements involving the X or other chromosomes are unlikely to be present. Previous interphase FISH data using clones distal to *PLP1* indicated that this family had a duplication that extended to a similar region as in family 1, with a breakpoint possibly near or within the distal LCRs (Figures 3.4., 4.1. and 4.7.) (Woodward *et al.*, 2000).

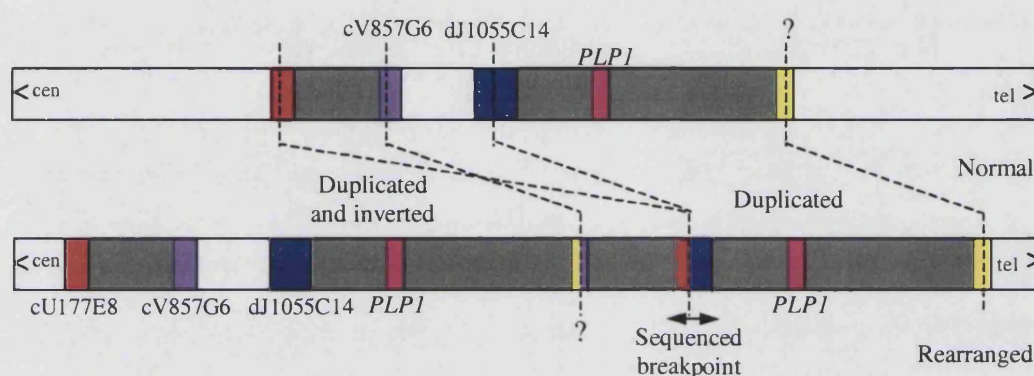


Figure 5.15. Diagram showing a possible rearrangement in family 2, based on the sequencing, UPQFM-PCR and interphase FISH data gathered on this family to date. The upper part of the diagram shows a normal X chromosome (not to scale), with the various coloured boxes indicating regions involved in the rearrangement. Dashed lines show the position of breakpoints (either based on sequence and UPQFM-PCR data, or just theoretical). cU177E8 is red, cV857G6 is lilac, dJ1055C14 is blue, *PLP1* is coloured pink, and a region around a theoretical distal breakpoint is yellow. The grey areas show the areas of the sequence that have may have been duplicated in the rearrangement. The lower part of the diagram shows the possible rearrangement. A length of sequence from cU177E8 to cV857G6 is duplicated and inverted, with sequence from cU177E8 joined to dJ1055C14, as has been found by sequencing (Figure 5.4.). A breakpoint distal to *PLP1* joins on to the end of the duplicated sequence from cV857G6, resulting in inverted and duplicated sequence originating proximal to *PLP1* being between two directly repeated stretches of sequence including *PLP1*.

5.8.2. Nature and possible mechanisms of the rearrangement

As with the duplication breakpoint in family 1, the breakpoint sequenced from family 2 appeared most compatible with a mechanism of non-homologous end joining. There was no major sequence homology found between any of the breakpoints or flanking sequences involved in the rearrangement, and the different breakpoints contain a short sequence overlap (2-4bp), all of which is typical of repair by NHEJ (Figures 5.4. and 5.13.) (reviewed in (Lieber *et al.*, 2003)). As discussed before, in chapter 4, NHEJ is a mechanism used for the repair of double-strand breaks in DNA, and the unusual rearrangement in family 2, which has been partially characterized in this chapter, may have resulted from illegitimate joining of the ends of different DSBs on the X chromosome by NHEJ (see section 4.12.4.). It is also a possibility that a similar mechanism to that proposed for the simple tandem duplications, with homologous strand invasion and non-homologous repair could also be involved in the rearrangement in family 2 (see section 4.12.5.). The sequence close to the sequenced breakpoints has been investigated for the presence of various sequence motifs and other factors that may have contributed to the generation of DSBs. Some intriguing possibilities have been noted and are discussed further in sections 5.8.2.2., 5.8.2.3. and 5.8.2.4.

5.8.2.1. Insertion of a short sequence at the breakpoint

The junction between cU177E8 and dJ1055C14 sequence also contains 49bp of additional inserted sequence, from cU177E8, originating from a sequence expected to lie 1.3Kb distal to the breakpoint within the duplicated region (see section 5.4.3.). Although the majority of breakpoints repaired by NHEJ are joined at sites of microhomology, approximately 10-20% of NHEJ junctions in mammalian cells contain insertions of extra nucleotides at the breakpoint, termed “filler DNA” (Roth *et*

al., 1989;Roth *et al.*, 1991;Lieber *et al.*, 2003). Most filler DNA sequence is very short, typically just consisting of a single nucleotide, but in rarer cases it can be much longer, with tens or hundreds of extra nucleotides added between the two sequences (Roth *et al.*, 1989;Merrihew *et al.*, 1996). As has been found with the breakpoint in family 2, filler DNA sequences can originate from elsewhere in the genome (Williams and Fried, 1986;Merrihew *et al.*, 1996;Sargent *et al.*, 1997;Lin and Waldman, 2001a;Little and Chartrand, 2004). Inserted filler DNA can also originate from extrachromosomal DNA, such as viral genomes present in transformed cell lines, retrotransposable elements, introduced oligonucleotide fragments and plasmids (Roth *et al.*, 1991;Sargent *et al.*, 1997;Lin and Waldman, 2001a;Lin and Waldman, 2001b).

5.8.2.2. Recombination/rearrangement associated motifs

Various short sequence motifs that have been found to be involved in recombination, rearrangements and DNA breaks were searched for near the breakpoints and some were found at higher than expected frequencies (see sections 5.7.7.1. and 5.7.9.1). However, it is unclear whether any of these motifs could be involved in the breakage and rearrangements in this family. Generally, each breakpoint region examined had a different complement of motifs in the vicinity, whether at an increased frequency or not (see section 5.7.7.2. and 5.7.9.1., Appendix C). It is difficult to determine if any of these motifs could be involved in the breakage and rearrangement, but this cannot be discounted at this stage.

5.8.2.3. Polypurine tracts and short repeats

One of the most striking features of the sequence near the breakpoint in family 2 is the polypurine tract just within the duplicated sequence from cU177E8 (Figure 5.13.). As well as this sequence having low nucleotide complexity (83bp of either adenine or

guanine, interspersed by two isolated thymines), there are also some internal repeat sequences, both direct and symmetric, found near the breakpoint (see section 5.7.9.3, Figures 5.13. and 5.14.). The analysis in Figure 5.14a. is only concerned with the 38bp region of the purine tract closest to the breakpoint and does not include the whole of the tract. When the 85bp of purine-rich sequence is looked at in isolation using the same method and settings to find internal repeated sequences, a large overlapping direct repeat (27 nucleotides in length, with 4 mismatches) was found, as well as a perfect symmetric repeat 16bp in length (Figure 5.16a.). There does seem to be an association between stretches of purine residues and the other breakpoint regions so far identified in family 2, with polypurine tracts greater than 10bp in length found close to most breakpoints (Figure 5.14.). Polypurine tracts have been found to be over-represented at both deletion and translocation breakpoints and may themselves be involved in the mechanisms of DNA rearrangement (Abeyasinghe *et al.*, 2003).

5.8.2.4. Triplex DNA and secondary structure

The presence of a perfect symmetric (or mirror) repeat is interesting, as sequences of this nature have been shown, under various conditions to form a DNA triplex structure (Mirkin *et al.*, 1987; Kohwi and Kohwi-Shigematsu, 1988; Kohwi, 1989; Bernues *et al.*, 1990; Dayn *et al.*, 1992; Kohwi and Kohwi-Shigematsu, 1993; Gilbert and Feigon, 1999). The triplex formation of DNA involves the binding of single stranded DNA in the major groove of the conventional B-DNA double helix, with the third DNA strand able to specifically bind to the base pairs in the double helix via Hoogsteen hydrogen bonds (Figure 5.16c.). There are several different possible types of DNA triplex, including one where a third homopyrimidine strand binds to the homopurine strand within the double helix in a parallel orientation,

forming the triplet bases (T-A-T and C⁺-G-C), and another possible triplex structure occurs when a single homopurine strand binds within the major groove in an antiparallel orientation, forming the triplets (A-A-T and G-G-C) (Reviewed in (Gilbert and Feigon, 1999)). The pyrimidine-purine-pyrimidine triplex structure forms more readily under acidic conditions, as one of the triplex base pairs requires a protonated cytosine (C⁺) on the third strand (Mirkin *et al.*, 1987; Gilbert and Feigon, 1999). Other factors can affect the propensity of susceptible sequences to form triplex structures, such as the degree of DNA supercoiling and the presence of some metal ions, eg Mg²⁺, *in vitro* (Kohwi and Kohwi-Shigematsu, 1988; Kohwi, 1989; Bernues *et al.*, 1990; Dayn *et al.*, 1992; Kohwi and Kohwi-Shigematsu, 1993).

The mirror repeat near the breakpoint in cU177E8 could lead to the formation of a DNA triplex, as described above, and one possible triplex conformation of this DNA sequence is shown in Figure 5.16b. DNA replication has been shown to terminate at triplex structures associated with homopurine tracts and other triplex-forming sequences (Lapidot *et al.*, 1989; Dayn *et al.*, 1992). One source of a double-strand break in this sequence could possibly be as a result of replication termination at a secondary structure formed at the mirror repeat in the polypurine tract. The triplex-forming potential of some sequences within human DNA has been proposed as a mechanism in the pathogenesis of some disorders. The formation of cysts during the progression of autosomal dominant polycystic kidney disease has been suggested to be triggered by the loss of function of the normal allele of the *PKD1* gene, and the high rate of cyst formation has been attributed to the presence of a long (2.5Kb) 97% pyrimidine tract, within which replication blockage has been shown to occur frequently *in vitro* (Patel *et al.*, 2004). Triplex DNA has also been suggested to be involved in large-scale mitochondrial DNA rearrangements and may also trigger

recombination (Kohwi and Panchenko, 1993; Rooney and Moore, 1995; Rocher *et al.*, 2002; Napierala *et al.*, 2004). Triplex forming DNA sequences may be involved in bringing distant sequences together in the nucleus and in other aspects of chromatin structure (Ohno *et al.*, 2002).

Another way in which a triplex secondary DNA structure could be involved in the rearrangement seen at the cU177E8-dJ1055C14 breakpoint in family 2 is by stabilizing the rearrangement. As well as the perfect mirror repeat in the polypurine tract, there is an imperfect copy (13/16 identical nucleotides) of one of the repeat units within the 49bp inserted between the two main sequences at the breakpoint (also from cU177E8, but originating from 1.3Kb away distal to the breakpoint) (Figure 5.16a.). It is possible that this short sequence containing the partial copy of the repeat could interact with the triplex-forming mirror repeat, and the resultant proximity of these sequences may have been important in the rearrangement, or the formation of a triplex may have stabilized the rearrangement (Chuzhanova *et al.*, 2003). Some possible interactions between the various repeats are shown in Figures 5.16d and 5.16e. A triple helix could form between the two sequences, e.g. double stranded DNA from one of the mirror repeats could form a triple helix with the one of the DNA strands with a partial copy of the repeat (Figure 5.16d.). Another potential configuration could be a triplex forming from the symmetric repeat sequence in the polypurine tract, as described before in Figure 5.16b, and one strand of the partial repeat could perhaps form a short double helix with the unpaired strand (Figure 5.16e.).

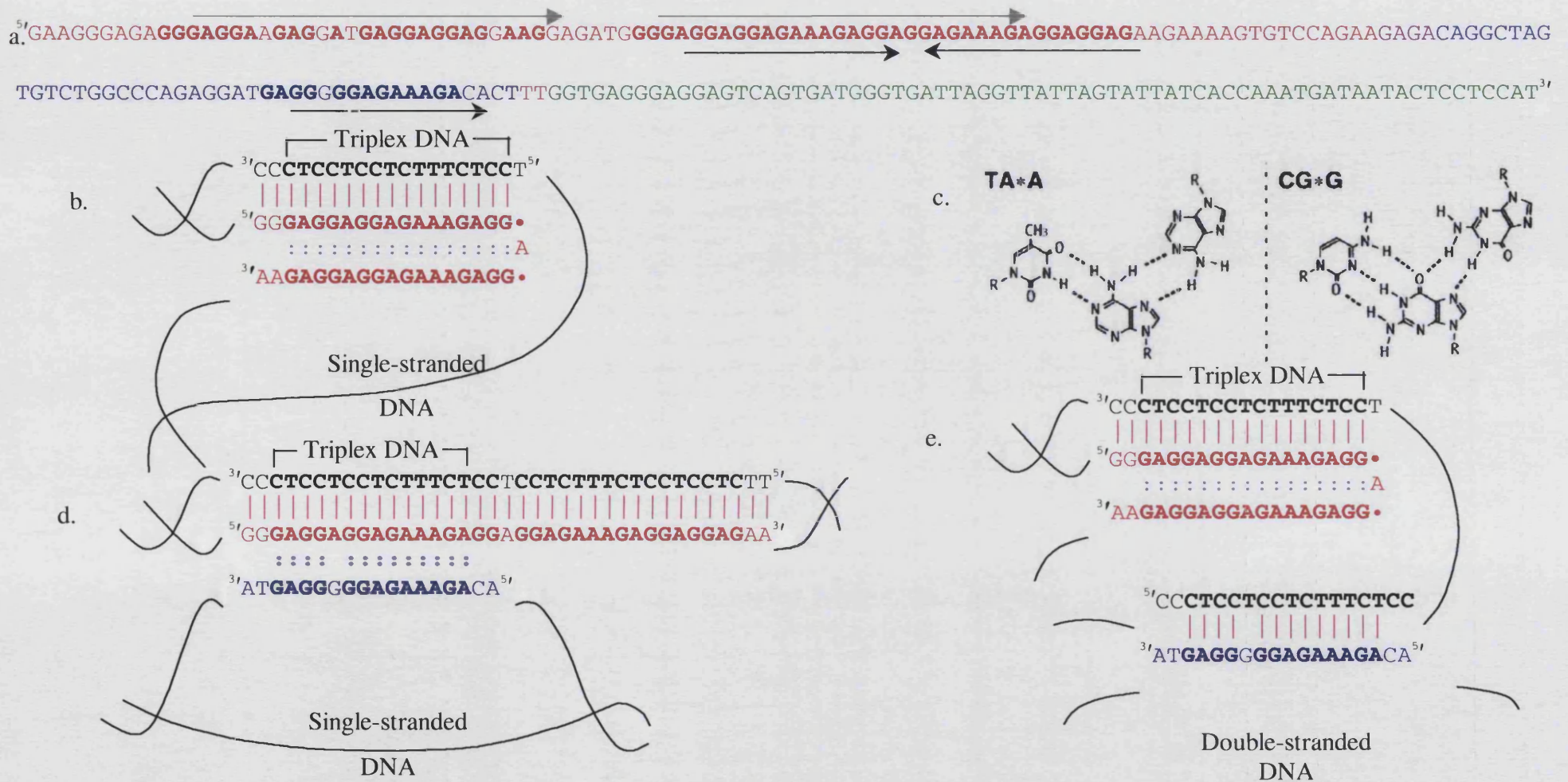


Figure 5.16. Symmetric repeats, Hoogsteen base pairs, and possible triplex DNA configurations. Figure legend on next page.

Figure 5.17. Legend. Figure showing location of repeats within the polypurine tract close to the breakpoint in family 2, and some possible secondary structures resulting from the symmetrical repeats. (a.) shows the sequence from the cU177E8-dJ1055C14 breakpoint, and all of the polypurine tract near the major breakpoint in cU177E8. The various sequences are shaded as in Figure 5.12., sequence up to position 37335 in cU177E8 is red, between 36010-36052bp in cU177E8 (inserted sequence) is blue and sequence from dJ1055C14, starting at position 59811bp is green. Overlaps between the sequence at the breakpoints are shaded purple or pink. Internally repeated sequences within the polypurine tract are shown by arrows, grey and above the sequence for direct repeats, black and underneath the sequence for symmetric repeats. Nucleotides contained in the repeats are highlighted in bold type. A partial match to one of the symmetric repeats within the inserted cU177E8 sequence is also underlined. (c.) Some Hoogsteen base pairings that occur in triplex DNA structures. Taken from (Vetcher *et al.*, 2002). (b.) A triplex DNA structure that could be formed by one (single-stranded) half of the mirror repeat folding back on itself and forming Hoogsteen bonds with the other (double-stranded) half of the homopurine symmetric repeat, in an antiparallel orientation, leaving the remaining strand single-stranded. The various sequences are coloured as in (a.), and sequences from the complementary polypyrimidine strand are coloured black. Watson-Crick hydrogen bonds are shown as pink vertical lines and Hoogsteen bonds as lilac double dots. (d.) A possible triplex secondary structure involving part of the inserted cU177E8 sequence and one of the two halves of the mirror repeat. (e.) Another possible secondary structure that could be formed between the two cU177E8 sections of sequence. A triplex involving the mirror repeat is formed (as in b.) and the partial copy of one of the repeats from the inserted sequence could base-pair with the unpaired loop containing the complimentary copy of one of the mirror repeats.

5.8.2.5. DNA replication and rearrangement in family 2

Near the main junction sequence in family 2 from dJ1055C14 sequence, there was found to be a short deletion of 13bp, which in the genomic sequence for this clone is flanked by a 3bp (ATG) direct repeat (Figures 5.4. and 5.12.). While this deletion could result from repair of a DSB by NHEJ, another possible cause of this type of small sequence deletion is replication slippage, which may occur following polymerase pausing or stalling on the template (Viguera *et al.*, 2001). It is also a possibility that the other breakpoints could be attributed to replication termination at secondary structural motifs causing a DSB.

5.8.2.6. Capture of long sequences at DSBs

Most studies of sequence capture at DSBs have only characterised short sequences inserted into a few NHEJ repair events (Roth *et al.*, 1989; Merrihew *et al.*, 1996; Lin and Waldman, 2001b). However, many of the protocols used to identify filler sequence capture at DSBs have been PCR-based, so if a large-scale rearrangement or large insertion of sequence into the DSB has occurred this may either be too large to be amplified by PCR, or may no longer be flanked by the appropriate primer-binding sites. The occurrence of these types of events would hence be underestimated, and other reports have suggested that a significant proportion of NHEJ DSB repair events may involve either a large insertion of filler sequence at the break, or may trigger a large-scale rearrangement (Lin and Waldman, 2001a; Allen *et al.*, 2003). The rearrangement discovered in family 2 could be a result of a large DNA fragment, e.g. one originating from between cU177E8 and cV857G6, being involved in NHEJ repair of a more distal DSB event (Figure 5.15.). There is at least one more, as yet unsequenced, breakpoint in the *PLP1* region in this family, and as the sequences just proximal to the breakpoint and end of the duplicated region in dJ1055C14 are not

duplicated, it is unlikely that the proximal cU177E8-cV857G6 duplicated segment is simply inserted into a DSB in this region (Figure 5.7., Tables 5.3. and 5.4.). Fragile sites are loci where gaps or breaks in chromosomes can be induced under certain conditions (Rassool *et al.*, 1991). It has been shown that exogenous DNA integrates preferentially into some fragile sites, which may be as a result of capture of sequences during DSB repair (Rassool *et al.*, 1991; Lin and Waldman, 2001a).

It is still possible that the rearrangement in family 2 could result from a process of strand invasion by free DNA ends, followed by replication, similar to that described in section 4.12.4. Although the microhomologies observed at the breakpoints in family 2 have been attributed to NHEJ repair, a “copy-join” model has been proposed for random integration of extrachromosomal or transfected DNA, where the ends of the transfected DNA prime synthesis on a single-stranded chromosomal region by binding to a few homologous nucleotides within the single stranded region (Merrihew *et al.*, 1996). In this manner, segments of genomic sequence from several different sites could be picked up by a single free end, initially generated by a DSB, and joined together by regions of microhomology (Allen *et al.*, 2003).

5.9. Summary

Through a combination of experimental techniques, a rearrangement breakpoint has been sequenced in family 2. Contrary to expectations based on existing data, the *PLP1*-containing duplication in this family was discovered not to be an uncomplicated tandem duplication event (as found in several other cases of *PLP1* duplications) (Woodward *et al.*, in preparation; Iwaki *et al.*, 2003). Instead the rearrangement involved at least two non-contiguous duplicated regions, one of which was inverted with respect to the original sequence during the rearrangement. These data from family 2 demonstrate the importance of determining the actual sequence present at these types of duplications for enabling understanding of the mechanisms behind these events.

6.1. ANALYSIS OF FAMILY 3

Previous work had found that an affected male (3:6) and his carrier mother (3:4) had an unusual duplication of *PLP1*, with the duplicated copy of *PLP1* located in Xq26 (Figure 2.2.) (Hodes *et al.*, 2000; Woodward *et al.*, 2003). This rearrangement was not visible on G-banded chromosomes. In addition, the female carrier was found to carry a mosaic deletion of part of Xq in the chromosome that harboured the duplication (Hodes *et al.*, 2000; Woodward *et al.*, 2003). This deletion encompassed both copies of *PLP1*. Various strategies were pursued in an attempt to characterise the various duplication and deletion breakpoints in this family, to gain insight into the molecular mechanisms involved, and also to determine if the duplication and Xq deletion were unrelated events, or if they were connected in some way.

6.2. Xq26 *PLP1* insertion family mosaic deletion analysis

6.2.1. Metaphase FISH

Mapping of the size of the deletion and the position of the breakpoints in the mosaic female carrier with a *PLP1* duplication and a large deletion was initially carried out using metaphase FISH. Previous work had shown that both an *XIST* FISH probe mapping to Xq13.3, and an Xq subtelomeric FISH probe, gave a signal on the X chromosome containing the deletion (Woodward *et al.*, 2003). Biotinylated genomic PAC and BAC clones were hybridised to metaphase chromosomes, and fluorescently detected using FITC conjugated to avidin. Using this strategy, the proximal breakpoint was found to lie within bA346E8 located in Xq21.1 (Table 6.1a, Figure 6.1.). The distal breakpoint was found to be between the adjacent clones bA183K14 and bA79A21, in the Xq27.3 cytogenetic band (Table 6.1b, Figure 6.2.). According to data from the Ensembl genome browser, this interstitial deletion included approximately 65Mb, and comprised over half of the long arm of the X chromosome.

(a)

Proximal clones	band	+/-
dJ570L12 (<i>PGK1</i>)	Xq21.1	+
dJ795G23	Xq21.1	+
bA102P23 (<i>SH3BGRL</i>)	Xq21.1	+
bA346E8	Xq21.1	+/-
bB52K8	Xq21.1	-
dJ2A2	Xq21.1	-
dJ717L12	Xq21.1	-
dJ326L13 (<i>POU3F4</i>)	Xq21.1	-
dA43C13 (<i>REP1</i>)	Xq21.2	-

(b)

Distal clones	band	+/-
dJ260J9 (<i>FGF13</i>)	Xq26.3	-
bA51C14 (<i>SOX3</i>)	Xq27.1	-
<u>dJ177G6</u>	Xq27.1	-
dJ507I15	Xq27.1	-
dJ406C18	Xq27.2	-
dJ73H14	Xq27.2	-
dJ357K22	Xq27.3	-
bG256O22	Xq27.3	-
dJ145B12	Xq27.3	-
bA550B3	Xq27.3	-
bA159A24	Xq27.3	-
bG278N14	Xq27.3	-
bA183K14	Xq27.3	-
bA79A21	Xq27.3	+
dJ203P18	Xq27.3	+

Tables 6.1a and b. Results of deletion breakpoint analysis. Clones are listed in order along the X chromosome, centromere to Xq telomere. + indicates that a FISH signal was seen for that clone on the X chromosome containing the deletion, - indicates that a FISH signal was not seen for that clone on the X chromosome containing the deletion. Some of the known genes located in the clones are in brackets after the clone name.

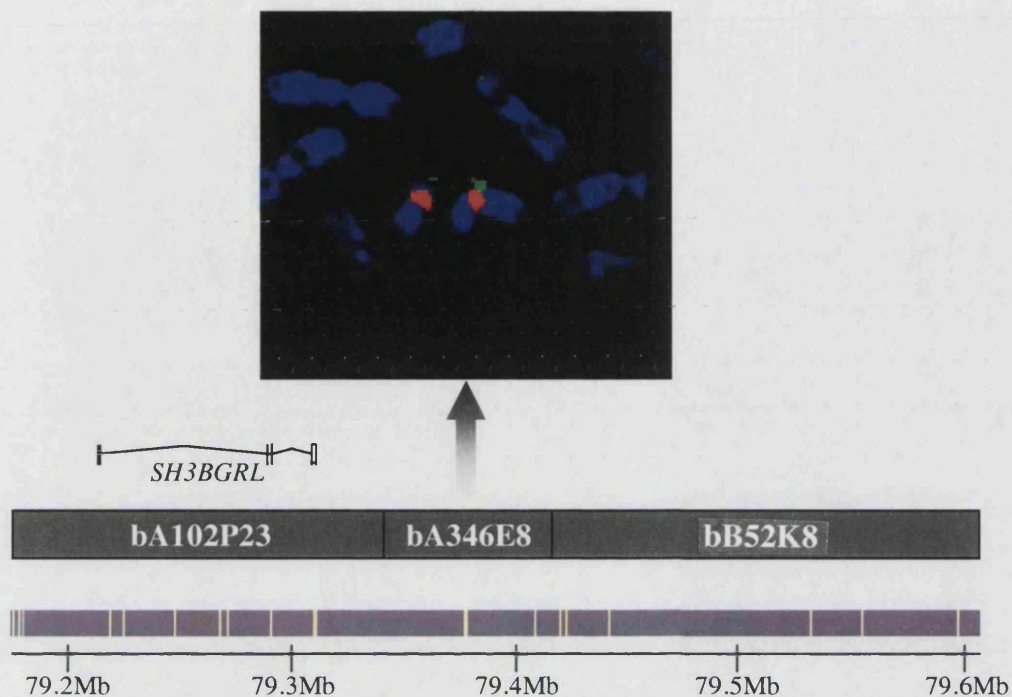


Figure 6.1. FISH mapping of the proximal deletion breakpoint. Metaphase chromosomes from the mother in the Xq26 insertion family, with probe bA346E8 signal shown in green (Xq21.1) and a red X centromere probe. The signal for bA346E8 is much fainter on the deleted X chromosome, making it likely that this clone contains a deletion breakpoint. The positions of bA346E8 and adjacent clones are shown as dark grey boxes, the scale shows the position on the X chromosome according to the Ensembl genome browser (version 21.34d.1). The location of the *SH3BGRL* gene, which is close to the breakpoint-containing clone, is shown, and interspersed repeat content (as displayed on the Ensembl genome browser, calculated using Repeatmasker) is shown just above the scale bar.

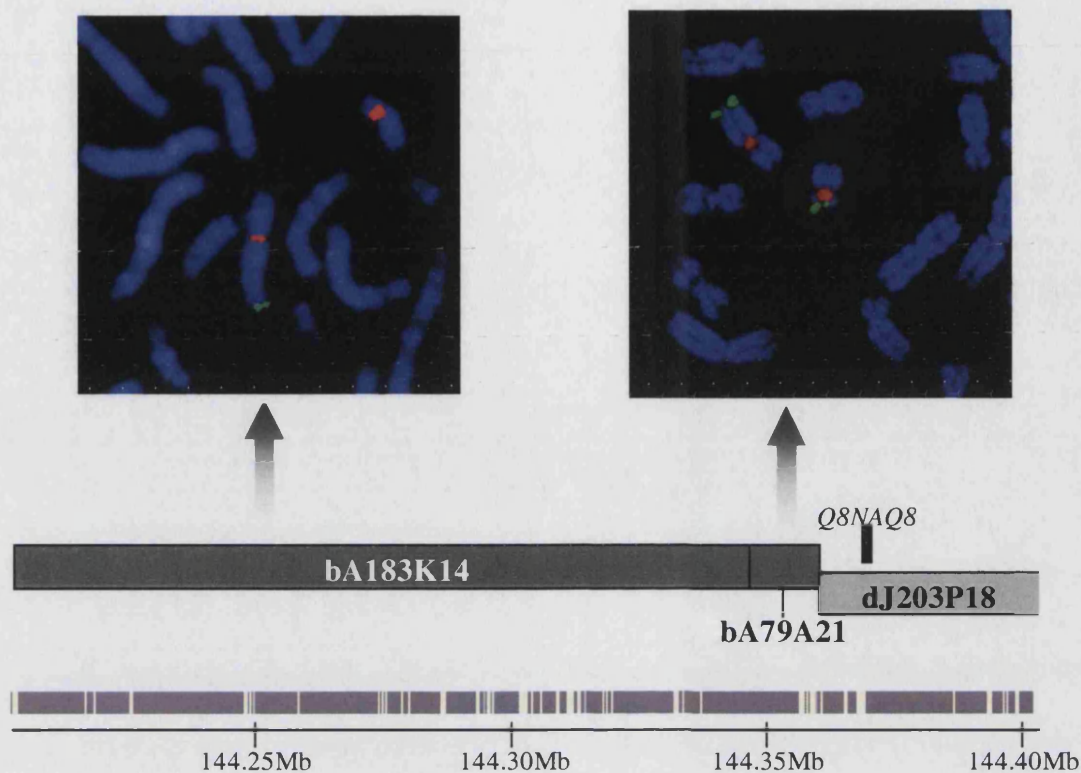


Figure 6.2. FISH mapping the distal deletion breakpoint. Metaphase spreads are shown from the mother in the family with an insertion of *PLP1* into Xq26. The X centromere probe is shown in red and the two genomic clones are each shown in green on the metaphase chromosomes. The positions of bA183K14, bA79A21 and the proximal portion of an adjacent clone, dJ203P18, are shown as grey boxes, the scale shows the position on the X chromosome according to the Ensembl genome browser (version 21.34d.1). The location of a nearby transcript, *Q8NAQ8*, which is close to the breakpoint-containing clone, is shown, and interspersed repeat content (as displayed on the Ensembl genome browser, calculated using Repeatmasker) is shown just above the scale bar.

6.2.2. STS analysis on flow-sorted chromosomes

A cell line from the carrier female that was known to contain both the deleted and duplicated cells was flow-sorted (by Nigel Carter, Sanger Institute), so that the two differently sized populations of X chromosomes were separated. This enabled further mapping of the precise locations of the ends of the deletion using STS mapping (sequence tagged site), as STS sequences that were contained within the deletion would only be able to be amplified from the full sized chromosomes, not from the chromosome with the deletion. Flow sorting was not able to separate the normal X and the X with the duplication, as they only differed in size by approximately 600kb. Only a small amount of DNA was recovered following flow sorting (approximately 50 chromosomes), so to provide more material for analysis, the DNA from the flow-sorted chromosomes was further amplified using degenerate oligonucleotide primed PCR (DOP-PCR, see section 2.2.1.2.) (Telenius *et al.*, 1992). Primer pairs were then designed to amplify short regularly spaced stretches of sequence (STSs), that were within the genomic regions where the ends of the deletion had been mapped using metaphase FISH. These STS primers were then used in PCRs with the DOP-PCR amplified flow-sorted chromosomes as a template. This made it possible to narrow down the location of the breakpoints further, as STSs that failed to amplify from the deleted chromosome but that did give a PCR product from the full size X chromosomes were most probably included in the mosaic deletion (Table 6.2.). For the proximal breakpoint, the PCR results agreed with the FISH results, placing the breakpoint within bA346E8, between positions 1142–8732bp in the submitted sequence for this clone. bA346E8 had shown a reduced signal on the deleted X by FISH (Figure 6.1.).

6.2.2.1. Disparity between FISH and STS mapping of deletion breakpoint

At the distal breakpoint the FISH and STS-PCR results were apparently discordant. By metaphase FISH bA79A21 was present on the deleted X, with no obvious reduction of signal intensity (Figure 6.2.). However, the STS used that was within the published sequence for bA79A21 did not amplify by PCR from the deleted X chromosome, which was consistent with bA79A21 being within the deleted region (Table 6.2.). STS-PCR results for the adjacent distal clone, dJ203P18, showed that the proximal deletion breakpoint was within that clone, between positions 97926-98798bp in Z97180, the sequence submitted for this clone (Table 6.2.). A substantial overlap between these clones was confirmed by using some of the STS primers that had been used for the deletion mapping on DNA from the two clones bA79A21 and dJ203P18 (Figures 6.3a and 6.3b). This mapped the telomeric end of clone bA79A21 to between 52464bp and 4928bp in Z97180, which is within the telomeric half of the submitted sequence for dJ203P18 (Figure 6.3b). This indicated that at least 40Kb of the human genomic DNA present in BAC bA79A21 fell beyond the deleted region, which would probably be enough to account for the metaphase FISH results showing that this clone was not reduced in signal intensity on the deleted X chromosome (Figure 6.2. and Table 6.2.). In addition, data showing the relative sizes of the clones in this region, based on restriction fragment fingerprint data available from the Sanger Institute shows that bA79A21 and dJ203P18 do overlap substantially (Figure 6.4.).

Clone	Strand	Position of STS in clone	Full size chromosomes	Deletion chromosome
Proximal deletion breakpoint				
bA102P23	+	132693-132870	+	+
bA102P23	+	133192-133394	+	+
bA346E8	+	952-1142	+	+
bA346E8	+	8386-8763	+	-
bA346E8	+	17767-17970	+	-
bA346E8	+	26616-26815	+	-
bA346E8	+	69417-69621	+	-
bA52K8	+	23495-23699	+	-
Distal deletion breakpoint				
bG278H4	+	159160-159378	+	-
bA183K14	+	84019-84233	+	-
bA183K14	+	129708-129900	+	-
bA79A21	+	10058-10260	+	-
dJ203P18	-	126544-126349	+	-
dJ203P18	-	116587-116365	+	-
dJ203P18	-	106044-105705	+	-
dJ203P18	-	98827-98450	+	-
dJ203P18	-	97926-97768	+	+
dJ203P18	-	96946-96708	+	+
dJ203P18	-	95757-95518	+	+
dJ203P18	-	95565-95369	+	+
dJ203P18	-	87934-87737	+	+
dJ203P18	-	79512-79234	+	+
dJ203P18	-	75170-74970	+	+
dJ203P18	-	52664-52464	+	+
dJ203P18	-	4928-4745	+	+

Table 6.2. Results for STS mapping of Xq deletion breakpoint. (+) indicates that an STS did amplify from the flow-sorted chromosomes, (-) that it failed to amplify.

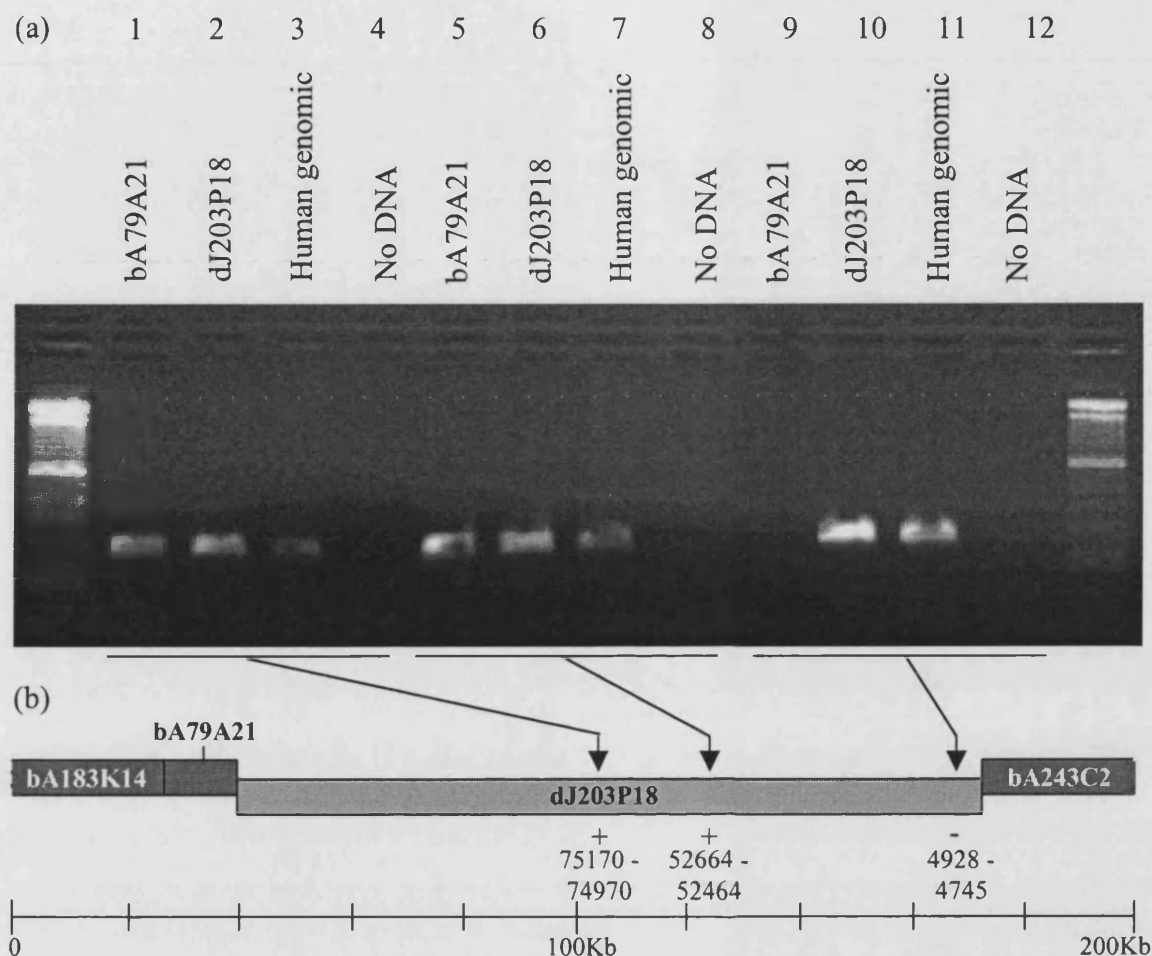


Figure 6.3. bA79A21 overlaps substantially with dJ203P18. Figure 6.3a shows agarose gel electrophoresis of PCRs using selected STS primers from within Z97180 (Z97180, the sequence submitted for dJ203P18, is on the reverse strand). Lanes 1-4 are using primers that amplify sequence between 75170-74970bp, lanes 5-8 are using primers for 52664-52464bp and lanes 9-12 is using primers for sequence between 4928-4745bp (All positions given within Z97180). PCR product was seen in all reactions using clone dJ203P18 and human genomic DNA as a substrate, but for bA79A21 product was only seen for the two proximal sets of primers, and not for the primers from the most telomeric end of dJ203P18. 100bp ladder size marker is in the two outside lanes. Figure 6.3b shows a 200Kb region including bA79A21 and dJ203P18, with the approximate location of the STS primers used indicated by the arrows underneath dJ203P18. A plus sign (+) underneath a primer pair indicates that it did amplify from bA79A21, a minus sign (-) indicates that the primer pair did not amplify from bA79A21 DNA. Clones that have been sequenced on the forward strand are shown as dark grey boxes, dJ203P18 has been sequenced on the reverse strand and is shown as a lighter grey box.

Detailed view: X:144.0Mbp-146.0Mbp



Figure 6.4. Screenshot from Sanger Institute X chromosome fingerprinted contig map based on restriction analysis of the whole clones showing relative positions and sizes of genomic clones, including dJ203P18 and cU79A21, in a 2Mb window. Relative sizes of genomic clones are shown by the horizontal boxes, which are colour-coded according to the sequencing status at the time this map was last modified (on 12/12/2002 according to information on the website).

URL: <http://www.sanger.ac.uk/cgi-bin/humace/fpcwebmap.cgi?mode=map&map=bac.X.144.html>

6.2.3. Amplifying deletion breakpoint by LR-PCR

As the two deletion breakpoints were mapped to relatively short genomic regions (just over 7.5Kb for the proximal end, and 872bp at the distal end), long-range PCR was carried out to obtain a junction product spanning the deletion breakpoint (Figure 6.5.). Three primers were designed in the 7.5Kb candidate region for the proximal breakpoint in clone bA346E8 and each was used in a separate reaction with a single reverse primer from dJ203P18 (Figure 6.5.). Products were seen for all three reactions after PCR and agarose gel electrophoresis using DNA template from 3:4 who carried the mosaic deletion and also a cell line from that individual (Figure 6.5b). As expected, no PCR products were produced from a reaction containing normal male control DNA, confirming that a novel breakpoint junction was being amplified (Figure 6.5b).

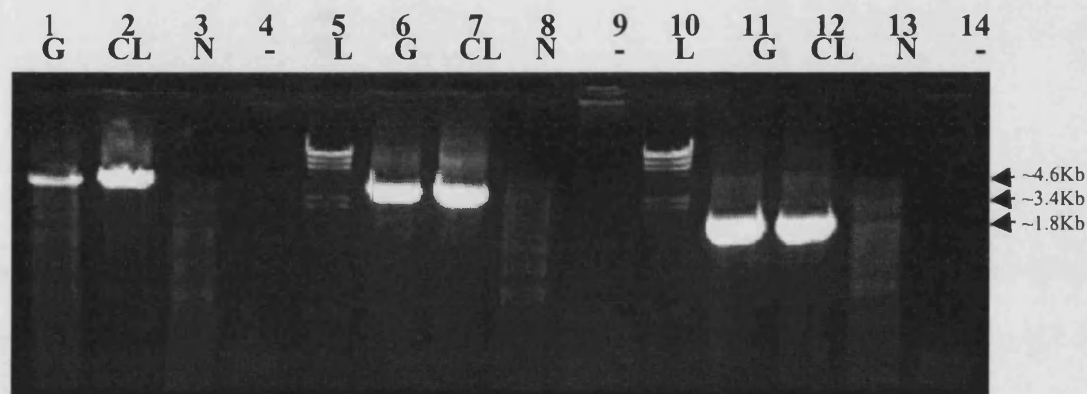
6.2.4. Sequencing of deletion breakpoint

The junction PCR products from the carrier female were gel-purified, and the shortest product was sequenced (see sections 2.2.1.4.4. and 2.2.4.) (Lane 11 in Figure 6.5b). Initial sequencing did not reveal the deletion breakpoint, so further nested primers were designed (starting at positions 4815 in bA346E8 and 98454 in dJ203P18). These primers amplified a shorter fragment (485bp) from carrier female genomic DNA and also from the previous 1.8Kb LR-PCR product, and they were used to completely sequence the breakpoint in the shorter nested PCR product (see section 2.2.4.).

(a)



(b)



Figures 6.5a and 6.5b. Long PCR across the Xq deletion breakpoint. (a.) shows the position of the primers used for long PCR and sequencing relative to the proximal and distal deletion breakpoint regions (not to scale). The reverse primer used within dJ203P18 had previously been used for the deletion breakpoint mapping and was used in all three reactions. The names of the two primers that were used to sequence the breakpoint are shown in bold type. (b.) shows the PCR products using these primers after agarose gel electrophoresis and staining with ethidium bromide. G, genomic DNA from 3:4; CL, DNA from a cell line from 3:4 with a large proportion of cells containing the deletion; N, normal male genomic DNA; -, no DNA; L, λ /Hind III ladder. Lanes 1-4 are from a reaction using the 1446F primer, and show PCR products of approximately 4.6Kb in the genomic and cell line DNA, but not in a reaction using normal control DNA. Similarly, lanes 6-9 were from reactions containing the 2596F primer and lanes 11-14 used the 4250F primer, and showed PCR products at sizes of approximately 3.4Kb and 1.8Kb respectively, only in DNA from 3:4.

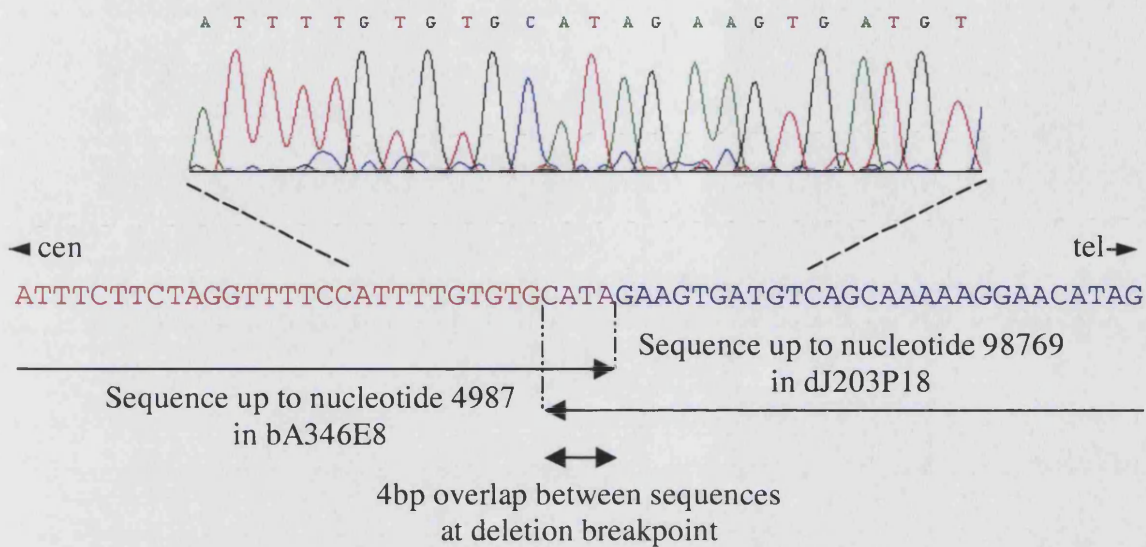


Figure 6.6. Sequence from the deletion breakpoint in the female carrier. The electropherogram shows the breakpoint and ten nucleotides on either side from a sequencing reaction using a proximal breakpoint primer. More of the sequence found around the breakpoint is shown underneath, with sequence that originated from bA346E8 (Xq21.1) highlighted in red, and sequence from dJ203P18 (Xq27.3) coloured blue. At the breakpoint there is a 4bp overlap between the two clones, which is in purple type.

6.2.5. Analysis of deletion breakpoints

6.2.5.1. Interspersed repeats and G+C content

Analysis of the sequence of the deletion breakpoint, using BLASTn, showed that the 485bp junction fragment consisted of 173bp of sequence originating from the proximal clone bA346E8 and 316bp of sequence from the distal clone dJ203P18, with a 4 nucleotide overlap at the junction between the two sequences (Figure 6.6.). At the proximal end, the breakpoint fell within a long interspersed repeat element (L1P4), at position 3414 in the repeat consensus (Figure 6.7., Table 6.3.). The distal end of the deletion was in apparently unique sequence, with the nearest sequence features that were picked up by Repeatmasker analysis being an almost complete *AluSx* element, which was 320bp proximal to the breakpoint, and a (TA)_n simple repeat, which was 1092bp distal to the breakpoint (Figure 6.7., Table 6.4.). Both the proximal and distal breakpoint regions contained numerous repetitive elements. In the 5Kb surrounding the proximal breakpoint interspersed repeat sequences made up almost all of the sequence (99.42% interspersed repetitive elements, 0.46% simple repeats, with 6 nucleotides remaining unmasked by Repeatmasker), and 57.08% of the 5Kb of sequence around the distal breakpoint was also characterised as belonging to interspersed repetitive elements by Repeatmasker software (Tables 6.3. and 6.4.). The especially high repeat content near the proximal deletion breakpoint consisted mainly of fragments of an L1P4 repeat element (Table 6.3.). The 5kb surrounding the proximal deletion breakpoint had a G+C content of 36.66%, slightly lower than the average for the surrounding 1Mb (37.01%), and the 5Kb region around the distal deletion breakpoint had a G+C content of 39.82%, higher than the 36.55% G+C content in the surrounding 1Mb.

6.2.5.2. Comparing sequences near the deletion breakpoints

There was no obvious sequence similarity, apart from the 4bp overlap, between the sequences around the proximal and distal deletion breakpoints. Both BLAST2 and BLASTz were used to compare 5kb centred on each breakpoint against each other, and no large regions of similarity were found between the two sequences (Figures 6.8. and 6.9.). The only major similarities between the two sequences corresponded to regions containing *Alu* repetitive elements (Figure 6.9.). Alignment of the normal genomic sequences immediately surrounding the two deletion breakpoints by ClustalW did not show any great similarity between the two sequences, with no contiguous alignments of greater than four nucleotides seen (Figure 6.8.). In total there was 40.82% identity between the two sequences over the regions that were compared immediately around the breakpoint, and not including the 4bp overlap at the junction (Figure 6.8.).

6.2.5.3. Gene content near deletion breakpoints

Both the proximal and distal deletion breakpoints were located in gene-poor regions of the genome. At the proximal end, a 1Mb region centred on the deletion breakpoint only contained one annotated gene, *SH3BGRL* (SH3 domain-binding glutamic acid-rich-like protein), which was located 55.6Kb centromeric to the breakpoint (the location of this gene is shown in Figure 6.1.). The 1Mb region around the distal breakpoint contains similarly few potential genes, with just one predicted gene annotated, *Q8NAQ8*, which was 24.9Kb centromeric to the breakpoint (Figure 6.2.). *Q8NAQ8* is a single exon transcript, coding for a protein of unknown function, without any significant homology to other known proteins.

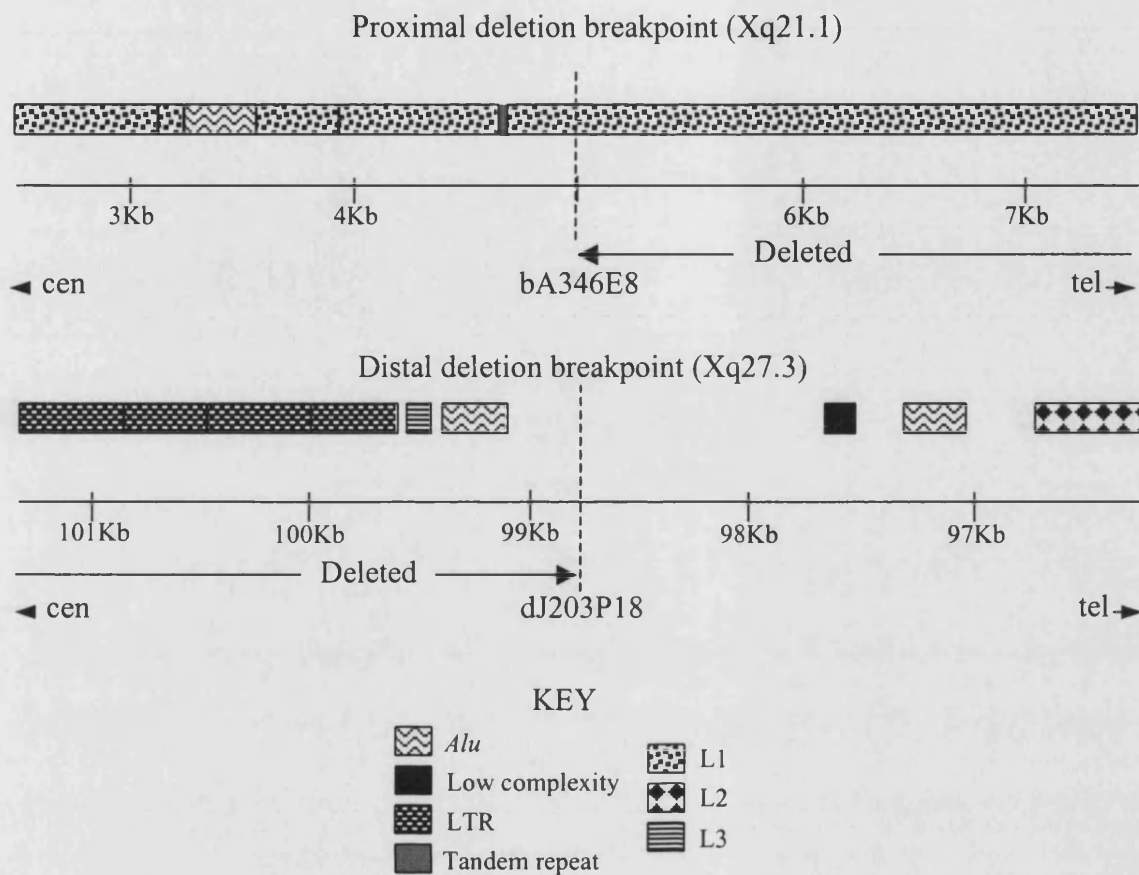


Figure 6.7. Figure showing 5Kb around the proximal and distal deletion breakpoint and the interspersed repeat content of these regions. The scale bars show distances from the start of the available sequence for each clone. The dashed vertical lines indicate the position of both breakpoints. Shaded and patterned boxes show the various types of repeats present.

Distance from breakpoint (bp)	Repeat type	Position in repeat consensus sequence	Orientation
(-) 2493-1922	L1P (LINE/L1)	4212-4788 (577/6146)	-
(-) 1931-1738	L1P4 (LINE/L1)	5307-5498 (192/6165)	-
(-) 1737-1423	<i>AluJb</i> (SINE/ <i>Alu</i>)	3-308 (306/312)	-
(-) 1422-1031	L1P4 (LINE/L1)	4906-5303 (398/6165)	-
(-) 1030-330	L1P4 (LINE/L1)	3714-4399 (686/6146)	-
(-) 339-307	(TTG) _n (simple repeat)	2-24	+
(-) 306-4683 (+)	L1P4 (LINE/L1)	5-3714 (3710/6146)	-

Table 6.3. Repeat content (using Repeatmasker software, Repbase version 7.4) of the 5Kb genomic region centred on the proximal deletion breakpoint (at position 4987) within bA346E8. The first column shows how far in base pairs each repeat element is from the deletion breakpoint, (-) indicates that the repeat element is proximal to the breakpoint and (+) that the repeat is distal relative to the breakpoint. The type of repeat and the class of repeat element to which it belongs are given in the second column. The third column shows which portions within the appropriate repeat consensus sequence each repeat has similarity to, and also how many bases out of the total repeat unit are present in the sequence. In the fourth column, the orientation of each repeat is shown; + shows a repeat is on the forward strand (i.e. running from centromere to telomere), - the repeat is on the reverse strand. If a repeat element was found to coincide with the beginning or end of the 5kb segment, the adjacent sequence was also analysed with Repeatmasker, until the true end of the repeat was found.

Distance from breakpoint	Repeat type	Position in repeat consensus sequence	Orientation
(-) 2500-2030	THE1D-int (LTR/MaLR)	630-1095 (466/1580)	+
(-) 2025-1670	THE1B (LTR/MaLR)	6-364 (359/364)	+
(-) 1669-1186	THE1D-int (LTR/MaLR)	1090-1580 (491/1580)	+
(-) 1185-812	THE1D (LTR/MaLR)	1-381 (381/381)	+
(-) 769-657	L3 (LINE/CR1)	412-535 (124/1577)	+
(-) 615-319	<i>AluSx</i> (SINE/ <i>Alu</i>)	1-297 (297/312)	-
(+) 1093-1203	(TA) _n	1-111	+
(+) 1203-1224	(GA) _n	2-23	+
(+) 1439-1725	<i>AluSg</i> (SINE/ <i>Alu</i>)	1-299 (299/310)	+
(+) 2020-2493	L2 (LINE/L2)	2268-2755 (488/3314)	+

Table 6.4. Repeat content (using Repeatmasker software, Replibase version 7.4) of the 5Kb genomic region centred on the distal deletion breakpoint (at position 98769) within dJ203P18. The first column shows how far in base pairs each repeat element is from the deletion breakpoint, (-) indicates that the repeat element is proximal to the breakpoint and (+) that the repeat is distal relative to the breakpoint. The type of repeat and the class of repeat element to which it belongs are given in the second column. The third column shows which sections within the appropriate repeat consensus sequence each repeat has similarity to, and also how many bases out of the total repeat unit are present in the sequence. In the fourth column, the orientation of each repeat is shown; + shows a repeat is on the forward strand (i.e. running from centromere to telomere), - the repeat is on the reverse strand.

```

bA346E8 4937 -TGTTTCCATGAATTTATTCATTTCT-TCTAGGTTTTCCATTTTGTGTGCA
Junction      -TGTTTCCATGAATTTATTCATTTCT-TCTAGGTTTTCCATTTTGTGTGCA
dJ203p18 98818 CCAGGCAGACAAATAAAGAAAATGATATTTATGGTCACTTTATGTGACCA

bA346E8      TACTTGTGTTCTTGGTAGTCTCTGAGGTTTTTAGT--ATTCTGTGTGTTCAA 5036
Junction      TAGAAGTGATGTCAGCAAA----AAGGAACATAGCTGGATACTAAGTAAAAAA
dJ203p18      TAGAAGTGATGTCAGCAAA----AAGGAACATAGCTGGATACTAAGTAAAAAA 98719

```

Figure 6.8. Alignment of deletion breakpoint junction and 100bp flanking the two deletion breakpoints. Each breakpoint region was individually aligned by ClustalW to the sequence from the junction breakpoint and the two alignments were then combined together manually. Junction sequence that originated from bA346E8 (Xq21.1) is highlighted in red, and nucleotides aligned with this sequence from the genomic sequence around the distal breakpoint are also highlighted in red. Similarly, junction sequence originating from dJ203P18 (Xq27.3) is highlighted in blue, and any bases from the proximal genomic sequence aligned with this sequence are also shown in blue. The 4bp overlap at the junction between the two sequences is shaded purple.

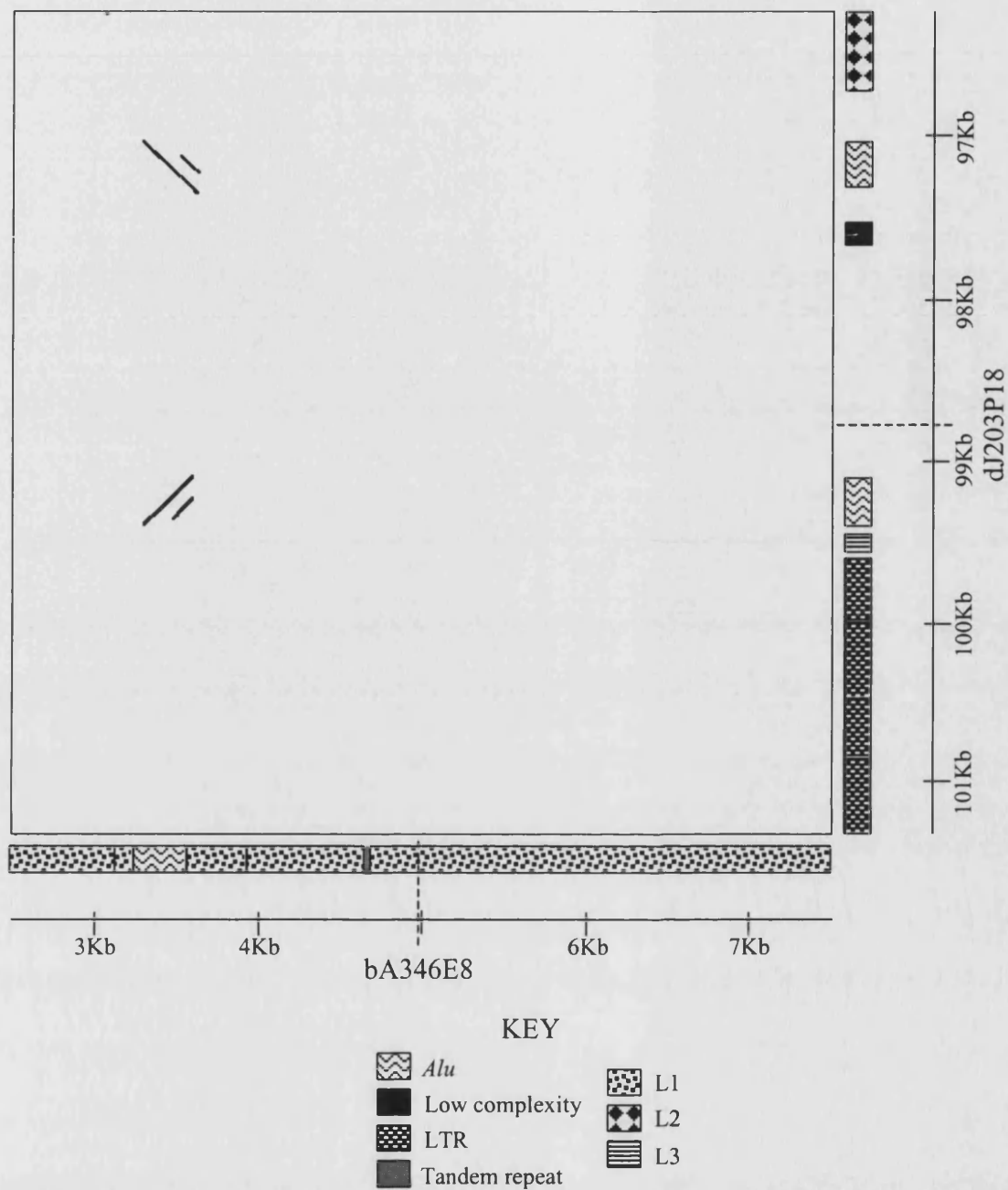


Figure 6.9. Dotplot output from Pipmaker comparing 5kb regions around the proximal and distal deletion breakpoints. Regions of sequence similarity are shown by black lines in the dotplot. Scale bars show the position within the two genomic clones bA346E8 (at the proximal end) and dJ203P18 (at the distal end). The position of the two breakpoints is shown by the dashed lines. Interspersed repeat content of these regions is shown by the shaded and patterned boxes. Similarities between the two sequences are shown as diagonal lines on the dotplot, upwards facing diagonals (/) indicate directly repeated sequences, and inverted repeated sequences are downwards-sloping diagonals (\).

6.2.5.4. Analysis of deletion breakpoint regions for recombination-associated motifs

As for the other sequenced breakpoints in families 1 and 2, various motifs previously found to be associated with different genomic rearrangements were searched for in 5Kb regions around the breakpoints (see section 4.9.1.1. and Table 2.4.).

6.2.5.4.1. 5Kb regions around deletion breakpoints

Within the proximal 5Kb region centred on the deletion breakpoint in bA346E8, none of the sequence motifs looked for were found many more times than would be expected by chance (Appendix C). Two preferred topoisomerase cleavage sites were located at (CAT) or just distal to the proximal deletion breakpoint (CTY) (Appendix C) (Been *et al.*, 1984). Within the 5Kb region surrounding the distal deletion breakpoint in clone dJ203P14 some sequence motifs were found more than five times as would be expected. Two of these were AT rich motifs that were not found close to the breakpoint, the *S. cerevisiae* ARS consensus (WTTTATRTTTW), and a consensus scaffold attachment region sequence (WADAWAYAWW) (Appendix C) (Broach *et al.*, 1983; Maundrell *et al.*, 1988; Dobbs *et al.*, 1994). Two copies of an 8bp motif (CAGRCAGR) which has been found to correspond to a recombination hotspot found in the murine MHC locus were present near the distal deletion breakpoint (Appendix C) (Steinmetz *et al.*, 1986; Shiroishi *et al.*, 1995). One copy of the motif was 371bp proximal to the breakpoint, and the other was just 82bp distal to the breakpoint. Although not found to be over-represented, two short topoisomerase cleavage sites were overlapping the distal deletion breakpoint, (CAT) and (GTY) (Appendix C) (Been *et al.*, 1984).

6.2.5.4.2. Matrix attachment regions

Both 5Kb regions around the two deletion breakpoints were searched for potential MAR regions using MAR-Wiz. The proximal deletion breakpoint region did not contain any regions that reached the threshold matrix-binding potential of 0.6, but the distal deletion breakpoint within dJ203P18 did contain one region that was a potential MAR, starting 700bp downstream of the deletion breakpoint, and spanning 900bp (Figure 6.10.). As the MAR potential predicted using MAR-Wiz can be dependent on the sequence context, a larger (20Kb) region surrounding the breakpoint was also tested for the presence of MARs, and an area of increased MAR potential was observed in the same approximate location (Namciu *et al.*, 2004). This potential MAR included an 111bp stretch of (TA)_n simple repeats, which had been detected by Repeatmasker (Table 6.4.).

6.2.5.4.3. Detailed analysis of 100bp around each deletion breakpoint

6.2.5.4.3.1. Alternating purine/pyrimidines, polypurine and polypyrimidine tracts

One 11bp purine/pyrimidine tract was found at the proximal deletion breakpoint, which included the 4bp overlap at the breakpoint (Figure 6.11.). No homopurine or homopyrimidine tracts of greater than or equal to 10bp were found close to either deletion breakpoint. Alternating purine/pyrimidine tracts can cause DNA to adopt alternative conformations, such as Z-DNA, and have been shown to be over-represented at deletion breakpoints (see section 4.9.2.2.) (Abeyasinghe *et al.*, 2003).

6.2.5.4.3.2. Inverted repeats and secondary structures

Using a program that searches for various types of repeats in DNA sequences - direct, direct complementary, inverted and symmetric repeats (Oligorep), some repeats were found (Figure 6.12.). As before, 100bp centred on both breakpoints was analysed, as

well as the actual recombinant breakpoint sequence (Figure 6.12.). The longest repeat found was an inverted repeat, one half of which included the distal deletion breakpoint and covered 19 nucleotides, with four mismatches (Figure 6.12.). Inverted repeats could potentially form secondary structures such as hairpin loops, which could be cleaved by topoisomerases (Froelich-Ammon *et al.*, 1994). Other types of repeat are also present within the sequences examined, including some inversions of inverted repeats, which could potentially be involved in stabilising the rearrangement (Chuzhanova *et al.*, 2003).

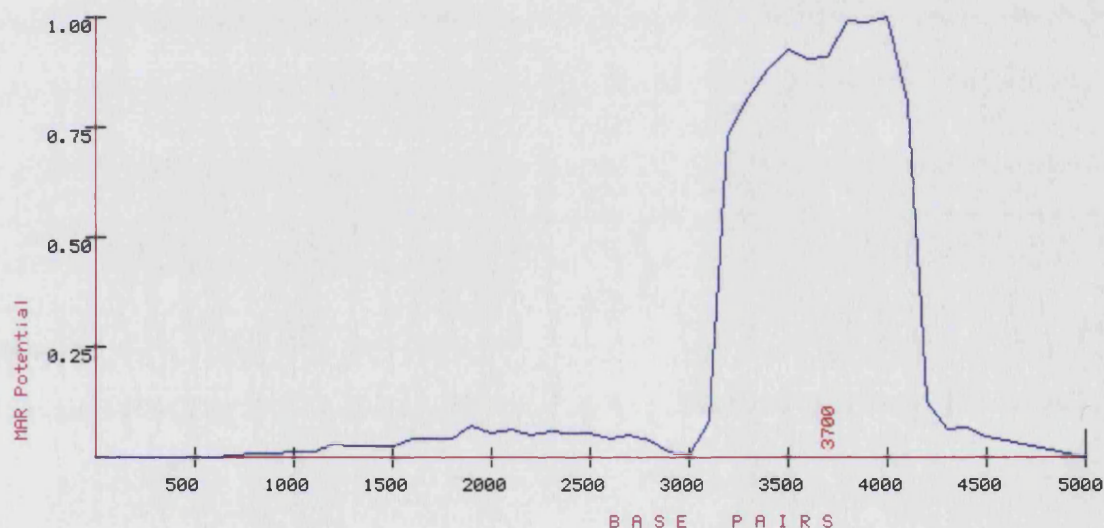


Figure 6.10. MAR potential in the 5Kb surrounding the distal deletion breakpoint (at position 2500 in this sequence) as found by MAR-Wiz.

Proximal deletion breakpoint

bA346E8 4937-5036bp

TGTTTCCATGAATTTATTCATTTCTTCTAGGTTTCCATTTTGTGTGCATACTTGTGTTCTTGGTAGTCTCTGAGGGTTTTAGTATTTCTGTGTGTTCAA
RYRRRRRYRYYYRRRYRRRYRRRRRRRRYYYRRRRRRYRRRRYRYRYRYRYRRRYRYRRRRRYRYYYRRRRRRYYYYYRRRRYYRYRRRRRYRYRYRRRY

Distal deletion breakpoint

dJ203P18 98818-98719bp

CCAGGCCAGACAAATAAGAAAAATGATATTTATGGTCACATTTATGTGACCATAGAAGTGATGTCAGCAAAAAGGAACATAGCTGGATACTAAGTAAAAAA
RYYYRRYYYRYRYYYYYYYYRYRYRRRYRYRRRYRYRRRYRYRYRRRYRYRYRRRYRYRYRRRYRYRYRRRYRYRYRRRYRYRYRRRYRYRYRRRY

Figure 6.11. Purine/pyrimidine content in the region immediately surrounding both deletion breakpoints. The 100bp of sequence flanking the two deletion breakpoints is shown on the top of the pairs, and the classification of each nucleotide is shown on the lower line of the pairs. Tracts of alternating purines/pyrimidines (R/Y), or pyrimidines/purines (Y/R), of 10 nucleotides or greater, are underlined.

Proximal deletion breakpoint bA346E8 4937-5036bp

TGTTTCC**ATGAATTTATTCAT**TTCTTCTAGGTTTTCCATTTTGTGTGCATACTTGTGTTCTTGGTAGTCTCTGAGGGTTTTAGTATTTCTGTGTGTTCAA

Distal deletion breakpoint dJ203P18 98818-98719bp

CCAGGCCAGACAATAAAGAAA**ATGATATTTATGGTCACATTTATGTGACCATAGAAGTGAT**GTGCAGCAAAAAGGAACATAGCTGGATACTAAGTAAAAAA

Deletion junction

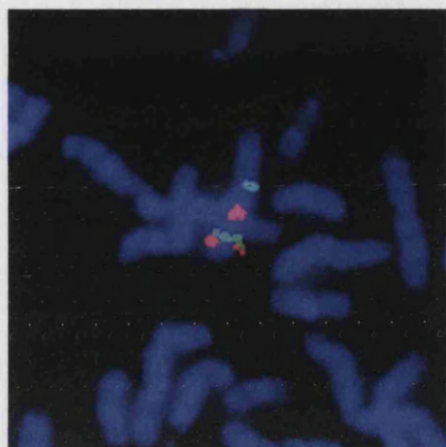
TGTTTCCATGAAT**TTATTCATTTCTT**CTAGGTT**TTCCATTTT**GTGTGCATAGAAGTGATGTCAG**CAAAAAGGAAC**ATAGCTGGAT**ACTAAGTAAAAAA**

Figure 6.12. Inverted repeats around the deletion breakpoints, the deletion junction fragment and the reciprocal rearrangement. The normal sequence is shown from the proximal deletion breakpoint (bA346E8) and from the distal deletion breakpoint (dJ203P18). Also shown is the recombinant deletion breakpoint, with the proximal half (shown in red) originating from bA346E8 and the distal half (shown in blue) originating from dJ203P18. The 4 nucleotides that form the deletion breakpoint are shaded purple. Repeats were searched for using Oligorep (see section 2.2.10.6.).

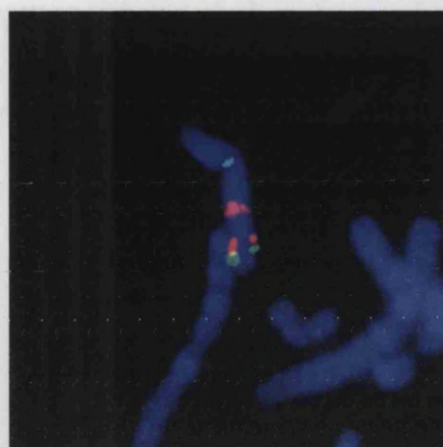
6.3. Xq26 insertion mapping

6.3.1. Metaphase FISH

Mapping of the insertion site of the duplicated copy of *PLP1* was carried out using metaphase FISH with direct dual colour labelling of cosmid, PAC and BAC clones. In experiments containing a Spectrum Red labelled *PLP1* clone and a Spectrum Green labelled Xq clone, the position of the inserted copy of *PLP1* could be deduced with respect to the genomic clone by observing if the green signal was telomeric or centromeric to the duplicated copy of *PLP1* on metaphase chromosomes. This mapped the inserted sequence to between Xq26.2 and Xq27.1 (Table 6.5., Figure 6.13.). A clone positioned in Xq26.3 (dJ119E23) was not able to be mapped with respect to the insertion as on most metaphases examined the red *PLP1* signal and the green signal were co-localised, making it likely that the duplicated copy of *PLP1* was located very near this clone.



(a)



(b)

Figures 6.13a and 6.13b. Metaphase spreads showing dual labelling to map Xq26 *PLP1* insertion. *PLP1* probe is labelled red, X centromere is aqua. Figure 6.13a shows dJ154J13 (Xq26.2) with a green signal and Figure 6.13b has dJ656F14 (Xq27.1) labelled in green.

Clone	Band	above or below distal copy of <i>PLP1</i>
dJ75H8	Xq23	above
dJ378P9	Xq24	above
dJ404F18	Xq24	above
dJ136O17	Xq26.1	above
dJ297J13	Xq26.2	above (5/7)
dJ154J13	Xq26.2	above (7/10)
dJ119E23	Xq26.3	3/8 below, 4/8 were on top, 1/8 above
dJ656F14	Xq27.1	below (8/10)
bA51C14	Xq27.2	below
dJ507I15	Xq27.2	below

Table 6.5. Table showing results from insertion mapping using dual FISH. A clone was recorded as being above the distal *PLP1* signal if the order of signals on the metaphase X chromosome was: centromere probe (aqua)/red/green/red, as shown in Figure 6.13a; and below if the order of signals was: centromere probe (aqua)/red/red/green, as shown in Figure 6.13b. Number of metaphase spreads scored is shown in brackets, with the number of spreads that showed that arrangement indicated. Unless otherwise stated, in the other metaphase spreads examined the green signal could not be resolved from the more distal *PLP1* signal. Where numbers of metaphase spreads scored is not given, only a small number (<5) were scored due to lack of metaphases on the slide or poor hybridisation and faint signals.

6.3.2. Duplication mapping

6.3.2.1. Metaphase FISH

The extent of the duplication in this family was initially characterised using interphase and metaphase FISH with genomic clones known to map to the *PLP1* region. This appeared to show that the duplication breakpoint proximal to *PLP1* was contained within the genomic clone dJ1055C14, which is located 58.5kb upstream relative to *PLP1*. In most metaphase spreads using this clone for FISH, the additional copy of this clone could be seen in Xq26, but the second signal was noticeably weaker than the copy in Xq22, suggesting that only part of the clone was contained in the duplication (Figure 6.14.). In support of this, FISH results showed that the clone immediately proximal to dJ1055C14 was not duplicated, and the next distal clone appeared to be entirely duplicated (Figure 6.14.).

6.3.2.2. UPQFM-PCR and fine mapping of proximal duplication breakpoint

Finer mapping of the proximal duplication breakpoint was carried out using universal primer quantitative multiplex PCR using primers along the clone suggested by FISH to contain the breakpoint, dJ1055C14 (see section 2.2.1.3. and Table 6.6.). This narrowed the region containing the breakpoint down to 1474bp (Table 6.6.).

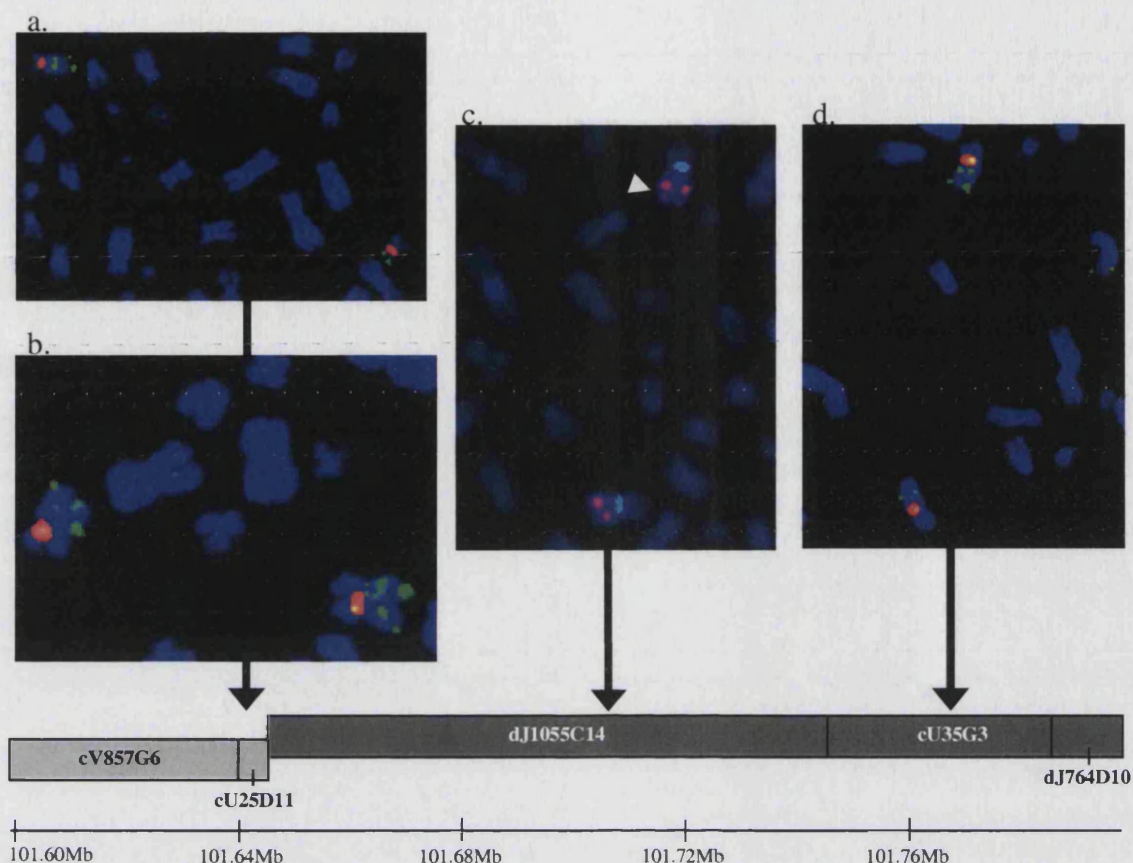


Figure 6.14. FISH on metaphase chromosomes from 3:4 using probes around the proximal duplication breakpoint. (a) cU25D11 shown in green, X centromere is red. cU25D11 maps just proximal to dJ1055C14, but also maps to the Xq telomere, as can be seen on this metaphase spread, where there is a green signal at the end of the deleted Xq arm, and also on the telomere of the normal X chromosome, which also has a signal for this probe at Xq22. (b) From the same hybridisation as (a) above, but with two full sized chromosomes, one of which contains the *PLP1* duplication. Both X chromosomes have cU25D11 signals at Xq22 and near the telomere, with no clear signals showing a duplicated copy at Xq26-27. (c) dJ1055C14 shown in red, X centromere probe shown in aqua. The duplicated copy of dJ1055C14 (white arrow) gives a reduced signal compared to the full-length copy in Xq22. (d) cU35G3 shown in green, X centromere is red. The additional copy of this clone appears to be of equal or greater strength than the signal at Xq22, suggesting that this whole clone is contained within the duplication. Contig data was adapted from the Ensembl human genome browser, release 22.34d.1., and shows a 200Kb region just upstream of *PLP1*.

Position in dJ1055C14 of UPQFM primer pairs	Mean ratio of dJ1055C14 PCR product compared to...		Number of experiments performed
	PLP1	CF	
46009-46237bp	0.59	1.15	4
54449-54682bp	0.62	1.31	3
62946-63105bp	0.55	1.01	2
67104-67331bp	0.55	0.87	2
68598-68849bp	1.12	1.57	2
69896-69955bp	1.16	2.83	1
75773-76033bp	1.01	2.05	3
88473-88702bp	1.01	1.68	2

Table 6.6. UPQFM-PCR results for primers amplifying sequences within dJ1055C14, the genomic clone thought to contain the proximal duplication breakpoint. Dosage quotient ratios were calculated as described (see section 2.2.1.3.1.) and compared to the PLP1 and CF primer pairs. The mean dosage ratios that were within the duplicated range are highlighted in bold type. The presumed location of the breakpoint is indicated by the zigzag line.

6.3.2.3. Inverse PCR and sequencing of insertion breakpoint

6.3.2.3.1. Inverse PCR strategy

Inverse PCR (iPCR) was carried out to obtain the sequence for the proximal duplication and insertion junction using knowledge of the sequences close to the end of the duplication (see section 2.2.1.4.) (Triglia *et al.*, 1988). Primers closest to the proximal duplication breakpoint but still within the duplicated region were chosen from the UPQFM-PCR experiment (Table 6.6.). The primers used for inverse PCR were the reverse complement of the closest primers to the breakpoint used in the UPQFM-PCR experiments (without the universal tag sequences) that were known to be duplicated, which amplified the DNA between positions 69896 and 69955bp in the human genomic clone dJ1055C14 (Figure 6.15.). Two restriction enzymes, *EaeI* and *MspI*, were used because they both had restriction sites that were very near the forward inverse PCR primer (within the known duplicated sequence) and also each had another site within the duplication breakpoint region, as close as possible to the centromeric boundary of the duplicated region. Following digestion of genomic DNA with either enzyme, the DNA was ligated, and the inverse PCR primers were used in a PCR reaction to only amplify the ligated products that contained the target sequences. As the normal sequence in the region was known, the exact size of the inverse PCR product for each enzyme digestion could be predicted, and if there were a breakpoint between the two restriction sites in the genomic DNA, then an altered sized band should be seen as well as the expected size band. The digestion, ligation, and PCR procedure was expected to produce a 1095bp band for *MspI*, and a 1324bp band for *EaeI* digestion. For both restriction enzymes, a smaller band was seen in the carrier mother and the affected boy after agarose gel electrophoresis of the inverse PCR reactions (Figure 6.16.). These bands were gel-extracted and sequenced in a reaction using the iPCR primers (see sections 2.2.1.4.4. and 2.2.4.).



Figure 6.15. Diagram showing strategy for inverse PCR. The dashed line indicates the breakpoint region within dJ1055C14 as defined by UPQFM-PCR results. The positions of amplicons used in UPQFM-PCR, and the duplication status of these regions, are shown underneath. The locations of restriction sites for *EaeI* and *MspI* enzymes are indicated by *E* and *M* respectively.

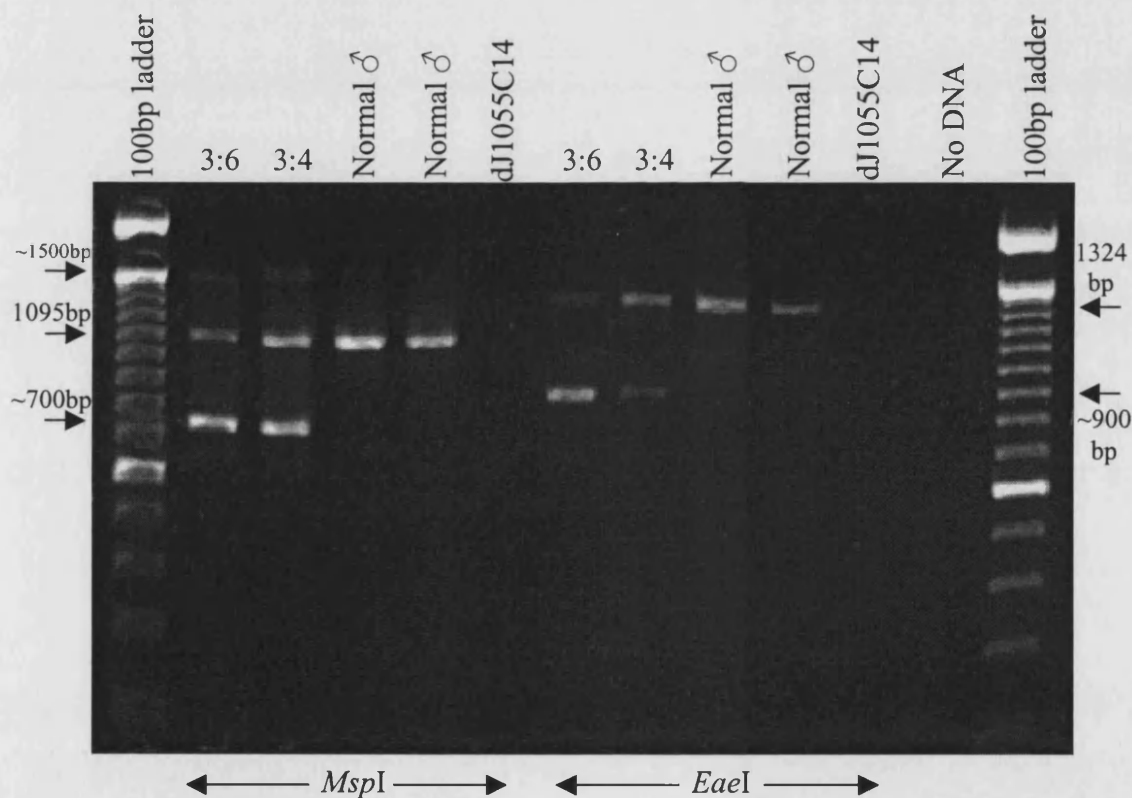


Figure 6.16. Inverse PCR run on agarose gel. Bands of the expected size are seen for the genomic DNAs for both digests (1095bp for *MspI*, 1324bp for *EaeI*), altered bands (~750bp for *MspI*; ~900bp for *EaeI*) are only seen for the mother and son who carry the duplication. Faint larger bands, about 1500bp in size, were also present after inverse PCR using *MspI* digestion, in both 3:4 and 3:6, but not in the normal males. These products may be concatamers formed during the ligation reaction.

6.3.2.3.2. Sequence results for inverse PCR

Sequencing results for the inverse PCR products revealed a complex breakpoint at the insertion point. Over 280bp of readable sequence was obtained from the *MspI* inverse PCR band, which was compared against all human sequences using BLASTn. The results showed that there were four matches to different stretches of human genomic DNA within this sequence, three hits to sequences located in Xq26.2 (dJ453F18, dJ197O17 and dJ305B16), and one match to dJ1055C14, where the primers used in the inverse PCR procedure mapped (Figures 6.15. and 6.17.). The breakpoint region was then also sequenced from genomic DNA from the family, using a PCR product from a reaction using a primer in dJ1055C14 (R68208) and a primer in bA453F18 (R24491) as template. The same sequence as found from the inverse PCR product was produced from the genomic DNA from 3:6 in both directions, showing that the complicated nature of the breakpoint was not just an artifact of the inverse PCR reaction.

6.3.2.3.3. Confirmation of insertion by PCR

For confirmation that there were no further rearrangements in the vicinity of the insertion, oligonucleotide PCR primers were designed, one within dJ1055C14 close to the breakpoint, and two within bA453F18 sequence, distant from the breakpoint (Figure 6.18.). Products of the expected sizes of 887bp and 2264bp were seen following PCR and ethidium bromide stained agarose gel electrophoresis using genomic DNA from individuals in this family carrying the duplication/insertion (Figure 6.18.). This suggested that there were probably no further rearrangements within the next 2.25Kb of bA453F18, although other more distant rearrangements could not be ruled out.

ggtcaggctggtcttgaactccgacctcaggtaatccaccgccttggcctcccaaagggtgagattaaaggcgtgag
 ← 25075-24955 in bA453F18 →
 ccactgctccaggttgctgatacatttgataatacttctt | t | agtagtatggtagtggaagcacagcatgctgttgg
 → ← 69379-69433 in dJ197O17
 gaattagaggagggacagc | ctgttttactcaatgtggaagtactaggagccttcccactaaaatcaagagg | agc | ttg
 → ← 44079-44028 in dJ305B16 → ←
 atggcagaaagggcactacaattagcgaatttatataattataa →
 ◀ 68184-68231 in dJ1055C14 and onwards... →

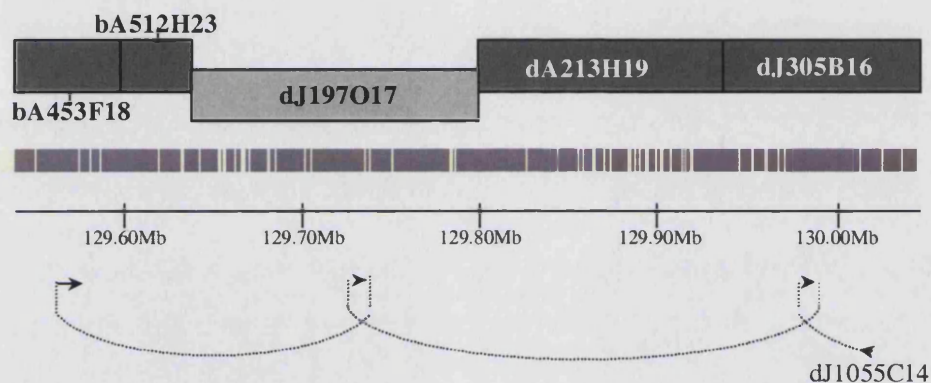


Figure 6.17. The annotated sequence from the *MspI* iPCR band. The arrows and annotation beneath the sequence data show which human genomic clones these sections of the breakpoint have a similarity to as picked up by BLASTn. Underneath is the relevant section of Xq26.2, adapted the Ensembl human genome browser (Release 22.34d.1), showing the clones, position on the chromosome and repeat content of this region. Arrows beneath show the approximate locations of the sequences from the breakpoint junction and the dotted lines indicate how these sections of sequence have joined to the proximal end of the Xq22.2 duplication in dJ1055C14.

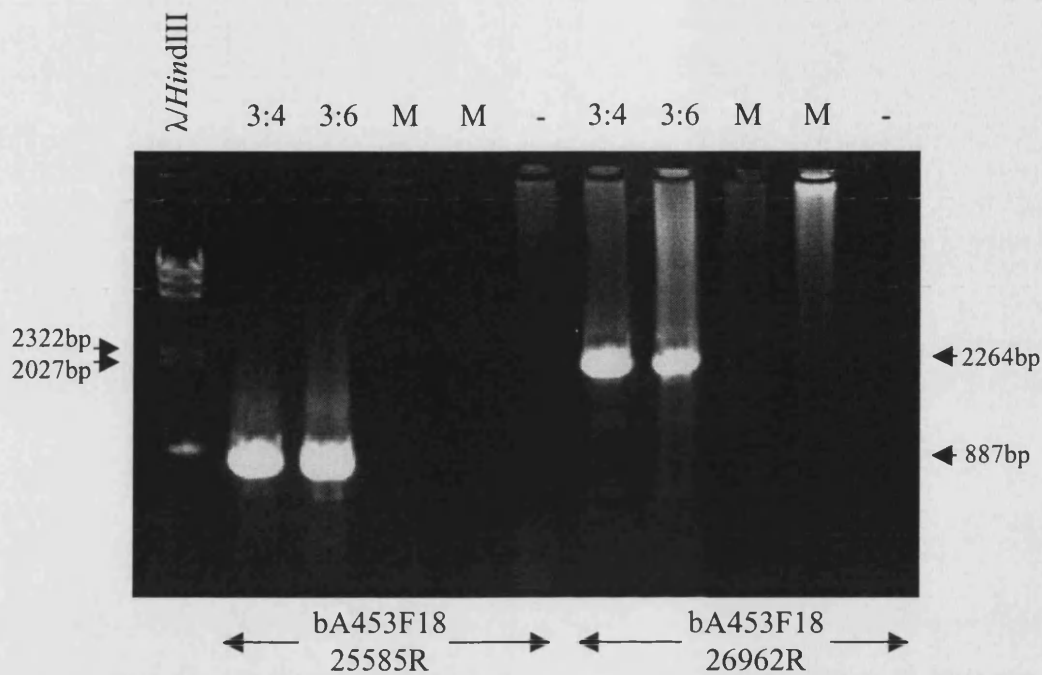


Figure 6.18. Agarose gel showing long-range PCR from dJ1055C14 to bA453F18. Two combinations of primers spanning the breakpoint were used; both reactions used the same primer from dJ1055C14 (68617R) and a different primer from bA453F18 (25585R or 26962R). PCR products of the expected size were only seen in individuals carrying the duplication/insertion and not in normal males (M). Sizes of the products as calculated from sequence data are shown, and the sizes of the closest bands in the *λHindIII* ladder are shown. There is an extra band seen in the size standard lane, this is presumed to be overspill from the adjacent lane in the gel.

6.3.3. Analysis of sequences around the duplication and insertion breakpoints

To determine whether there were any features of the sequences near the breakpoints involved in the Xq26.2 duplication insertion, sequences at both ends of the breakpoints were analysed using a variety of bioinformatic tools.

6.3.3.1. Genomic features near Xq22.2 proximal duplication breakpoint

The duplication/insertion breakpoint within clone dJ1055C14 falls within the first intron of the *MORF4L2* gene (Figure 6.19.). The Xq22 breakpoint was near, but not contained within, interspersed repeat sequences (Figure 6.19.) and was located between a partial L1MC5 repeat and an MIR repeat (Table 6.7.). The G+C content of the 5Kb of sequence around the breakpoint was 41.44% and the interspersed repeat content amounted to 21.52%. The G+C content was close to the average for the 1.1Mb proximal to *PLP1* (40.47%), but the interspersed repeat content was low compared to the rest of the region (55.67%), due to the coding sequence present in this region.

6.3.3.2. Genomic features near Xq26.2 insertion breakpoints

Much of the sequence from Xq26.2 that was found at or near the insertion breakpoints consisted of repetitive elements. The G+C content of a 1Mb region centred on the insertion point and including the two short inserted sequences was 39.08% and the interspersed repeat content was 52.88%.

6.3.3.2.1. Genomic features near dJ305B16 inserted sequence

Most of the 52 nucleotides that mapped to the clone dJ305B16 were part of a partial L1M4 repeat (4034-4079 in the repeat consensus), which maps between positions 43631 and 44073 in this clone (Table 6.8.). The 5Kb examined from this clone contained a

variety of interspersed repetitive elements, making up 64.38% of the sequence, which had a G+C content of 36.90% (Repeatmasker data, Repbase version 7.4) (Figure 6.19. and Table 6.8.).

6.3.3.2.2. Genomic features near dJ197O17 inserted sequence

The 55bp that had homology to dJ197O17 is part of an L2 interspersed repeat (3076-3093 in the L2 repeat consensus), located between bases 69215 and 69757 in this clone (Figure 6.19. and Table 6.9.). This was located in a 5Kb region that was 82.70% interspersed repeats, and had a G+C content of 35.20% (Repeatmasker data, Repbase version 7.4).

6.3.3.2.3. Genomic features near bA453F18 insertion point

The region surrounding the breakpoint in bA453F18 is rich in interspersed repeats (Figure 6.19.). In the 5Kb region around the insertion sequence, 95.86% of the sequence is classified as being interspersed repeats by Repeatmasker (Table 6.10.). The actual insertion breakpoint sequence is within an LTR element (MER51, at position 2330 in the consensus) and is also very close to an almost complete *AluSp* copy, which starts less than 30 bases after the breakpoint (Table 6.10.). This region had a G+C content of 40.76%.

6.3.3.3. Comparison of sequences surrounding breakpoints

6.3.3.3.1. 5Kb surrounding breakpoint sequences

The 5Kb of sequence surrounding each breakpoint from the Xq26.2 insertion were all compared against each other using BLASTz, to determine if there was any large-scale sequence homology present at the breakpoints. Only very limited amounts of sequence similarity was found between two pairs of sequences, and this was all accounted for by

sequences present in common repetitive elements. When 5Kb from bA453F18 and 5Kb surrounding the inserted sequence mapping to clone dJ305B16 were compared, some regions of sequence similarity were seen, all of which were within the *Alu* repeats present in both of these regions (Figure 6.19.). The only other pair of sequences that showed any sequence similarity were dJ1055C14 and dJ305B16, but this similarity only extended for 60bp and was part of an L1 repeat in both sequences.

6.3.3.3.2. 100bp surrounding breakpoints

Alignments were created using ClustalW of the 100bp of genomic sequence immediately flanking a breakpoint or inserted sequence and the sequenced insertion junction (Figure 6.20). Extensive sequence similarities were not present either side of the breakpoints in these sequences. The greatest percentage similarity seen between a junction and flanking sequence was between dJ197O17 and the sequence from bA453F18 from the junction sequence, at 51% of the aligned nucleotides, but most of the aligned nucleotides in this comparison were not contiguous and were not very close to the actual breakpoint (Figure 6.20).

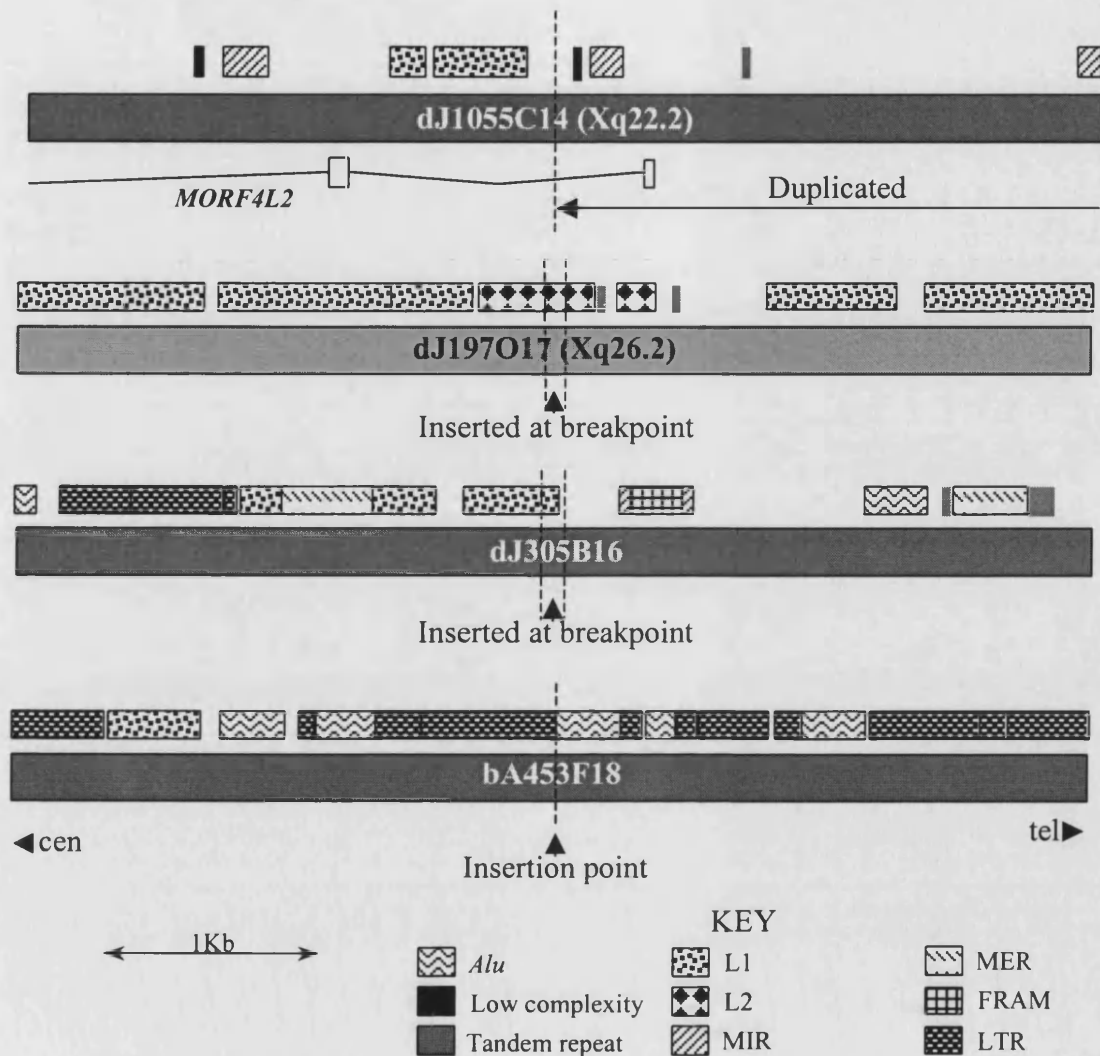


Figure 6.19. Diagram showing interspersed repeat content and genes in 5kb regions centred on the breakpoints associated with the Xq26 insertion. The locations of the breakpoints are shown by the dashed vertical lines. The positions of the first two exons of the *MORF4L2* gene within clone dJ1055C14 are indicated, and the patterned and shaded boxes represent repetitive sequences. All regions shown are in the same orientation, with more centromeric sequence at the left-hand side of the figure.

Distance from breakpoint (bp)	Repeat type	Position in repeat consensus sequence	Orientation
(-) 1727-1705	AT rich (Low complexity)	1-23	+
(-) 1597-1386	MIR (SINE/MIR)	2-241 (240/262)	-
(-) 825-654	L1MC5 (LINE/L1)	7783-7949 (166/7961)	-
(-) 602-187	L1MC5 (LINE/L1)	7527-7915 (388/7961)	-
(+) 30-52	AT rich (Low complexity)	1-23	+
(+) 108-268	MIR (SINE/MIR)	48-227 (179/262)	-
(+) 814-846	T _n (Low complexity)	1-33	+
(+) 2383-2486	MIR (SINE/MIR)	30-136 (107/208)	+

Table 6.7. Repeat content (using Repeatmasker) of the 5Kb genomic region surrounding the proximal duplication breakpoint within dJ1055C14. The first column shows how far in base pairs each repeat element is from the duplication breakpoint, (-) indicates that the repeat element is proximal to the breakpoint and (+) that the repeat is distal relative to the breakpoint. The type of repeat and the class of repeat element to which it belongs are given in the second column. The third column shows which portions within the appropriate repeat consensus sequence each repeat has similarity to, and also how many bases out of the total repeat unit are present in the sequence. In the fourth column, the orientation of each repeat is shown; + shows a repeat is on the forward strand (i.e. running from centromere to telomere), - the repeat is on the reverse strand.

Distance from breakpoint (bp)	Repeat type	Position in repeat consensus sequence	Orientation
(-) 2689-2384	<i>AluSg</i> (SINE/ <i>Alu</i>)	1-306 (306/306)	-
(-) 2268-1920	MLT1B (LTR/MaLR)	26-370 (345/390)	+
(-) 1919-1472	MSTB2 (LTR/MaLR)	2-457 (457/457)	+
(-) 1471-1451	MLT1B (LTR/MaLR)	370-390 (22/390)	+
(-) 1434-1239	L1MC5 (LINE/L1)	7617-7820 (204/7961)	-
(-) 1238-820	Tigger2a (DNA/MER2 type)	1-434 (434/434)	+
(-) 819-532	L1MC5 (LINE/L1)	7356-7617 (262/7961)	-
(-) 397-45 (+)	L1M4 (LINE/L1)	4034-4475 (442/6146)	-
(+) 282-356	MIR (SINE/MIR)	74-149 (76/208)	+
(+) 360-517	FRAM (SINE/ <i>Alu</i>)	1-155 (156/176)	-
(+) 525-618	MIR (SINE/MIR)	159-252 (94/262)	+
(+) 1404-1695	<i>AluJb</i> (SINE/ <i>Alu</i>)	9-298 (290/312)	+
(+) 1781-1804	(TA) _n (Simple repeat)	2-25	+
(+) 1816-2160	MER7A (DNA/MER2 type)	1-344 (344/346)	+
(+) 2163-2294	(TA) _n (Simple repeat)	2-138	+

Table 6.8. Interspersed repetitive elements from the 5Kb region surrounding the duplicated bases in the rearrangement from dJ305B16. For detailed explanation of categories, see Table 6.9. The L1M4 repeat that partially overlaps with the 52bp sequence of this genomic clone that is found within the rearrangement at the insertion breakpoint is highlighted in bold type, the L1M4 consensus sequence homology extends 45bp into the duplicated sequence.

Distance from breakpoint (bp)	Repeat type	Position in repeat consensus sequence	Orientation
(-) 2711-1602	L1MD5 (LINE/L1)	3174-4284 (1111/6146)	+
(-) 1531-706	L1MD5 (LINE/L1)	4394-5248 (855/6146)	+
(-) 704-352	L1MD5 (LINE/L1)	5882-6242 (6242/6242)	+
(-) 324 – 164 (+)	L2 (LINE/L2)	2752-3257 (506/3272)	+
(+) 169-200	(TGAA) _n (Simple repeat)	1-32	+
(+) 256-450	L2 (LINE/L2)	2433-2658 (226/3314)	+
(+) 530-561	(CA) _n (Simple repeat)	2-33	+
(+) 960-1557	L1M3e (LINE/L1)	1-1948 (1948/7030)	+
(+) 1690-3086	L1MA9 (LINE/L1)	4873-6312 (1440/6312)	+

Table 6.9. Interspersed repetitive elements from the 5Kb region surrounding the duplicated bases in the rearrangement from dJ197O17. The first column shows how far in base pairs each repeat element is from the duplication breakpoint, (-) indicates that the repeat element is proximal to the breakpoint and (+) that the repeat is distal relative to the breakpoint. The type of repeat and the class of repeat element to which it belongs are given in the second column. The third column shows which portions within the appropriate repeat consensus sequence each repeat has similarity to, and also how many bases out of the total repeat unit consensus is present in the sequence. In the fourth column, the orientation of each repeat is shown; + shows a repeat is on the forward strand (*i.e.* running from centromere to telomere), - the repeat is on the reverse strand. If a repeat element was found to coincide with the beginning or end of the 5kb segment, the adjacent sequence was also analysed with Repeatmasker, until the true end of the repeat was found. The repeat element that the duplicated breakpoint segment is contained within is highlighted in bold type. The figures given in the first column for the distance from the breakpoint for this repeat are for the distance from the closest end of the 52bp breakpoint segment to the end of the repeat, *i.e.* the (-) and (+) distances are measured from different nucleotides.

Distance from breakpoint (bp)	Repeat type	Position in repeat consensus sequence	Orientation
(-) 2691-2068	MER51-int (LTR/ERV1)	512-1221 (709/7816)	+
(-) 2041-1617	L1M2 (LINE/L1)	1578-2012 (435/6849)	+
(-) 1527-1230	<i>AluSg1</i> (SINE/ <i>Alu</i>)	1-296 (297/309)	-
(-) 1162-1082	MER51-int (LTR/ERV1)	1038-1121 (84/8120)	+
(-) 1081-798	<i>AluSq</i> (SINE/ <i>Alu</i>)	40-329 (290/329)	-
(-) 713-542	MER4A (LTR/ERV1)	276-517 (242/664)	+
(-) 539-26 (+)	MER51-int (LTR/ERV1)	1791-2357 (567/7816)	+
(+) 27-328	<i>AluSp</i> (SINE/ <i>Alu</i>)	3-305 (303/313)	+
(+) 329-432	MER51-int (LTR/ERV1)	2357-2460 (104/7816)	+
(+) 445-588	<i>AluSp</i> (SINE/ <i>Alu</i>)	6-151 (146/313)	-
(+) 590-707	MER51-int (LTR/ERV1)	2466-2589 (124/7816)	+
(+) 701-1018	MER51-int (LTR/ERV1)	2923-3265 (343/7816)	+
(+) 1037-1170	MER51-int (LTR/ERV1)	1660-1783 (124/6741)	+
(+) 1171-1464	<i>AluSq</i> (SINE/ <i>Alu</i>)	6-299 (294/313)	+
(+) 1465-1996	MER51-int (LTR/ERV1)	1783-3823 (2041/6741)	+
(+) 1994-2100	MER51-int (LTR/ERV1)	2340-2446 (107/5217)	+
(+) 2110-2641	LTR10A (LTR/ERV1)	1-547 (548/609)	+

Table 6.10. Interspersed repetitive elements from the 5Kb region centred on position 24956 in clone bA453F18. For detailed explanation of categories, see Table 6.9.

Junction sequence **AGGCGTGAGCCACTGCTCCAGGTTGCTGATACATTTGTATAATACTTCTT-T-AGT**

bA453F18 25004 **AGGCGTGAGCCACTGCTCCAGGTTGCTGATACATTTGTATAATACTTCTTAT-TTT**

dJ197O17 69329 **AAGACTGAGCC-TGAGTC**TGACAA**GT****TGAT--ATTTACAAAAGAGTACAG-TGAGT**

Junction sequence **AGTATGGTAGTGGCAAGCACAGCATGCTGTTGGGAATTAGAGGAGGGACAGCCTGT**

bA453F18 **ACTTTAATCAAGACTAAGAACTTTAACTATGAGAA**TGTT**CAATTAG** 24905

dJ197O17 **AGTATGGTAGTGGCAAGCACAGCATGCTGTTGGGAATTAGAGGAGGGACAGC**TAAC

dJ305B16 43978 **TGCC**TAATCAAGTTT**AAATGTGC**TTTT**AAAAAGTTTCATTCAAAAGCCA--CTGT**

Junction sequence **TTTACTC---AATGTGG-AAGTACTAGGAGCCTTCCCACTAAAATCAAG-AGGAG**

dJ197O17 CCAGATTTGGG**ATGTGTTGAGTA**TTAA**AGCTCCCAACAAGAGTG** 69483

dJ305B16 **TTTACTC---AATGTGG-AAGTACTAGGAGCCTTCCCACTAAAATCAAG-AGGGA**

dJ1055C14 68134 ---GG**TA---AAAATTG**-GTAT**ATCTATTTTATCTTTAATTTATCTAGGAAAGAA**

CTTGATGGCAGAAAGGGCACT-ACAATTAGCGAATTTATATAATTTATAATT Junction sequence

-----**AAAAAGGATGCTCAC**TC**TTGCCACTATTATGTAACAATGCATTGAAGG** 44029 dJ305B16

GTTGATGGCAGAAAGGGCACT-ACAATTAGCGAATTTATATAATTTATAATT 68233 dJ1055C14

Figure 6.20. Alignment of insertion breakpoint junction and 50bp either side of all the breakpoints or short inserted sequences. Legend on next page.

Figure 6.20. Legend.

Each breakpoint region was individually aligned by ClustalW to the sequence from the junction breakpoint and the two alignments were then combined together manually. Junction sequence that originated from bA435F18 (Xq26.2) is highlighted in red, and nucleotides aligned with this sequence from the genomic sequence around the other breakpoints are also highlighted in red. Similarly, junction sequence originating from dJ197O17 (Xq26.2) is highlighted in blue, dJ305B14 (Xq26.2) sequence is shaded purple, and dJ1055C14 (Xq22.2) sequence is shaded green. The 4bp overlap at the junction between the two sequences is shaded purple. Positions within the clone are shown at the start and end of each sequence. Nucleotides that appear to have been inserted at the junctions are coloured black and are in bold type.

6.3.3.4. Recombination/rearrangement associated sequence motifs

6.3.3.4.1. 5Kb around dJ1055C14 duplication breakpoint

Three of the different motifs that were found more often than the expectation within the 5Kb region centred on the duplication breakpoint in dJ1055C14 were all associated with eukaryotic origins of replication. All three motifs (the human replication origin consensus (WAWTTDDWWWDHWGWHMAWTT); the *S. cerevisiae* ARS motif (WTTTATRTTTW); and the *S. pombe* ARS sequence (WRTTTATTTAW)) were each only found once within the 5Kb region (Appendix C) (Maundrell *et al.*, 1988; Dobbs *et al.*, 1994; Dechering *et al.*, 1998). This human replication origin motif, out of the three associated with replication origins, was the closest to the duplication breakpoint in family 3, at 844bp proximal to the breakpoint (Appendix C).

Other sequence motifs located in the 5Kb around the duplication breakpoint in dJ1055C14 and found at a greater frequency than expected included 7 copies of a scaffold attachment consensus (TTWTTWTTWTT) and 7 copies of an 8bp motif (GCCCWCCW), which is recognised and bound to by the translin protein and has been found at the breakpoints of chromosomal translocations in human lymphoid malignancies (Appendix C) (Dobbs *et al.*, 1994; Aoki *et al.*, 1995).

6.3.3.4.2. 5Kb around dJ305B16 inserted sequence

Several different recombination or rearrangement-associated motifs were found at a higher frequency than was expected in the 5Kb region around the 52bp inserted at the junction originating from clone dJ305B16.

A match to the consensus binding site (PUR) of the Pur α DNA binding protein (GGNNGAGGGAGARRRR) was located 1130bp proximal to the start of the 52bp sequence inserted at the breakpoint. The Pur α protein was originally isolated as a factor that binds to the purine-rich strand of the consensus sequences found upstream of the human c-myc gene (Bergemann *et al.*, 1992; Bergemann and Johnson, 1992). PUR sequences are also present in several promoter sequences and Pur α binding sites have also been found at origins of replication in both mammalian genomic DNA and viral genomes (Bergemann and Johnson, 1992; Smith *et al.*, 1998, reviewed in Gallia *et al.*, 2000). This motif is predicted to occur by chance only rarely, approximately once every 29Mb in sequences with the same nucleotide composition as this 5Kb stretch from dJ305B16. There are no genes present in the near vicinity of this sequence element; the nearest known transcript (annotated on Ensembl as NM_194277) is 14Kb away (Figure 6.25.). Although the lack

of proximity to gene and promoter sequences makes it less likely that this sequence has a function in the regulation of gene transcription in this region, it is also possible that this sequence may be involved in DNA replication in this region. Pura is able to destabilise the DNA double helix and was originally isolated as a factor binding to regions of initiation of DNA replication (Bergemann and Johnson, 1992; Darbinian *et al.*, 2001). The presence of this PUR sequence may contribute to initiation of replication, a possibility that is increased by its localisation within the first of two putative matrix attachment regions as predicted by the program MAR-Wiz (see sections 4.9.1.2., 6.3.3.5., Figure 6.21. and Table 6.11.) (Singh *et al.*, 1997). Some origins of replication have been mapped to the same regions as matrix or scaffold attachment sites (Razin *et al.*, 1991; Lagarkova *et al.*, 1998; Girard-Reydet *et al.*, 2004). As previously mentioned, rearrangements involving the *PLP1* gene could be connected to the process of replication (see section 5.8.2.5. and 6.4.2.). Some other sequence motifs were found to be over-represented within the 5Kb region centred on the inserted sequence from dJ305B16, but none of these were located in particular proximity to the breakpoints (Appendix C).

6.3.3.4.3. 5Kb around dJ197O17 inserted sequence

Just one sequence motif was found at a higher frequency than expected in the 5kb region centred on the middle of the 53bp of sequence inserted into the Xq26.2 insertion breakpoint from this clone (also from Xq26.2), a 6bp sequence (ACCCCA) that is a DNA polymerase frameshift mutation hotspot (Kunkel, 1985a; Kunkel, 1985b). This motif occurred 6 times in the 5Kb examined, compared to the expected frequency of just over 1 copy of this motif (Appendix C). The nearest copy of this sequence was 150bp distal from the inserted sequence and all the other copies were over 1Kb distal to the inserted sequence.

6.3.3.4.4. 5Kb around bA453F18 insertion point

Several different sequence motifs were searched for in the 5Kb surrounding the sequenced insertion point within clone bA453F18 (see Table 2.4.). Some motifs that were over-represented in the region were the fission yeast ARS consensus (WRTTATTTAW) and two scaffold attachment motif consensus sequences, (AATAAAYAAA, TTWTWTTWTT) (Maundrell *et al.*, 1988; Dobbs *et al.*, 1994). The closest of these three motifs to the breakpoint was one of the scaffold attachment motifs, 318bp distal to the insertion point.

6.3.3.5. Matrix attachment regions

Potential MARs were found near all of the 5Kb regions examined around the breakpoints (see section 4.9.1.2., Figure 6.21. and Table 6.11.). As the MAR potential predicted using MAR-Wiz can be dependent on the sequence context, a larger (20Kb) region surrounding the breakpoint was also tested for the presence of MARs, and areas of increased MAR potential were again observed in the same approximate locations (Namciu *et al.*, 2004).

6.3.3.6. *In silico* analysis of sequence from 50bp regions either side of breakpoints

6.3.3.6.1. Purine/pyrimidine content

The 100bp regions around the duplication and insertion breakpoints, and the inserted sequences were examined for purine and pyrimidine content (see section 4.9.2.2.). There were polypurine tracts (10bp or greater in length) found at or very close to three breakpoints, but no long tracts were found near the duplication breakpoint in dJ1055C14 (Figure 6.22.). The longest polypurine tract was 16bp in length, and overlapped with one

of the ends of the inserted sequence from dJ305B16 (Figure 6.22.). Purine tracts have been found to be over-represented at both translocation and deletion breakpoints and can form unusual DNA structures and may be able to stimulate homologous recombination (see section 5.8.2.4.) (Rooney and Moore, 1995).

6.3.3.6.2. Inverted repeats and secondary structure near breakpoints

The sequences around the breakpoints of the duplication/insertion and the sequenced junction were all searched for the presence of direct, inverted, symmetric and inversions of inverted repeats by the Oligorep program (see section 4.9.2.3.). Only a few repeated sequences were found in the areas searched (Figures 6.23. and 6.24.). Generally a few more repeats were seen in the junction than the original sequences, which may have been involved in mediating or stabilising the rearrangement, or this apparent increase may just have been related to the fact that a longer sequence was being examined (Figures 6.23. and 6.24.) (Chuzhanova *et al.*, 2003).

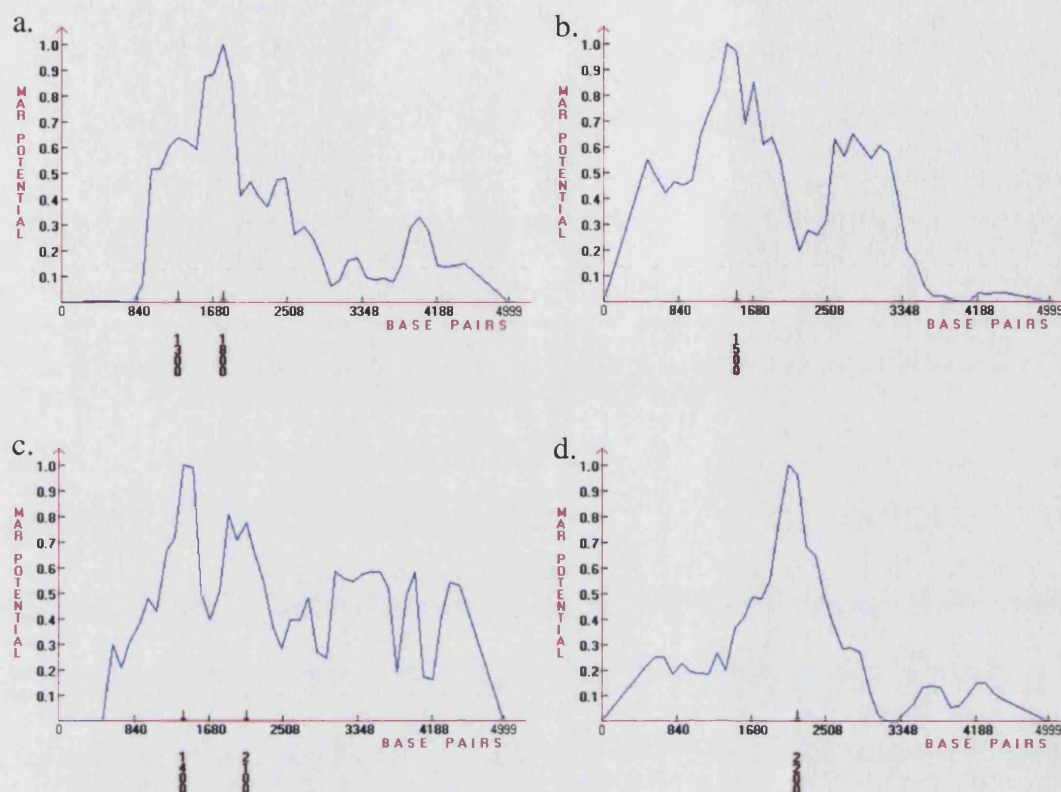


Figure 6.21. Regions of MAR potential in the regions around the duplication and insertion breakpoints in family 3 according to the MAR-Wiz program. (a.) shows 5Kb around the insertion point in bA453F18, (b.) is the 5Kb of sequence surrounding the short insertion from dJ197O17, (c.) shows the results from the 5Kb of sequence around the short inserted sequence from dJ305B16, and (d.) shows the MAR-potential plot for 5Kb of sequence surrounding the proximal duplication breakpoint in family 3 in clone dJ1055C14. All breakpoints or inserted sequences are located at (or centred on) position 2500bp on the plots.

Clone	Region within clone	Location of potential MAR	Average strength	Integrated strength
bA453F18	22455-27454	24055-24155	0.878	264.211
dJ197O17	71905-66906	70705-70205	0.760	380.707
dJ305B16	41553-46552	42853-43053	0.808	162.396
		43453-43653	0.655	131.684
dJ1055C14	65684-70683	67684-68084	0.793	318.085

Table 6.11. Regions of relatively high MAR potential near the sequences involved in the Xq26.2 insertion, as found by MAR-Wiz. Positions are given in base pairs from the start of the sequence of the relevant genomic clone. Average strength is defined by the MAR-Wiz program as the potential per 1Kb sequence window, and is worked out from all the contiguous windows that have a potential higher than the threshold normalised MAR-potential value of 0.6. Integrated strength is the total MAR potential over the whole length of the putative MAR.

dJ197O17 69329-69483

AAGACTGAGCCTGAGTCTGACAAGTTGATATTTACAAAAGAGTACAGT**GAGTAGTATGGTAGTGGCAAGCACAGCAT****TGCTGTTGGGAATTAGAGGAGGGACAG**
 CTAACCCAGATTTGGGATGTGTTGAGTAATTAAAA**AGCTCCCAACAAGAGTG**

dJ305B16 43978-44029

TGCCTAATCAAGTTTAAATGTGCTTTTTTAAAAAGTTTCATTCAAAAAGCCACTGTTTTACTCAATGTGGAAGTACTAGGAGC**CTTCCCACTAAAAT**CAAGAGGG
 AAAAAAGGATGCTCACT**CTTCCCACTATAT**GTAAACAATGCATTGAAGG

dJ1055C14 68134-68233

GGTAAAAATTGGTATATCTAT**TTTATCTTTAATTATTCTAG**GAAAGAAGTTGATGGCAGAAAGGGCACTACAATTAGC**GAATTATATAATTTATAATT**

Figure 6.23. Internally repeated sequences around the breakpoints in family 3, looking at the genomic sequence around the breakpoints or inserted sequences. 50bp on either side of a breakpoint is shown. No repeats were found in the region around the bA453F18 insertion point. Sequences found at the breakpoint are coloured as before in Figure 6.17., and the sequences not found in the final rearrangement are shaded grey. Nucleotides that form part of repeat sequences are shown in bold. The type of repeat found is shown by the arrows above and below the sequence: direct (→), inverted (←) and symmetric (↔). Repeats within the sequence were searched for using Oligorep (see section 2.2.2.10.6. and 4.9.2.3.).

Junction sequence



Figure 6.24. Internally repeated sequences around the sequenced breakpoint junctions in family 3. 50bp on either side of a breakpoint is shown. Sequences found at the breakpoint are coloured as before in Figure 6.17., and the sequences not found in the final rearrangement are shaded grey. Nucleotides that form part of repeat sequences are shown in bold. The type of repeat found is shown by the arrows above and below the sequence: direct (→), inverted (←) and symmetric (↔). Repeats within the sequence were searched for using Oligorep (see sections 2.2.10.6. and 4.9.2.3.).

6.3.3.7. Analysis of genomic features in the wider insertion region

6.3.3.7.1. Genes

The gene content of a 1Mb region including the insertion point in bA453F18 and the two short inserted sequences from dJ197O17 and dJ305B16 was examined (Figure 6.25.). There were five known genes in this region, and two novel genes (as annotated on the Ensembl genome browser, version 22.34d.1). One gene overlapped with the proximal end of this region, and coded for a member of the olfactory receptor gene family, *OR13H1* (Fuchs *et al.*, 2002; Malnic *et al.*, 2004). There are hundreds of copies of functional olfactory receptor genes and pseudogenes throughout the human genome, but this is the only one so far that has been mapped to the X chromosome (Malnic *et al.*, 2004). Transcripts corresponding to *OR13H1* have been submitted to Genbank under accession numbers Q8NG92 and Q96R21.

There were two genes in the region between the inserted sequences from dJ197O17 and dJ305B16 (Figure 6.25.). One, *MST4*, codes for a protein that is a member of the Ste20-like kinase family and contains an N-terminal kinase domain and a C-terminal regulatory domain (Qian *et al.*, 2001). There are two alternatively spliced isoforms of this gene, which are in the databases as accession numbers Q9BXC3 and Q8NBY1. The other gene, annotated as transcript *NM_194277* on Ensembl, was less well characterised but contained a band 4.1/FERM domain at in its N-terminal region (Chishti *et al.*, 1998). FERM domains are often involved in localising the protein to the plasma membrane (Chishti *et al.*, 1998).

A known gene was located at the distal end of clone dJ305B16 (Figure 6.25.). This gene codes for a member of the Ras-like family of small GTPases, RAP2C. Members of this

large protein family have diverse roles in signal transduction (Stork, 2003). One further annotated gene was just within the distal end of this 1Mb region (Figure 6.25.). *MBNL3* (Muscleblind-like protein 3) is involved in the regulation of myogenesis and may be involved in the pathogenesis of myotonic dystrophy (Squillace *et al.*, 2002; Fardaei *et al.*, 2002). The MBNL3 protein contains 4 Cys3His zinc finger domains, which may be involved in RNA binding (Squillace *et al.*, 2002).

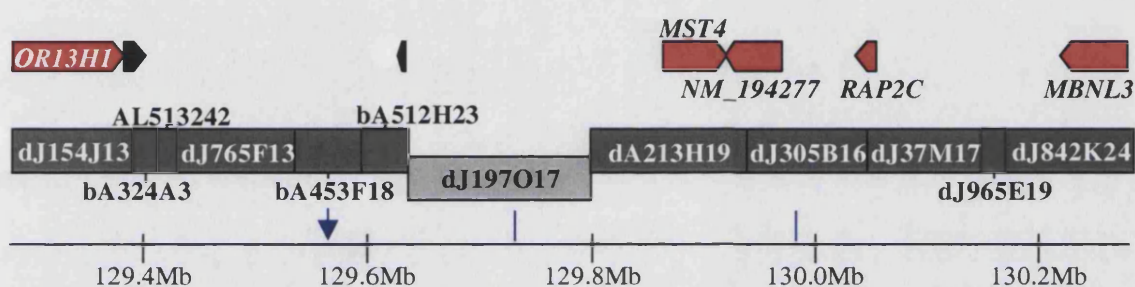


Figure 6.25. Genes present in the insertion region, adapted from the Ensembl genome browser. Genomic clones from a 1Mb region including the insertion sequence are shown as grey boxes, dark grey when sequenced on the forward strand, light grey when sequenced on the reverse strand. Distance from the Xp telomere is shown in Mb underneath the contig. Genes in the region are shown as arrows above the contig, the direction of transcription is the same as the orientation of the arrow. Known genes are shaded red and labelled, novel uncharacterised genes are shaded black. The position of the insertion of Xq22.2 sequence into Xq26.2 is shown by the blue arrow above the scale bar and the two original locations of the two ~50bp sequences inserted at the breakpoint are shown by the blue bars underneath dJ197O17 and dJ305B16.

6.3.3.7.2. Region – specific repeats

1Mb of genomic sequence containing the insertion site was compared against itself using Pipmaker on sequence that had been masked for interspersed repeats by Repeatmasker (Figure 6.26.). This revealed that part of this region had a complex repetitive nature, with many repeats, both direct and inverted, and that the insertion of Xq22.2 sequence had occurred within this highly complex repetitive region (Figures 6.26. and 6.27.). Some of this repetitive region was contained within bA453F18, and part of the distal section is in another genomic clone involved in the breakpoint, dJ197O17 (Figure 6.26.). The rest of the repeats were within the clone that separated bA453F18 and dJ197O17, bA512H23 (Figure 6.26.). Another Pipmaker dotplot, of just bA453F18 (masked for interspersed repeat sequences) sequence against itself, to give a more detailed view of the local repeats in this region, showed that this clone contained six copies of a small inverted repeat, and also had large stretches of direct repeats (including the inverted repeat sequence) in the distal half of the clone (Figure 6.27.). Comparison with the track showing the location of interspersed repeat elements in this genomic clone, using the Ensembl genome browser, shows that almost all of the sequence in this clone is repetitive, either as interspersed repeats, or the region-specific repeats (Figure 6.28.). This highly repetitive region appears to be evolutionarily conserved, as similar patterns of repeated sequences are seen when either the syntenic regions of the *Mus musculus* or *Pan troglodytes* X chromosome are compared by BLASTz in a dotplot (Figure 6.28.). The human and chimpanzee repeats are similar in nature as when the syntenic regions from the two genome were compared the similarities between the two were very much like the pattern seen in Figure 6.28a (data not shown). However, when the mouse sequence was compared against either the human or chimpanzee sequence, hardly any similarities between these sequences were found by BLASTz comparisons (data not shown). The

syntenic region of the rat X chromosome also contains a region including arrays of a short, directly repetitive sequence, revealed by BLASTz comparisons, with a very similar organisation to the repeats in the mouse (not shown). It seems that the repetitive nature of this genomic region is conserved between the three species, but that the actual sequence within the repeats may not be so well conserved. One obvious difference between the mouse and the two primate sequences is that the murine repeats appear to be purely direct repeats, whereas the human and chimpanzee repeats also contain some inverted repeat sequences (Figure 6.27., and 6.28.). Although the human and mouse sequences for this 200Kb region are contiguous, it is important to note that the chimpanzee genome sequence is not yet complete, and there are several gaps in the sequence in this region, so these regional repeats may be more extensive in this genome sequence, and also be present in the as yet unsequenced gaps (Figure 6.28a.). Interestingly, when this region is viewed on the Ensembl genome browser, it appears that some open reading frames predicted by Genscan, and an EST, coincide with many of the inverted repeats within bA453F18 (Figure 6.27.).

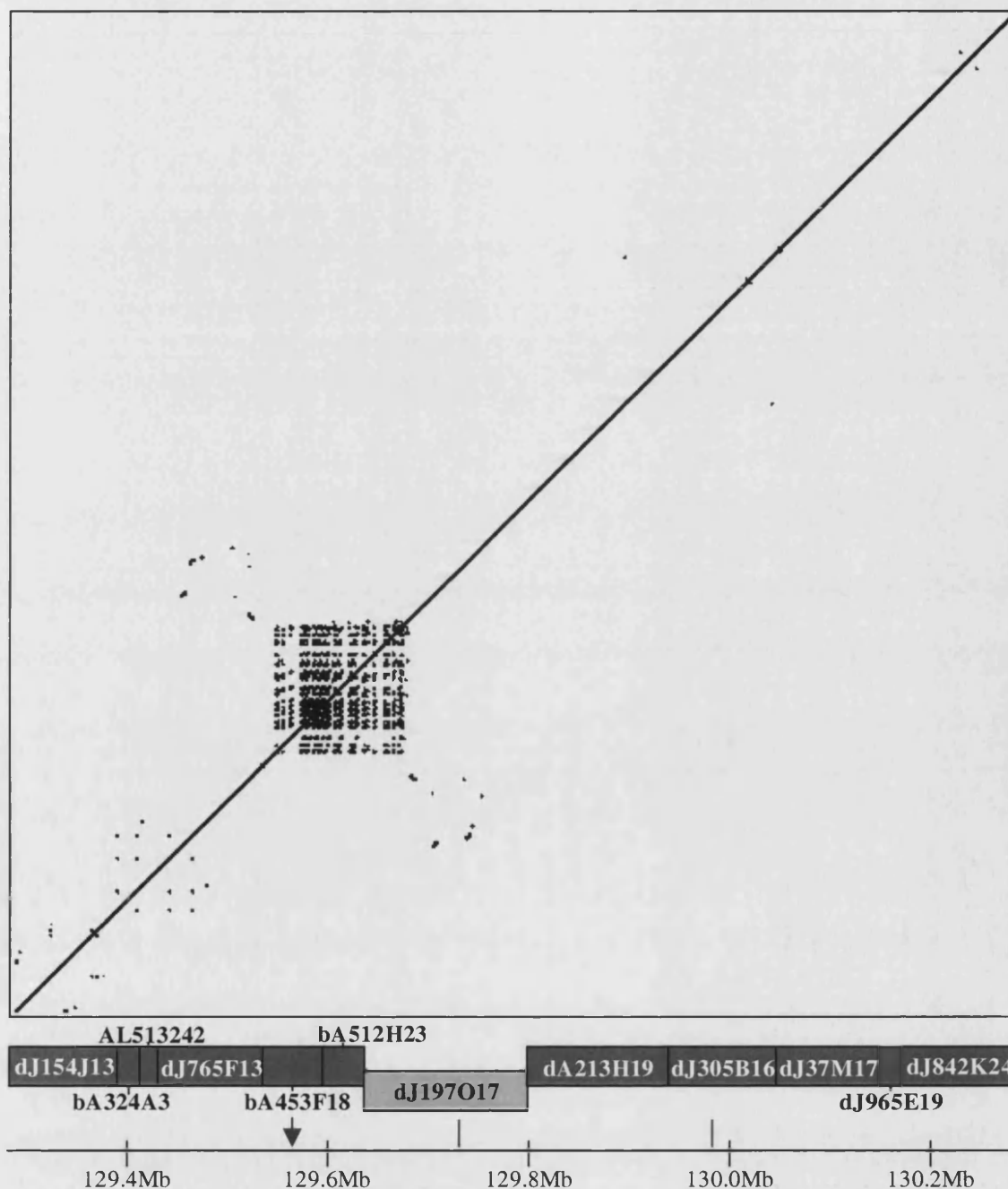


Figure 6.26. Dotplot (Pipmaker) of the region around the Xq26.2 insertion point. The contig for the region is shown underneath the Pipmaker output. The scale bar shows distances in Mb from the Xp telomere (NCBI Build 34). The positions of human genomic clones for this region are shown by the labelled grey boxes underneath the dotplot. 1Mb genomic sequence, including the insertion point (blue arrow) for the Xq22.2 duplication was used, masked for human repetitive elements before being compared against itself in the dotplot. The blue lines show the original locations of the two short sequences inserted at the breakpoint.

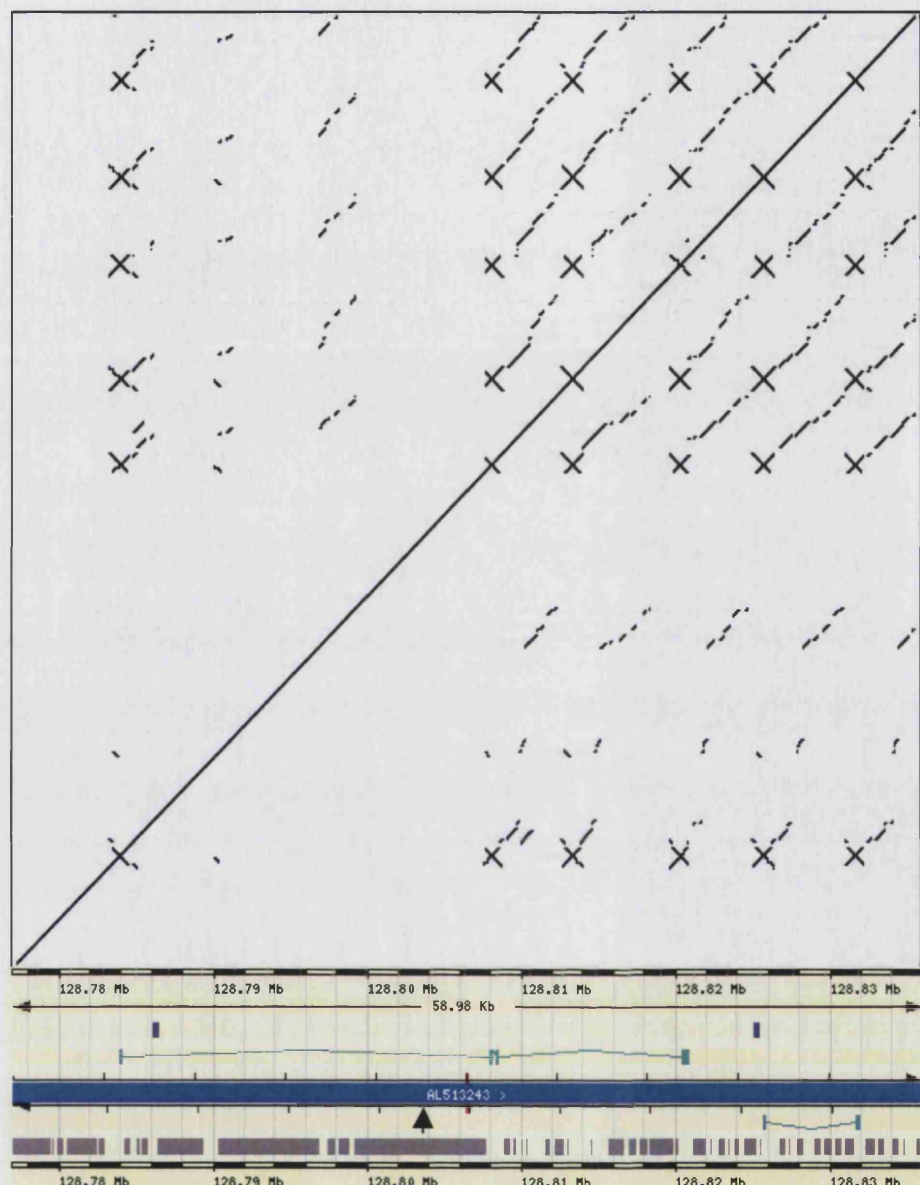


Figure 6.27. Dotplot output from Pipmaker for the whole (Repeatmasked) sequence of human genomic clone bA453F18, with Ensembl genome browser tracks lined up underneath (Release 16.33.1), showing position in Mb on the X chromosome, interspersed repeats in grey, ESTs in purple, and Genscan predictions are shown in green. The location of the insertion of Xq22 sequence containing *PLP1* is shown by the arrow.

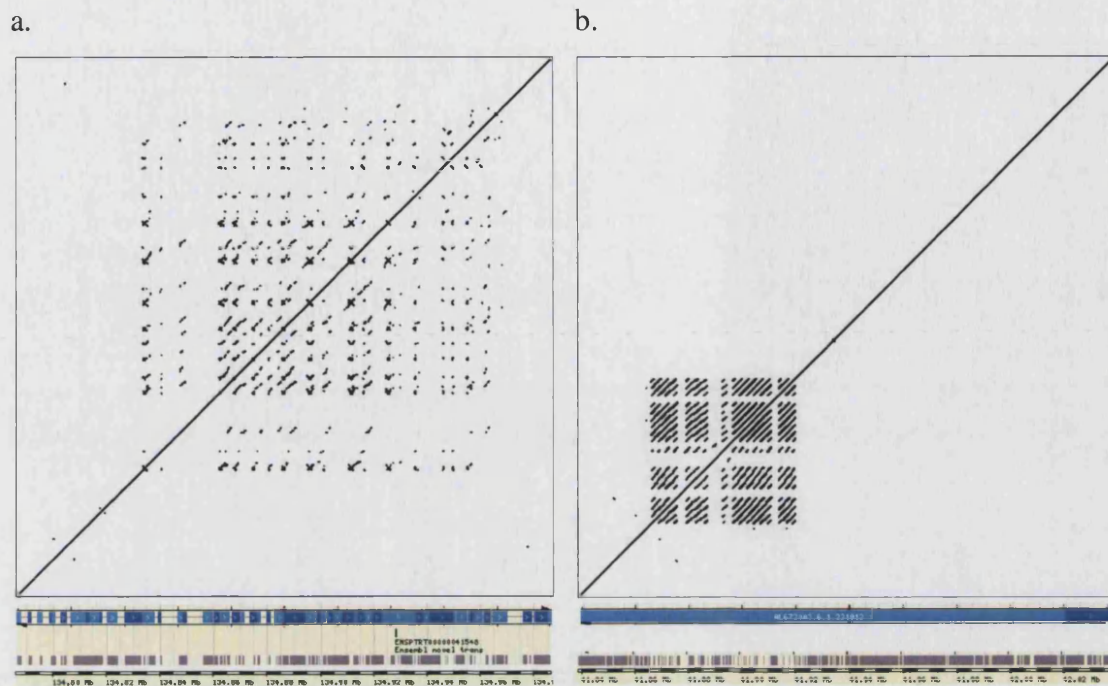


Figure 6.28 The syntenic regions of the chimpanzee (a.) and mouse (b.) genomes also contain numerous directly repeated sequence arrays. 200Kb of genomic sequence (from Ensembl genome browsers, version 22.1.1. for chimp and version 22.32b.1 for mouse) that appeared to be syntenic to the Xq26.2 insertion region were first masked for interspersed repetitive elements by Repeatmasker (Repbase version 7.4) then the masked sequence was compared against itself using Blastz/Pipmaker, and the resulting dotplot is shown here. Underneath the dotplot, parts of the Ensembl display for this region of the appropriate genome is shown, with the contig in blue, predicted genes in black, interspersed repeat content in grey, and the scale bar shows the position of this region in the chimpanzee and mouse genomes.

6.4. Discussion

6.4.1. Summary of rearrangement in family 3

Individuals from this family carried an atypical duplication including *PLP1*, which was duplicated into another region of the X chromosome, Xq26.2. The proximal end of the duplication in this family was located 90.6Kb centromeric to the start of the *PLP1* gene. The duplicated segment was also inverted in orientation with respect to the original sequence in Xq22.2 during the rearrangement (Figure 6.29.). The insertion of a segment of Xq22.2 sequence into Xq26.2 was also accompanied by the insertion of two short sequences (52bp and 55bp) into the breakpoint. These two short sequences were also from Xq26.2, distal to the insertion point, and both mainly consisted of known interspersed repetitive elements (Figure 6.19.). The other duplication/insertion breakpoint in this family has not been characterised, but previous work has shown that the distal duplication breakpoint may be in a similar region to that of family 1, and possibly family 2, and several other PMD duplications, *i.e.* in the region of the distal low-copy repeats (Woodward *et al.*, 2000; Woodward *et al.*, in preparation).

A mosaic deletion found in the carrier mother in family 3, which includes about half of the sequence from the long arm of the X chromosome, has been thoroughly characterised. The deletion breakpoints are both distant from the insertion region and the *PLP1* locus, which makes it unlikely that the deletion is connected to the duplication and insertion. The sequences obtained from the various breakpoints in this family so far show characteristics of NHEJ repair of DSBs, without large regions of homology between sequences, and microhomologies and inserted nucleotides at the breakpoint junctions (Figures 6.6. and 6.17.) (Lieber *et al.*, 2003).

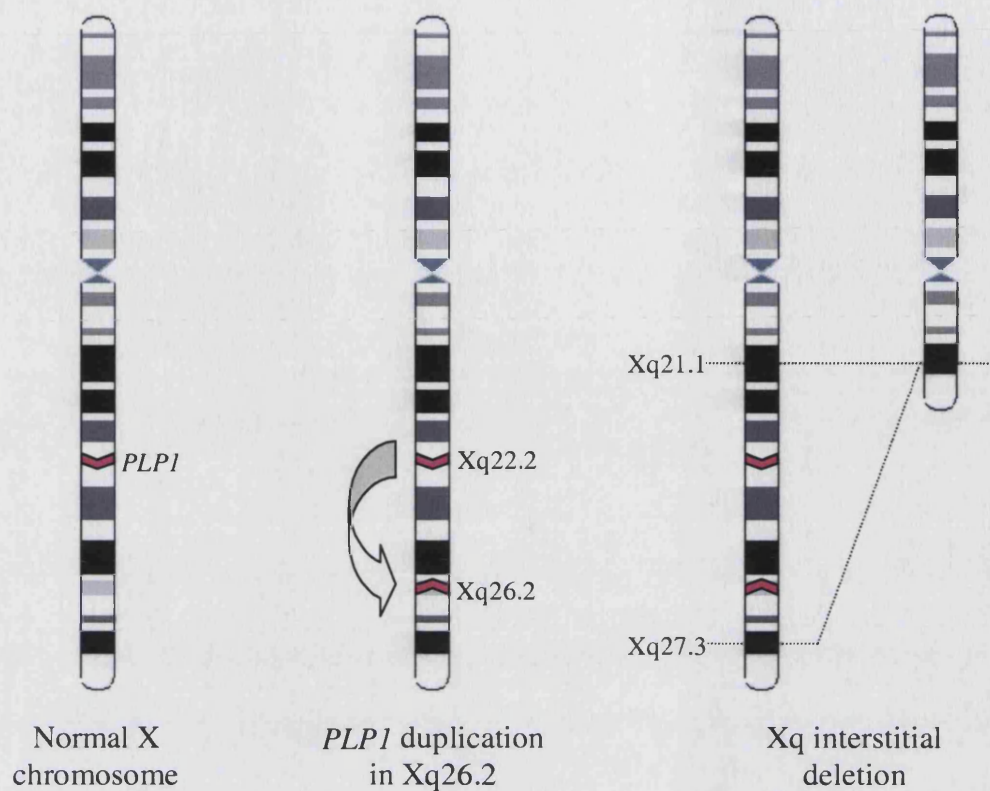


Figure 6.29. Summary of X chromosome rearrangements found in family 3. The location and orientation of the duplicated segment containing *PLP1* is shown by the pink chevron. X chromosome ideogram was adapted from the Ensembl genome browser.

6.4.2. Matrix attachment regions

MARs were found near all of the breakpoint regions associated with the duplication and insertion, and one potential MAR was also found just distal to the distal deletion breakpoint (Figures 6.10. and 6.21.). All of the potential MARs associated with the insertion breakpoints were slightly proximal to the breakpoints or insertion sequences, (Figure 6.21., Table 6.11.). This association of breakpoint sequences in this family with MARs could be relevant to the rearrangement mechanisms. Topoisomerase II cleavage sites are found in close association with MARs, and the MAR-Wiz program looks for topoisomerase II consensus sites as one of the criteria for determination of MAR-potential (Sperry *et al.*, 1989;Singh *et al.*, 1997). Cleavage of DNA by topoisomerase enzymes could be important for generating DSBs and stimulating rearrangements. Other genomic features that are sometimes associated with MARs are origins of replication (Singh *et al.*, 1997;Girard-Reydet *et al.*, 2004). Sequences including the human replication origin consensus and the PUR protein binding consensus sequence reported to be associated with human origins of replication were also found near some of the insertion breakpoints (see sections 6.3.3.4.1. and 6.3.3.4.2.) (Bergemann and Johnson, 1992;Dobbs *et al.*, 1994;Smith *et al.*, 1998;Darbinian *et al.*, 2001). DNA replication may occur in “replication factories”, distinct regions which are stationary within the nucleus and contain several polymerases and replication forks (Cook, 1999). The human genome is organised into groups of adjacent replicons which are replicated simultaneously and are associated with the same replication foci in the nucleus (Jackson and Pombo, 1998). Three of the sequences found at the insertion breakpoint originate from Xq26.2, and rearrangements such as those seen in family 3 involving these sequences could be due to different damaged and stalled replication forks associated with the same replication factory being misrepaired by NHEJ (Figure 6.17.).

6.4.3. Sequence motifs, G+C content and other features of the DNA sequence

Several sequence motifs that have been previously reported to be involved in other DNA rearrangements were found near the deletion and insertion breakpoints in family 3 (see section 6.3.3.4.). However, no single motif was observed to be particularly over-represented close to all of the breakpoints, and while it is possible that some aspects of the local DNA sequence could be involved in the generation of rearrangements, this could not be determined by this analysis.

Some short (10-16bp) polypurine and alternating purine/pyrimidine tracts were noted close to the various breakpoints in family 3 (Figures 6.11. and 6.22.). Again, this is in concordance with previous studies that have found purine/pyrimidine tracts to be over-represented at deletion breakpoints and polypurine tracts to be in excess at translocation breakpoints (Abeyasinghe *et al.*, 2003). Regions of unusual nucleotide composition may form alternative topological conformations which may stimulate rearrangements and recombination, and this may have been important for generating the rearrangements in this family (Abeyasinghe *et al.*, 2003).

A small number of repeats were found in the sequence around the breakpoints, which could contribute to secondary structure that could promote DNA rearrangements or recombination, or could help to stabilise the DNA during the process of the rearrangement (Figures 6.23. and 6.24.) (Chuzhanova *et al.*, 2003).

Gross deletion breakpoints are often located within A+T rich sequences (Abeyasinghe *et al.*, 2003). While both deletion breakpoints in family 3 were located in regions with generally high A+T content, a shorter stretch of sequence around the distal breakpoint

had a higher G+C content than the surrounding 1Mb (see section 6.2.5.1.). Translocation breakpoints have been reported to be G+C rich (Abeyasinghe *et al.*, 2003). Both the major duplication/insertion breakpoints in family 3, within clones dJ1055C14 and bA453F18 were located in regions that had a G+C content slightly higher than the surrounding 1Mb region (see sections 6.3.3.1. and 6.3.3.2.3.). The two short inserted sequences from Xq26.2 had much lower G+C content than the surrounding region, and all the sequences from Xq26.2 that were involved in the *PLP1* insertion event had a much higher interspersed repeat content than the surrounding 1Mb region (see section 6.3.3.2.). It is possible that some of the interspersed repeats found near these breakpoints could be involved in the breakpoint mechanism (see section 6.4.5.2.).

6.4.4. Insertion

One of the most intriguing aspects of the duplication in family 3 is the insertion of the duplicated sequence into an apparently unrelated region of the X chromosome. There does not appear to be any obvious similarity between the Xq22.2 and Xq26.2 sequences involved at the sequenced breakpoint. The distal duplication and insertion breakpoint from family 3 has not been characterised as yet, and further insights into the mechanism may be gained from investigation of this aspect of the rearrangements in this family.

6.4.5.1. Repetitive regions near Xq26.2 insertion

One interesting feature that has been noted in this study is the highly repetitive nature of some of the sequences close to the insertion point. The insertion breakpoint region in Xq26.2 is surrounded by a large array of direct and inverted repeated sequences in the human genomic sequence (Figures 6.26. and 6.27.). Arrays of repeated sequences are also evident in the syntenic genomic sequences of other primate and rodent genomes

(Figure 6.28.). The repeat arrays do not necessarily show interspecific homology between more distantly related species such as human and mouse.

Tandemly repeated sequences are found throughout the human genome, and many of these repeated sequences have functional and evolutionary significance. Some genes, such as those for ribosomal RNAs and small nuclear RNAs (snRNA), such as U2 snRNA, are found in extensive tandem arrays in the human genome (Van Arsdell and Weiner, 1984; Westin *et al.*, 1984). Sequences important for chromosomal function, such as centromeric satellite DNA and telomeric DNA comprise tandemly repeated sequences, and many polymorphic regions of the genome, such as micro-, mini-, macro- and mega-satellite sequences also occur in tandemly repeated sequence arrays. Concerted evolution is often noted to occur for tandemly repetitive sequences, where individual copies of a repeated sequence do not evolve independently but instead are subject to sequence homogenisation (Liao, 1999). Concerted evolution of a region containing an ancestral tandem repeat inherited from a common ancestor could result in tandemly repeated regions being present in syntenic regions of the genome in related species, but without any interspecies sequence homology found in the orthologous repeat units, as is observed for the repetitive sequences near the Xq26.2 insertion point in primate and rodent genomic sequence. Concerted evolution of tandem repeat arrays could occur by a process involving unequal crossover, either intra- or interchromosomal, or by gene conversion based mechanisms (Smith, 1976; Liao, 1999). Gene conversion has been proposed to occur by a process similar to the break induced repair/homologous strand invasion model (see section 4.12.5.), and to maintain homogenous tandem repeat arrays high levels of recombination or gene conversion are necessary (Pavelitz *et al.*, 1999; Lemmers *et al.*, 2004). A high rate of gene conversion within a tandem sequence

arrays may be driven by an excess of recombinogenic DSB formation within the array, and such regions could also be susceptible to other DSB-stimulated rearrangements, including insertions of DNA originating from elsewhere in the genome, as has occurred in family 3.

6.4.5.2. Mechanism of insertion

The two short Xq26.2 sequences inserted at the sequenced Xq22-Xq26.2 junction are both mainly composed of interspersed repetitive elements (see Figure 6.19.). It has been shown that retrotransposition of interspersed repetitive elements into pre-existing DSBs or repair of DSBs by capture of retrotransposable elements can occur (Lin and Waldman, 2001a; Morrish *et al.*, 2002). The two short inserted sequences (L1M4 and L2 elements) from the insertion breakpoint are not members of currently active retrotransposon families, but these sequences could still have been retrotranscribed in *trans* prior to capture and insertion at the breakpoint (Tables 6.8. and 6.9.) (Smit *et al.*, 1995; Smit and Riggs, 1995; Ejima and Yang, 2003). An almost full-length *Alu* sequence is present just distal to the insertion point in bA453F18 (see Figure 6.19. and Table 6.10.). *Alu*-mediated recombination and rearrangement has been proposed to be instrumental in the formation of segmental duplications, and this process may also have been involved in this duplication event (Bailey *et al.*, 2003; Babcock *et al.*, 2003; Jurka *et al.*, 2004).

As already mentioned previously, as well as short sequences being captured during the process of DSB repair, more substantial stretches of DNA can be inserted into DSBs (see section 5.8.2.6.) (Lin and Waldman, 2001a; Allen *et al.*, 2003). This type of mechanism could result in the insertion and duplication seen in family 3, either by NHEJ of isolated DNA ends, or possibly by a strand invasion and DNA synthesis model as has been

proposed for other PMD breakpoints (see sections 4.12.5. and 5.8.2.6.) (Woodward *et al.*, in preparation). A possible mechanism for the duplication and transposition of Xq22 sequence into Xq26.2 could be initiated by a non-homologous strand invasion event from one end of a DSB in Xq26.2 into Xq22, and replication could then be primed from the invading strand. Stalling of the replication fork once the *PLP1* region had been replicated, followed by non-homologous joining to the free end in Xq26.2, with the capture of some other short sequences from the region, would complete the insertional rearrangement.

6.4.6. Mild phenotype in family 3 – a possible position effect?

The phenotype in the affected boy in family is relatively mild compared to some other duplications (see section 2.3.3.). The size of the duplication in this family is relatively small compared to some *PLP1* duplications, and it is possible that smaller duplications may result in a milder phenotype (Inoue *et al.*, 1999; Woodward *et al.*, in preparation). Alternatively, the expression of the duplicated copy of *PLP1* away from Xq22 may be reduced compared to the original copy of *PLP1*, perhaps due to the local chromatin environment by a position effect (Kleinjan and van Heyningen, 1998). One case of an apparent position effect has been reported involving *PLP1* and a translocation (without duplication) to chromosome 19, where the autosomal copy appears to be expressed at either low levels, or not at all (Inoue *et al.*, 2002).

6.5. Conclusions

Characterisation of some of the breakpoints involved in the rearrangement in family 3 has shown that the duplication and insertion, and the mosaic deletion do not appear to be directly connected, as the various breakpoints are spatially separated and do not appear to share any obvious sequence characteristics. The breakpoints characterised so far do not exhibit much homology between the various sequences involved and it cannot be shown definitively from the breakpoint sequences how the rearrangement actually occurred, and several different pathways could have generated the observed breakpoints.

The occurrence of atypical PMD duplications such as in family 3 illustrates the importance of FISH as a diagnostic method for detecting duplications, as this is the only method that could detect the unusual nature of the duplication in this family. Due to the mosaic deletion in the carrier female, dosage detection methods such as quantitative PCR do not detect an increase in dosage at the *PLP1* locus (Woodward *et al.*, 1998). The carrier status of the mother in family 3 could only have been determined using a FISH methodology.

7.1. MAPH RESULTS

MAPH is a quantitative technique that can be used to determine copy number of genomic sequences (see section 1.5.4.) (Armour *et al.*, 2000). The MAPH technique was applied to the *PLP1* genomic region, with the aim of developing an alternative technique to interphase FISH for the diagnosis of duplications of *PLP1*. As well as assaying copy number of the *PLP1* gene itself, additional MAPH probes were also designed within the flanking genomic region, in order to be able to characterise the extent of the duplication in a single experiment, as duplications involving *PLP1* have been shown to be variable in size (Woodward *et al.*, 1998; Inoue *et al.*, 1999; Hobson *et al.*, 2003).

7.1.1 MAPH probe mix

A MAPH probe set was developed for the *PLP1* region, containing probes for the *PLP1* gene, the region flanking *PLP1* and various control probes. This probe set was used in a number of experiments to assess whether MAPH was a suitable method for determining gene dosage in PMD patients.

The probe mix consisted of 4 probes from *PLP1* - exons 3, 5, 6 and 7 (plp3, plp5, plp6, plp7); 6 autosomal control probes, one each from chromosomes 1, 4, 6 and 7 (ch1q24, ch4q26, ch6p24, ch7q31), and two from chromosome 17 (ch17q21, ch17p13); two sex-linked control probes, one from X (Xq12) and one from Y (sry); a non-human control probe from a *Xenopus* genomic sequence (XLnkx) and 5 probes from around the region immediately flanking *PLP1* (198p4, 79p11, 43h13, 240c2, 144a10) (Table 2.2.) . For positions of probe target sequences in or near *PLP1*, see Figure 7.1.).

7.1.2. MAPH experiments

The MAPH procedure was carried out on 58 samples in total in 18 separate experiments. Each experiment used between 3 and 7 normal controls, and between 3 and 10 test individuals, and the mean number of samples (normals and test) per experiment was 11.39. For the test samples, at least one result, but usually two measurements of dosage by MAPH, were obtained for 47 individuals, eight females and 39 males. In total 99 individual dosage results were obtained for each probe in from the test samples, 16 from females and 83 from males. 11 samples were used as normal controls, which were all used in at least two experiments, and some were used all the experiments.

7.2. Autosomal control probes

Normalised ratios were obtained for each probe, using the four control probes that were nearest in size to calculate the dosage ratio and a mean was taken for each individual sample over all experiments. For the six autosomal control probes the values of the mean and standard deviation are shown in Table 7.1. and Figure 7.2., and a histogram for each individual control probe using the individual values is shown in Figure 7.3.. The results from the control probes were each near 1, as would be expected given the method of normalisation, but some probes were more variable than others. In particular, the ch7q31 probe and the ch17q21 probe had the greatest standard deviations (Table 7.1.) and their means were the furthest away from the expected value of a normalised ratio of 1 (Table 7.1. and Figure 7.2.). Additionally, when the frequency distribution was plotted as a histogram for each control probe (Figure 7.3.), these two probes, especially ch17q21, did not appear to be normally distributed whereas the other four probes showed an approximately normal distribution. Because the ch7q31 and ch17q21 control probes did not appear to be

producing consistent results that were normally distributed, these two probes were not used when calculating normalised ratios for the other probes as they might skew the data. Instead, normalised ratios were obtained for all the test probes by using the four more reliable autosomal probes (ch17p13, ch1q24, ch6p24, ch4q26) as reference peaks instead of the four control probes that were nearest in size as had been used in the initial calculations.

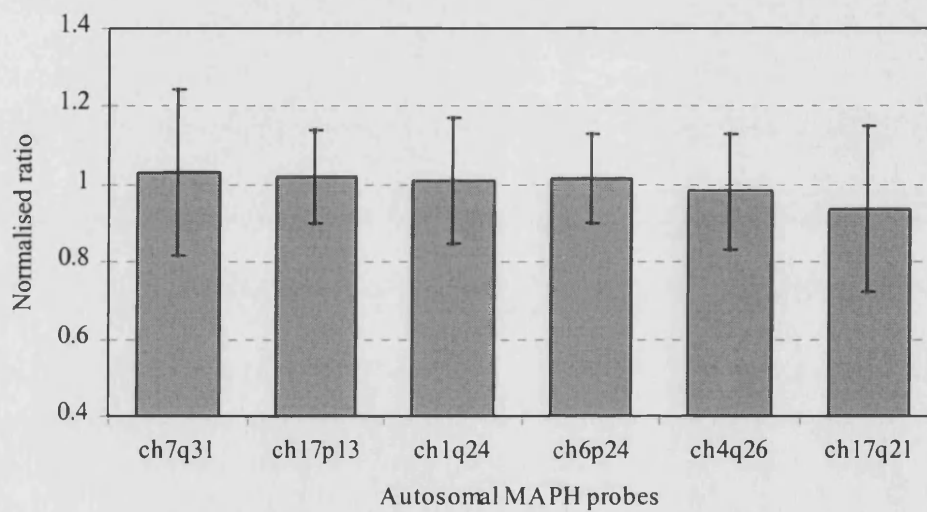


Figure 7.2. Graph showing the means \pm 1 standard deviation of the normalised ratio for the six autosomal control MAPH probes throughout all experiments.

Probe	ch7q31	ch17p13	ch1q24	ch6p24	ch4q26	ch17q21
Mean	1.03	1.02	1.00	1.01	0.98	0.93
Standard deviation	0.21	0.12	0.16	0.11	0.14	0.21

Table 7.1. Means and standard deviations for 6 MAPH autosomal control probes, using data from all experiments.

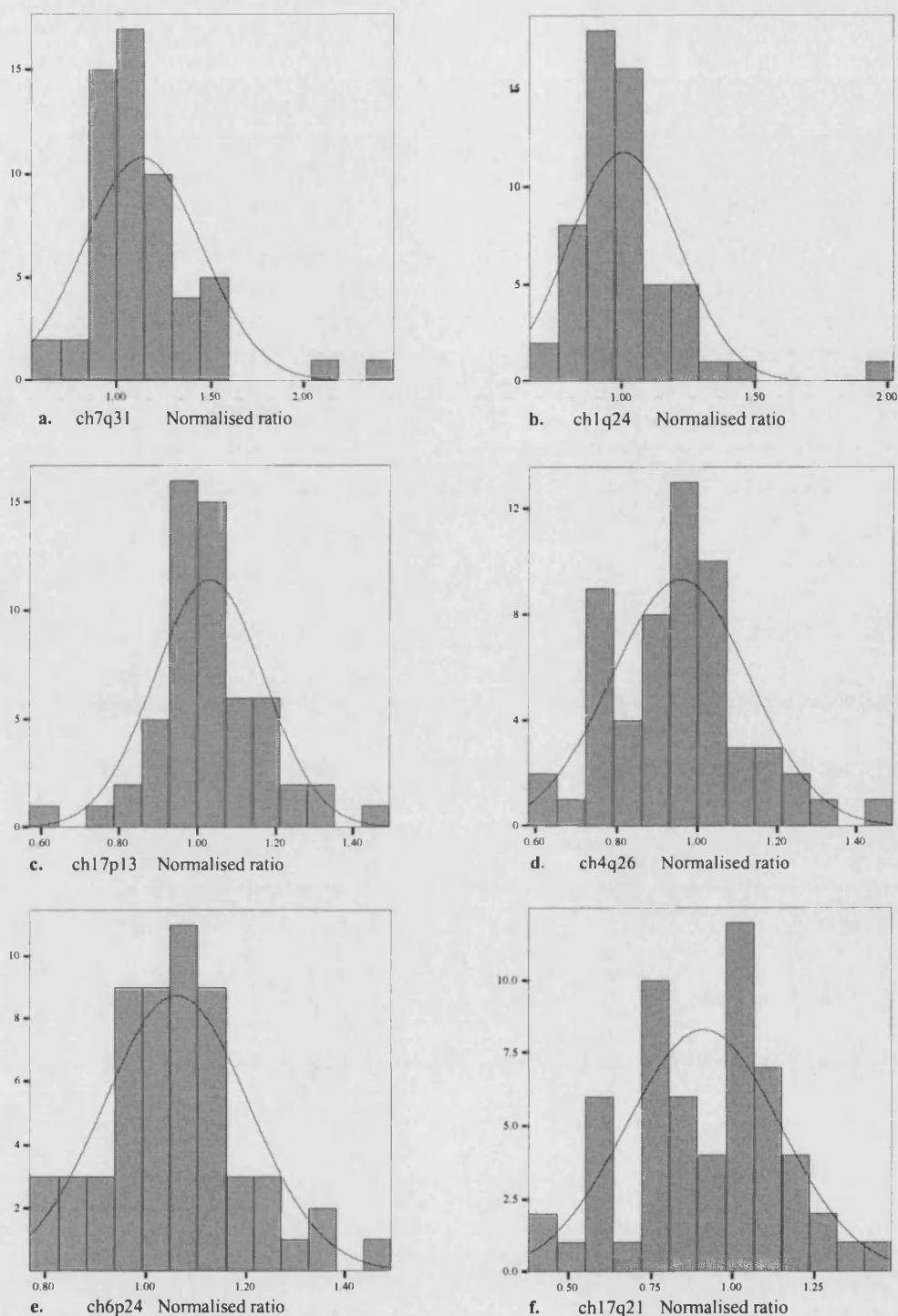


Figure 7.3. Histograms showing the frequency distribution of normalised ratios for the six MAPH autosomal control probes. The normal distribution curve that is closest to the data is shown superimposed on the histograms.

7.3. Sex-linked control probes

Two of the control probes were sex-linked, one from the Y chromosome, mapping to the *SRY* gene, and the other from the X chromosome, from the chromosomal band Xq12. The mean normalised ratios for 12 females and 45 males from several experiments (including the control individuals) are plotted in Figure 7.4. for these two probes. A clear difference can be seen between the sexes, most notably for the *SRY* probe where all the females had a normalised ratio of 0 and the male mean ratio was 0.82 (Figure 7.4.). Although the *SRY* probe did give some quite variable normalised ratios for males, the fact that it was never amplified in females showed that it was specific to its Y-linked target sequence. For the Xq12 probe (Figure 7.4.), the female mean was 0.97 and the male mean was 0.44, which were both close to the expected ratios of 1.00 for the normal female complement of two X chromosomes and 0.50 for the single male X chromosome. Similar results were also seen when looking at these data in just the control individuals. The *SRY* probe had a value of 0 for all female controls, and a mean of 0.97 for the male controls, and the Xq12 probe had a mean normalised ratio of 1.02 for the female controls and 0.48 for the male controls. The differences in normalised ratio between males and females for both Xq12 and *SRY* probes was significant using a t-test for equality of means ($p < 0.001$) for the data using either just the control individuals or all samples.

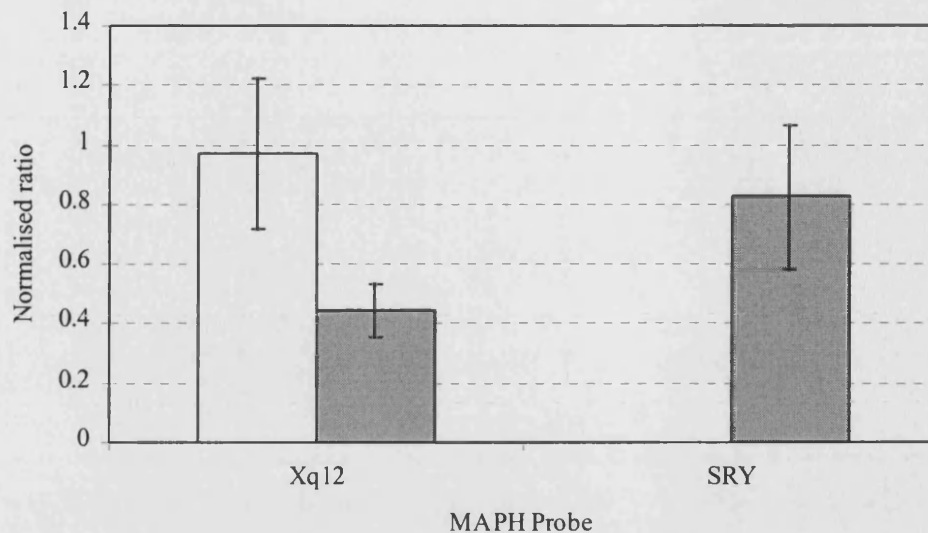


Figure 7.4. Mean normalised ratios \pm one standard deviation for the Xq12 and SRY MAPH probes for all samples. Male mean bars are shaded, females are unshaded.

7.4. Xq22 probes in control individuals

Mean values were taken for each of the Xq22 probes for each of the control individuals used in the MAPH experiments (7 males and 4 females) and the means for the males and females were then compared (Table 7.2., Figure 7.5.). Despite the small sample size, most probes did show a significant difference between the male and female controls. For the Xq22 probes where there was a significant difference in normalised ratios between the sexes, the range of female control means were between 0.878 and 1.074, all close to the expected ratio of 1, and the male control means ranged between 0.394 and 0.634. Two probes did not show a significant difference ($p < 0.05$) using a t-test for equality of means, between males and females, despite containing X-linked sequences. These were *PLP1* exon 3 and *PLP1* exon 7. The *PLP1* exon 7 probe showed the least difference between sexes with the two means being virtually the same (0.790 and 0.718). In addition, when MAPH was carried out on a male with a deletion that included the whole of the *PLP1* gene, the exon 7 probe

was amplified from this sample, but the other three *PLP1* probes in the mixture were not amplified as expected (Raskind *et al.*, 1991; Inoue *et al.*, 2002). Further analysis was not carried out on the results for these two MAPH probes as they were unable to detect the differences in sex chromosome dosage reliably in normal individuals, and the exon 7 probe also appeared to hybridise to at least one other site in the genome that was not in the immediate vicinity of *PLP1*.

Probe	198p4	79p11	43h13	PLP3	PLP5	PLP6	PLP7	240c2	144a10
Female mean	0.938	0.937	0.878	0.976	1.003	0.9916	0.790	1.075	1.053
Female SD	0.126	0.138	0.083	0.259	0.080	0.114	0.092	0.231	0.097
Male mean	0.513	0.465	0.634	0.601	0.523	0.473	0.718	0.518	0.394
Male SD	0.081	0.061	0.124	0.047	0.191	0.052	0.072	0.041	0.146
Significance	p=0.003	p=0.003	p=0.004	p=0.061	p<0.001	P=0.001	p=0.234	p=0.023	P=0.016

Table 7.2. Means and standard deviations (SD) for the X- and Y-linked MAPH probes for 4 normal females and 7 normal males. The significance of the differences between the means was tested using a t-test for equality of means.

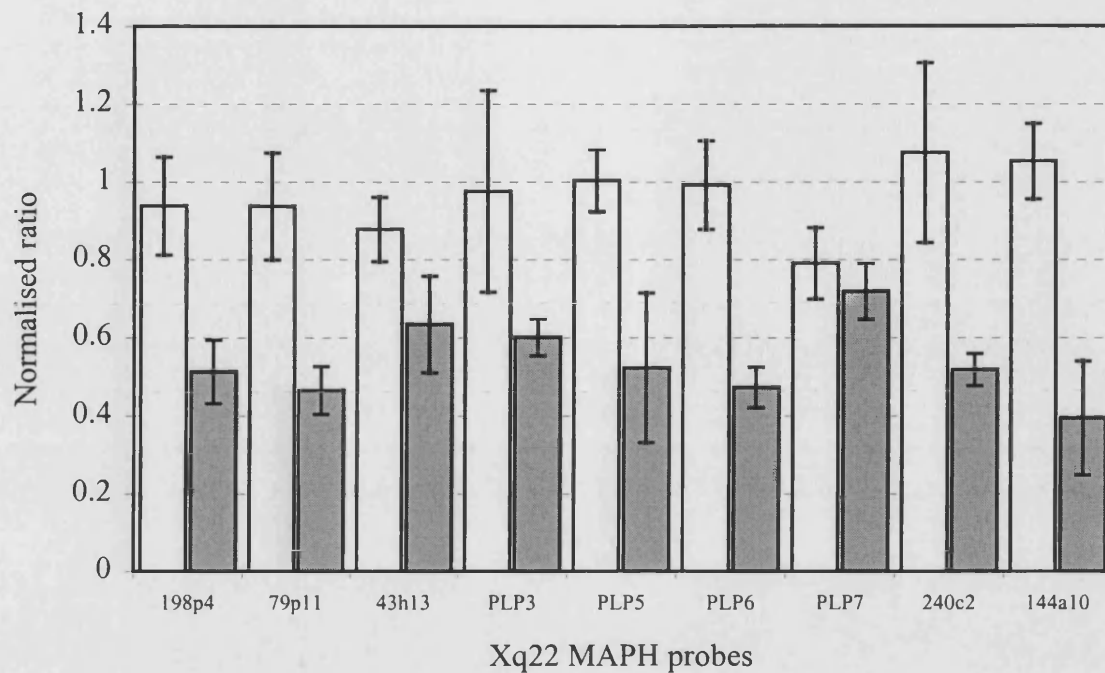


Figure 7.5. Bar chart showing means \pm 1 standard deviation for Xq22 probes in normal controls. Male means are shaded bars, female means are white.

7.4.1. *PLP1* probes

The two *PLP1* probes that appeared to be producing consistent results (plp5 and plp6) using control males and females were looked at in a cohort of patients who had a probable diagnosis of PMD, or if female, were possible carriers of PMD. Interphase FISH using a cosmid containing *PLP1* had been carried out on most of these individuals, so the *PLP1* duplication status was already known. Figures 7.6. and 7.7. show the mean normalised ratios for the different categories of individuals for both probes. Male controls and male PMD patients who had been found not to have a duplication by interphase FISH had normalised mean ratios for the plp6 MAPH probe close to 0.5; female controls and males known to have a duplication had ratios close to 1, corresponding to two copies of *PLP1*; and females who carried a duplication of

PLP1 (three copies of the gene), had a mean normalised ratio near 1.5. The *plp5* probe had similar results, but the differences between the male controls and duplicated males, and between the female controls and duplication carriers was not as great, and also the results were much more variable, especially for the female duplication carriers (Figure 7.7.). The MAPH protocol was carried out on a male who has been found to have a deletion of the *PLP1* gene, and a ratio of 0 was obtained for this individual for both *PLP1* probes (Table 7.3.) (Raskind *et al.*, 1991).

One family is suspected to carry a triplication of the gene, which has been detected by Southern blotting, quantitative PCR and interphase FISH (Ellis and Malcolm, 1994; Woodward *et al.*, 1998; Wolf *et al.*, in preparation). MAPH was carried out on an affected boy from this family and his carrier mother. The ratios for the boy for both *PLP1* probes were similar to the ratios for carrier females (Table 7.3.), and the results for the mother were a lot higher than the other carrier females (Table 7.3.), although considerably greater than the expected normalised ratio of 2, at 2.63 for *plp5*, and 2.33 for *plp6*. These data did support the idea that there is increased dosage of the gene in this family, greater than a duplication and most probably a triplication.

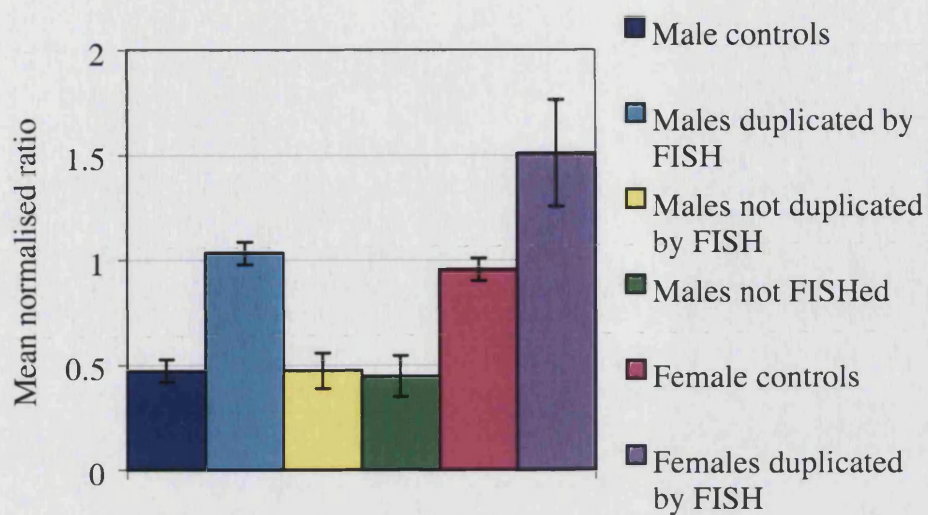


Figure 7.6. Mean normalised ratios for the MAPH probe plp6, with +/- one standard deviation shown by the error bars.

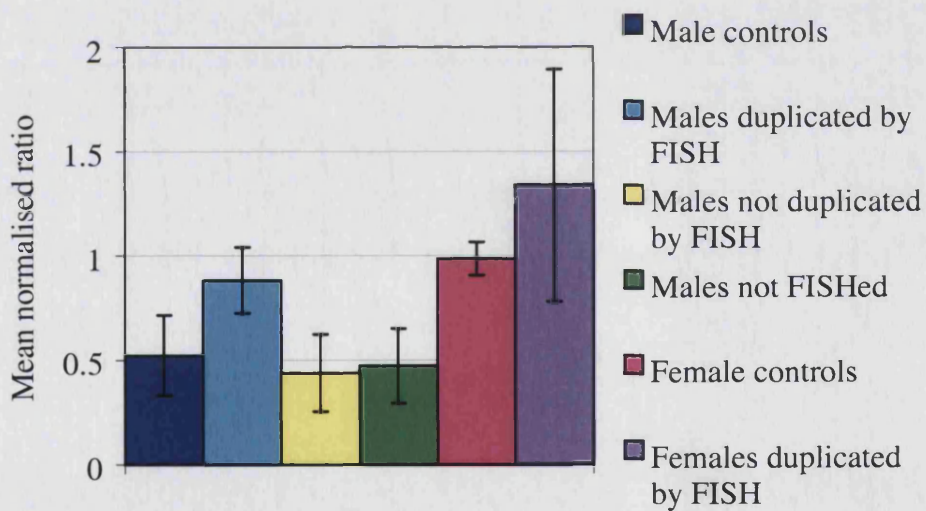


Figure 7.7. Mean normalised ratios for the MAPH probe plp5, with +/- one standard deviation shown by the error bars.

	N	plp5 mean	plp5 SD	plp6 mean	plp6 SD
Male controls	7	0.521	0.192	0.473	0.054
Males duplicated by FISH	16	0.882	0.158	1.03	0.092
Males not duplicated by FISH	16	0.439	0.185	0.474	0.085
Males not FISHed	4	0.473	0.178	0.448	0.098
Female controls	4	1.000	0.078	0.990	0.116
Females duplicated by FISH	6	1.340	0.558	1.510	0.252
Female not FISHed	1	1.06	-	1.04	-
Male deletion	1	0	-	0	-
Male triplication	1	1.33	-	1.38	-
Female triplication carrier	1	2.63	-	2.33	-

Table 7.3. Means and standard deviations for the different classes of PMD patients and controls for MAPH probes plp5 and plp6. N shows the number of individuals in each class.

7.4.1.1. Analysis of individual normalised ratios

Although when the results for each sample were averaged together, there was clearly a difference between the different categories, when carrying out a screen for changes in gene dosage, measures of significance need to be assigned to the quantitative results obtained for each individual (Figures 7.6. and 7.7.). A method that has been used to assess statistical significance of MLPA dosage quotients was adapted to analyse the dosage results using the *PLP1* MAPH probe set (developed by Andrew Wallace, National Genetics Reference Laboratory, Manchester, see <http://www.ngrl.org.uk/Manchester/Pages/mutationsdload.htm>).

This approach uses the standard deviation of the control probes from an individual experiment, and each normalised ratio is expressed as a number of standard deviations from the control mean normalised ratio for that probe. Significance for each probe for each individual was tested by using Student's t-distribution, which allows for small sample sizes, to calculate the probability of that value being contained within the t-distribution. Each individual experiment had been carried out with a number of test samples (up to ten) and some normal individuals of both sexes (usually two to four of each sex). As well as the probability of each probe having a normalised ratio within the same range as normal individuals, the probability of each probe being duplicated was also calculated. For male samples, the value expected for a duplicated sample was estimated by doubling the male control mean value. Each test value was then compared against this estimated duplicated ratio, using the same standard deviation as before. Similarly for female samples the female control mean value was multiplied by 1.5, to get an expected value for duplications, and compared to the test normalised ratios, using the same standard deviation that had been calculated from the normalised ratios obtained from the control individuals.

Not all the normalised results obtained from the MAPH experiments could be tested in this way, as in two experiments data had been collected from only one male control individual, as the other male control samples in that experiment did not happen to work on that occasion. In these instances, as the standard deviation cannot be calculated from a single number, significance was not able to be tested in this manner for the eight male samples from those two experiments, so only 75 male dosage results were used in this analysis.

7.4.1.2. Analysis of the significance of individual normalised ratios for MAPH probes plp5 and plp6

For both sexes, and also both *PLP1* probes, there was a large overlap in many of the normalised ratios, some of these were significantly different from the normal means and some were not significantly different. This was also the case when the normalised ratios were compared against the simulated duplicated ratios. For instance, some male normalised ratios with values close to 1 were found not to be significantly different from the normal male mean, and some strikingly high (normalised ratios greater than 2.5) dosage results from females for the plp5 MAPH probe were still not found to be significantly different from normal female control normalised ratios.

To be able to classify a result as duplicated with confidence, the ratio has to be significantly different from the normal mean value, and also not be significantly different from the duplicated mean value, with the converse being necessary for a result to be classified as normal. Out of the 91 individual samples tested, 45 could be categorised in this way using the normalised ratios from the plp6 MAPH probe, and 33 could be categorised using the plp5 dosage results. These results did not represent 91 separate individuals, as 47 different samples were tested in total, with 41 being analysed by MAPH more than once. For the individuals where more than one result had been obtained, to be confident about a dosage result, it must have been found to be the same each time the individual was tested. For the plp6 probe, only five individuals who had been tested more than once had the same result each time (normal or duplicated), and for the plp5 probe, only two individuals had consistent non-equivocal results each time they had been tested.

This difficulty in classifying dosage results adequately was mainly due to the high variability found in the control samples. The average standard deviation, taken from all experiments, for the MAPH probe plp5 was 0.14 for male controls and 0.27 for female controls. For the other *PLP1* probe, plp6, the mean standard deviation was 0.06 for male controls and 0.12 for female controls. The mean number of controls used in each experiment was 2.65 for males and 2.94 for females. For a difference in normalised ratios of 0.5 (the difference between the expected normalised ratio for normal and duplicated for both males and females) to give a p-value of less than 0.05, the difference expressed as a number of standard deviations must equal 4.30 or greater, using a two-tailed Student's t test with two degrees of freedom. Two degrees of freedom has been used for this example due to the mean numbers of controls used in the MAPH experiments. A difference of 4.30 standard deviations equates to a standard deviation of 0.1163. Thus if a cut-off point is set for an acceptable standard deviation at about 0.10, or 10%, differences from the normal mean ratio by a magnitude of 0.5 should be classified as significant by this method of analysis. The standard deviations for the plp5 probe were frequently greater than 0.10, especially for females. Also for probe plp6, relatively few experiments had female control values with a standard deviation less than or equal to 0.10. However, for male dosage of plp6, most experiments did have a standard deviation of less than 10% (13/16 experiments with at least two male controls). This high variability in the normalised control ratios in many experiments explains why many ratios, despite being very close to the expected value for a particular category (*e.g.* 1.0 for males with a duplication), are not classed as significantly different from the alternative (*e.g.* 0.5, normal male).

7.4.2. Xq22 probes flanking *PLP1*

Five MAPH probes were used flanking *PLP1* in Xq22 (Figure 7.1a). These were all included in the probe mix that was used in each experiment, with the aim of using dosage results from these probes to assess the extent of Xq22 duplications in a single test.

7.4.2.1. 198p4

This probe is the most centromeric to *PLP1* and as such, had only been found to be duplicated by using interphase FISH in a few individuals out of those tested (Figure 7.1.). Figure 7.8. shows the differences in mean normalised ratio between females, males not believed to have a duplication including this probe, and 3 males with a known duplication thought to extend past this genomic clone. Some results were not included because the duplication status was not previously known. Although the mean of the three “duplicated” males for this probe is quite close to the expected value for duplications (0.87), when these results are examined individually two of the three samples have mean ratios that are more consistent with this region not being duplicated (Figure 7.8.).

Further analyses were carried out on dosage quotient data from this MAPH probe. Significance values for each 198p4 dosage result were calculated as before (see section 7.4.1.1.). 34 of the individual male dosage ratios, and 4 of the female individual dosage ratios for this probe could be classified as either normal or duplicated in this way. 34 out of the 40 males were tested more than once, and of these, 7 had all dosage ratios classified as not duplicated. None of the three males that were affected with PMD and were duplicated for this clone by interphase FISH could be unambiguously classified in this way. For the sample with a mean dosage ratio

above 1, one of the dosage ratios was called as being duplicated, as it was significantly different from the normal control ratio for this probe, but did not differ significantly from the “duplicated” control ratio. The other dosage measurement using MAPH probe 198p4 on this individual was relatively high (1.72) and was thus found to be significantly different from both the normal control and simulated duplicated control ratios. The dosage results for this individual are more consistent with being duplicated than not. The MAPH technique measures sequence copy number at a much higher resolution than interphase FISH so it is possible that the MAPH dosage results were genuinely reporting the dosage of the target sequence at the sequence level, and perhaps only part of the region hybridised to by this genomic clone was in fact duplicated in the two males that were thought to be duplicated. None of the 7 female test samples for which there were at least two MAPH dosage ratios for probe 198p4 were consistently shown to be either duplicated or not duplicated. There were a few presumed false positive ratios among the males tested, three of the dosage ratios for individual males who had either been found not to have a duplication at all or had a duplication of *PLP1* which did not extend this far proximal to the gene. Although there is a possibility that there could be an additional duplication of this region, separate to the *PLP1*-containing duplication such as the complex rearrangement described in family 2 (Figure 5.15.), it was most likely that these were just spurious results, as other dosage results for this probe from these individuals did not appear duplicated.

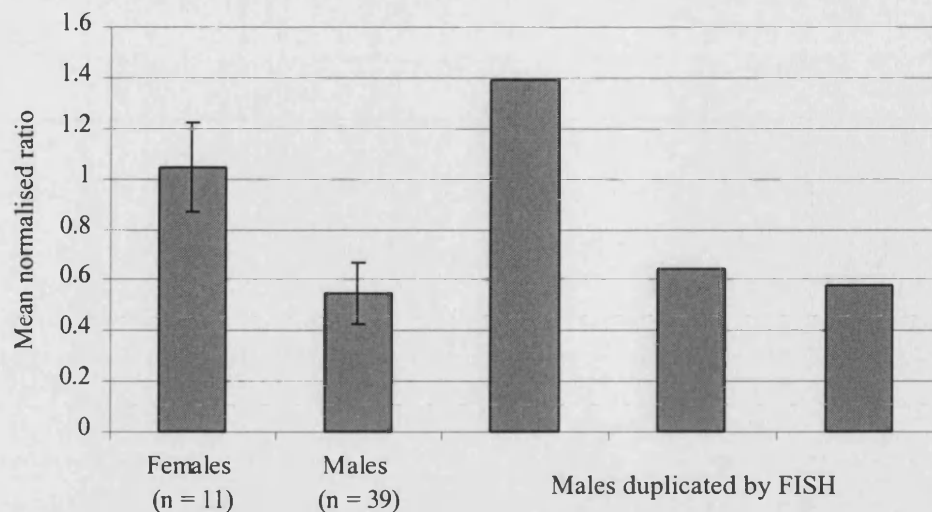


Figure 7.8. Mean normalised ratios for the MAPH probe 198p4, with +/- one standard deviation for the non-duplicated males and females shown by the error bars. The individual means from 3 males that had a duplication of this probe by FISH are shown as separate columns.

7.4.2.2. 79p11

As for the 198p4 MAPH probe, there were few individuals who were known to have a duplication encompassing this probe. When the mean dosage quotients from all the individuals tested were separated into groups, there was a clear difference between the groups (Figure 7.9.). Normal males, and males without a duplication of this region by interphase FISH had a mean for this probe of approximately 0.5, whereas the females, none of which were believed to have a duplication of this probe, and the males known to have a duplication including this region by FISH had means close to a normalised ratio of 1 (Figure 7.9.).

These data were also examined in the same way as the other MAPH probes, by comparing each dosage ratio for this probe against the appropriate control mean for that experiment, and assessing if there was a significant difference from the mean by

Student's t-test. 38 male and 8 female individual dosage measurements were classified as normal copy number using this method (see section 7.4.1.1.).

Only a few individuals where two or more normalised ratios had been obtained were consistently classified as duplicated or not duplicated for this probe, just 6/34 males and 2/7 females (see section 7.4.1.1.). Of the 6 males that were classified in this way, four were males who did not have a duplication of *PLP1*, one was a male with a deletion including *PLP1* but not the region of this probe, and the other was a male that did have a duplication of *PLP1* by interphase FISH. All these 6 males were classed as not being duplicated by these criteria. The two females that were categorised were both found to have a normal copy number of this probe target sequence. One of the females did have a duplication of *PLP1* from interphase FISH data whereas the other did not.

None of the 7 males that had been found by interphase FISH to have a duplication extending into the region contained in the MAPH probe 79p11 were found to be duplicated using these strict conditions. However, many of these individuals did show normalised ratios close to 1 or even higher, which is what is expected for a duplicated sequence. 5 out of these 7 individuals had all but one of their normalised dosage ratios classed as duplicated (2-4 results per person). Another individual had normalised ratios much above 1 in both experiments (1.30, 1.16), but these two dosage ratios were both classed as significantly different from both normal and duplicated means. The other male thought to have a duplication of this probe had one relatively low result (0.63) and one high dosage one (1.14), which was again found to be significantly different from both normal and duplicated.

In many of the experiments (12/15 for the male controls and 6/17 female controls) the standard deviation this probe was less than 0.1, so only a minority of experiments might be reasonably expected to classify normalised ratios by this method, as has been found to be the case for this probe. While an overview of these data shows that this MAPH probe (79p11) is able to detect a change in dosage of the target sequence on occasion, this is not particularly consistent. In general, the normalised ratios from this probe are too variable for effective statistical analysis. Data from this probe could be used as an indication as to whether this region proximal to *PLP1* is duplicated, but would have to be followed up with another technique to confirm any changes in gene dosage.

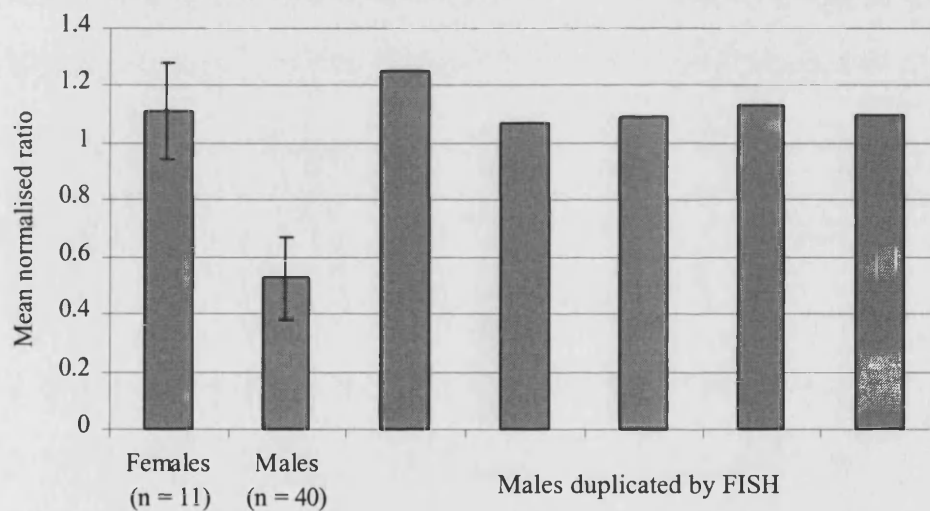


Figure 7.9. Mean normalised ratios for the MAPH probe 79p11, with +/- one standard deviation for the non-duplicated males and females shown by the error bars. The individual means from 5 males that had a duplication of this probe by FISH are shown as separate columns.

7.4.2.3. 43h13

This MAPH probe was the closest probe proximal to *PLP1*, located at 232Kb upstream to the transcription start site of the gene. Initial results for this probe were not promising, as although there was a significant difference between normal males and females in the normalised ratios obtained for this probe, the means of the two sexes were much closer together than the other MAPH probes (excluding the two *PLP1* probes that did not show a significant difference between the sexes) (Figure 7.5., Table 7.2.). The means from the controls also showed some of the greatest differences from the expected normalised ratios of 0.5 for males and 1.0 for females (Figure 7.5., Table 7.2.).

A similar analysis of the mean dosage frequencies in different categories of individuals was carried out, including cases that were believed on the basis of FISH and UPQFM-PCR data to have a duplication extending into the region where this probe hybridises (Figure 7.10.). As can be seen in Figure 7.10., any differences between the four groups (males, females, males with a duplication of this region, females with a duplication of this region) were not great. The difference in means between the males believed to have only a single copy of this genomic region and those known to be duplicated was only 0.107, and the range of values for the males with a normal copy number ranged from 0.2 to 1.9. Such variability in the normal individuals and the small differences between groups would make it very hard to be able to confidently ascertain the duplication status for this probe by MAPH. In accordance with this, there was very little difference apparent between individual dosage ratios that were classified in the various categories as assessed by a t-test (see section 7.4.1.1.). Only 4 of the 34 males where there were two or more dosage results

obtained could be consistently categorised, all as having a normal copy number for this probe. When the standard deviations of the normal controls for each individual experiment were examined, they were generally high with the mean standard deviation across all experiments for the male controls equalling 0.15 and for the female controls the mean was 0.25.

In general, this MAPH probe is not reliable enough to give consistent results regarding the copy number at this locus, due to the high variability of normalised ratios.

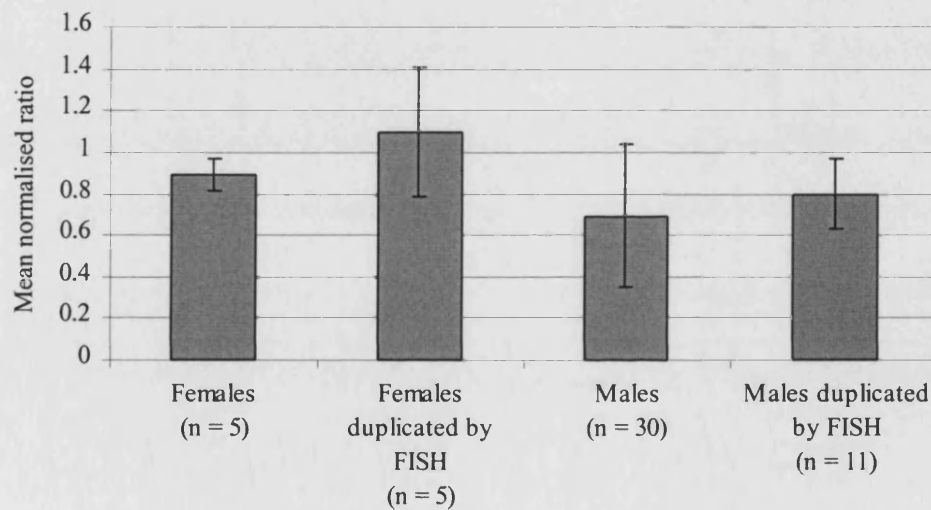


Figure 7.10. Mean normalised ratios for the MAPH probe 43h13, with +/- one standard deviation shown by the error bars.

7.4.2.4. 240c2

This MAPH probe contained sequence from the region containing all the distal low-copy repeats and mapped to the more proximal end of the sequence from human genomic clone cU240C2, between the LCRs PMDA and PMDB (Figure 3.4.). A large proportion of the distal boundaries of duplications including *PLP1* in PMD families have been mapped to this region of distal LCRs using other techniques, so a reliable method of dosage detection in this region of Xq22.2 could be especially useful for mapping duplication size (Woodward *et al.*, in preparation). When the mean dosage results for all the individuals (controls and PMD patients) were grouped appropriately and summarised, the means for the different groups were roughly as expected (Figure 7.17.).

However, for this MAPH probe to be useful, it needs to be able to reliably detect changes in dosage in individual normalised ratios. As before, this was assessed using a t-test on the difference between an individual normalised ratio and the mean of the controls of the same sex (see 7.4.1.1.). As a result of the generally quite high variability of the normalised ratios, few individual ratios could be satisfactorily classified; 6 out of the 16 dosage ratios for females could be classed as normal copy number or duplicated, and 28 out of 75 male normalised ratios in total could be classed using these criteria (see 7.4.1.1.). When all the results for each individual were considered, three males and one female had the same classification both times they were tested, all the males were not duplicated for the 240c2 MAPH probe, and the female did have a duplication of this region, in agreement with previous interphase FISH and other dosage data from this region for these individuals.

For this probe 240c2, as with the other MAPH probes, statistical analysis and classification of normalised ratios was hampered by the high variability seen within individual experiments.

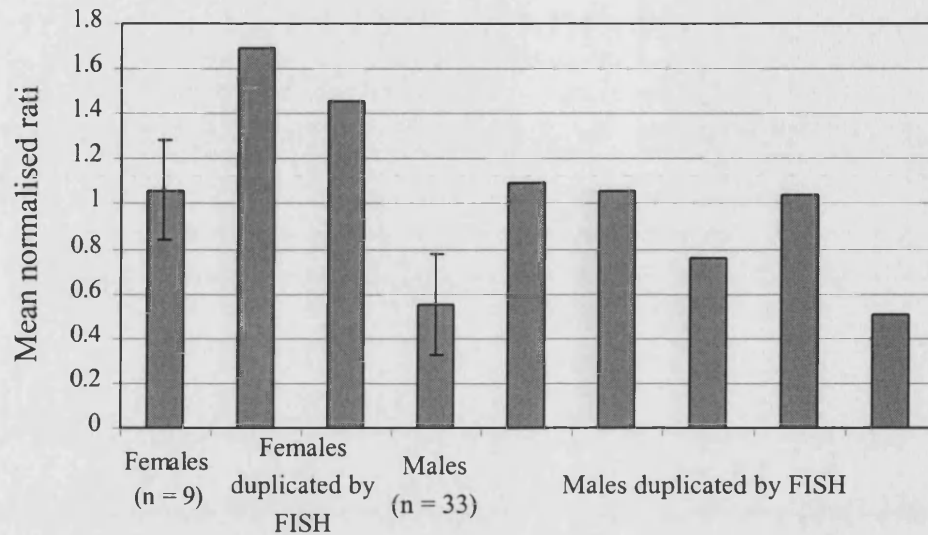


Figure 7.11. Mean normalised ratios for the MAPH probe 240c2, with +/- one standard deviation for the non-duplicated males and females shown by the error bars. The individual means from 2 females and 5 males that had a duplication of this probe by FISH are shown as separate columns.

7.4.2.5. 144a10

This probe was most distal to *PLP1*, and also maps within the cosmid clone that is routinely used as a control probe in interphase FISH, as it is only rarely found to be duplicated in PMD patients. However, amongst the cohort used in the MAPH experiments, there were two males and two females that were duplicated for this cosmid by interphase FISH, who did have increased mean normalised ratios for this probe (Figure 7.12.).

When individual normalised dosage ratios were assessed statistically, however, few results could be categorised (see section 7.4.1.1.). Just more than a quarter of male normalised ratios (20/75) were classed as either normal or duplicated, and 5/16 female normalised ratios could be similarly categorised. Only one male could be consistently classified (as having normal copy number) for this MAPH probe. This probe gave generally variable results, particularly in females, where the mean standard deviation across all experiments was 0.26. The mean standard deviation for the male normal controls was less than half this, at 0.12, but was still too large to produce consistent results. The four individuals who were known to have a duplication including this probe mostly did have normalised ratios within the range expected for a duplication, but only one of the ratios from the males and two of the ratios from the females were classed as being duplicated by the criteria used (see section 7.4.1.1.).

As with the other MAPH probes used in this study, dosage results from 144a10 did show an overall difference between normal males and females and those with duplications of this locus, but in most cases high variability of normalised ratios within an experiment has prevented a robust classification of individual dosage results as duplicated or normal copy number.

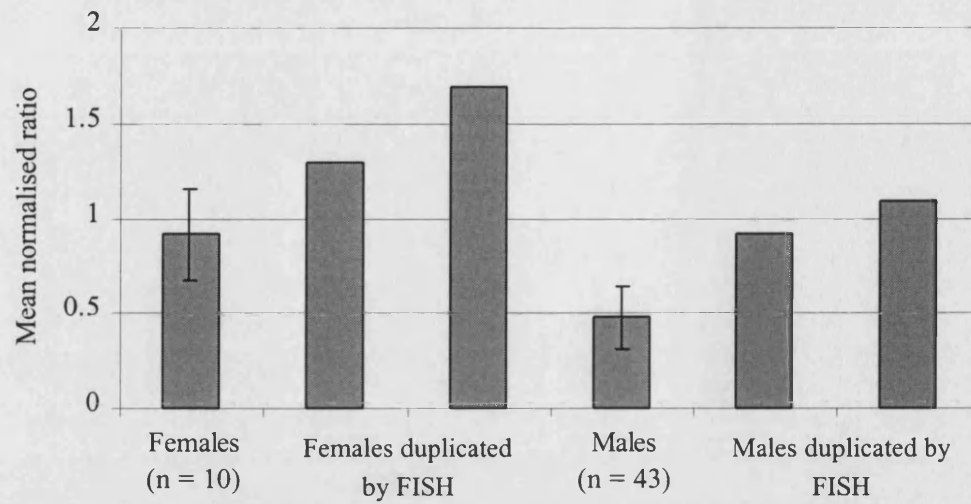


Figure 7.12. Mean normalised ratios for the MAPH probe 144a10, with +/- one standard deviation for the non-duplicated males and females shown by the error bars. The individual means from 2 males and females that had a duplication of this probe by FISH are shown as separate columns.

7.5. Discussion

A MAPH probe set including probes for *PLP1* and the surrounding genomic region was used on a cohort of individuals with a family history or symptoms consistent with PMD, as well as normal controls. Although changes in dosage were detected by MAPH, with increased normalised ratios seen for many different probes, high variability in dosage results has meant that most of these dosage changes are not statistically significant.

The MAPH protocol was shown to be specific, as a non-human probe (XLnkv) included in the probe mix was not amplified following hybridisation to human genomic DNA and stringent washing. Differences in male and female dosage ratios are apparent for X-linked probes, and the Y-linked control probe is not amplified from female samples (Figure 7.4. and 7.5.). Additionally, MAPH probes specific to *PLP1* were not recovered from genomic DNA from a male known to carry a *PLP1* deletion (Table 7.3.).

The high variability found in the MAPH results for this probe mix could be as a result of a number of different factors. The MAPH protocol consists of several stages, and errors and variability could be introduced into the results at any one of these points. Other users of the MAPH technique have reduced variability in MAPH results by ensuring that the wash solutions are at a constant 65°C, and adjustment of PCR conditions (Jess Tyson and Tamsin Majerus, personal communication). Another area where errors could perhaps be minimised is in the detection of the amplified probes. In this study this has been carried out by electrophoresis of fluorescently labelled PCR products in a slab gel electrophoresis system, but imperfections in the gel could lead

to errors in the results. Switching the detection stage to a capillary electrophoresis format would eliminate any variability introduced by slab gel imperfections.

Another factor that has had an adverse effect on the MAPH dosage results has been the small number of control individuals. Due to the requirement for a relatively large amount of DNA from each individual to be used in each experiment ($\sim 1\mu\text{g}$), and limited availability of normal control stocks with sufficient quantities of DNA, relatively small numbers of controls were used in each experiment (see section 7.1.1.).

Some of the probes showed different degrees of variability, which could be due to a number of factors (Figures 7.3., 7.4. and 7.5.). Although all potential probe sequences were checked for potential cross-hybridisation to other genomic sequences by BLASTn searches against human genomic sequence, there are some unsequenced regions of the human genome, such as heterochromatic regions, and uncloned gaps, that could be harbouring sequence with similarity to some of the MAPH probes. Hybridisation of MAPH probes to as yet unsequenced regions of the human genome could account for some variability and difficulties in detection of increases in dosage, such as those seen with one of the *PLP1* probes, *plp7* (Figure 7.5.).

Another factor which could be improved upon in the MAPH probe set is the number of control probes. Other MAPH probe sets in use in other laboratories can include approximately 40 probes, and contain several controls (Armour *et al.*, 2000; Hollox *et al.*, 2002; White *et al.*, 2002). In these larger probe sets, dosage quotients are calculated by dividing the area underneath each peak by the sum of the nearest 4 autosomal control peak areas (Hollox *et al.*, 2002). This has the effect of controlling for any differences in PCR amplification of probes based on size, as shorter probes

might be expected to amplify more efficiently. As the *PLP1* MAPH probe set only contained 6 autosomal control probes, two of which were excluded from the calculation of normalised ratios because of high variability, all four remaining control probes were used in the calculation of normalised ratios, abandoning this mechanism of controlling for size-based differences in probe amplification. Including more autosomal control probes in the *PLP1* MAPH probe set would eliminate this problem.

Only two of the *PLP1* MAPH probes, plp5 and plp6, showed a reasonable amount of specificity and discrimination in determining copy number (Figure 7.5.). Ideally a MAPH probe set for this gene would include a probe for each exon, as this would enable detection of small-scale dosage changes within the gene. Deletions or duplications of a few exons of *PLP1*, could account for some cases of PMD where the mutation is undefined, particularly in the case of a small internal duplication, as this would probably be undetected by conventional sequence-based mutation detection, and would also be undetectable by interphase FISH. The addition of reliable probes for the other 5 exons of *PLP1* would be a useful addition to this probe set.

Some of the results from MAPH probes flanking *PLP1* contradicted previous findings from interphase FISH. Although these could be false positives and negatives, due to the high variation seen in the dosage ratios for these MAPH probes, it is also possible that some of these supposedly spurious results may in fact reflect genuine copy number changes in these regions. As has been found in some PMD families, including family 2, not all *PLP1* duplications are simple tandem duplications and more complicated rearrangements may be present (see chapter 5). However, until the MAPH probes for this region have been made more reliable and show much less variation, these potential copy number changes will not be consistently detected.

7.6. Summary

Although MAPH can detect copy number changes in the *PLP1* region, high variability in the results means that this probe set is not currently suitable for use in the diagnosis of copy number changes in this region. However, further refinements and improvements to the probe mix and protocol could make this method more suited to a diagnostic setting, where the potential for high throughput and low cost could be advantageous compared to other methods that can be used for assaying copy number of *PLP1*, such as MLPA and interphase FISH (see section 1.5.).

8.1 *SOX3* DUPLICATION SCREENING AND ANALYSIS OF FAMILY 4

Interphase FISH was used to screen individuals from 15 families with suspected X-linked hypopituitarism for duplication of the *SOX3* locus (Figure 8.1.).

8.2 Interphase FISH screening

A BAC clone (bA51C14) mapping to Xq27.1 and containing the *SOX3* gene was used as a FISH probe in all the interphase FISH experiments. A control probe was also used on most individuals, either a clone from the *PLP1* region in Xq22, or another genomic clone from Xq27.1. At least 100 nuclei were scored for each probe, and 28 individuals were screened in total. The interphase FISH screen identified one family (Family 4) that appeared to carry a duplication of the *SOX3* locus (Figure 8.1.). This family included two half brothers (4:6, 4:7), both with a pituitary phenotype (see section 2.3.4.). Between 65-75% of the nuclei scored for each of the affected males contained 2-4 closely associated signals from the bA51C14 BAC FISH probe, which was consistent with this genomic region being duplicated in these two males (Figure 8.1. and Table 8.1.). The mother (4:4) and maternal grandmother (4:1) from family 4 were also checked for duplications, using the same genomic clone as a FISH probe. The mother was also found to carry a duplication of a region including this FISH probe, with the majority of nuclei containing signals consistent with increased copy number for this clone (Figure 8.1. and Table 8.2.). Interphase FISH results were inconclusive for 4:1, with on average roughly equal numbers of nuclei counted as duplicated or not duplicated, approximately 40% in each category (Table 8.2.). These ambiguous data were from 3 separate experiments, but in two of these relatively few nuclei were scored (57 and 75) compared to the third experiment in which 100 nuclei were scored. Subsequent analysis on 4:1 by UPQFM-PCR has confirmed that she does not carry the duplication (see sections 8.4., 8.6. and Figure 8.5.).

8.3. Interphase FISH mapping of Xq27.1 duplication

Interphase FISH was carried out using additional genomic clones mapping to Xq27.1 on peripheral blood lymphocytes from the affected males in this family that had been demonstrated to have a duplication including the *SOX3* locus. This mapped the centromeric end of the duplication to between genomic clone bA35F15 and bA364B14 and the distal telomeric boundary of this duplicated region to within the genomic clone dJ595A18 (Table 8.1.). dJ595A18, a clone mapping between bA51C14 (duplicated) and dJ177G6 (not duplicated) gave ambiguous results, with one affected male having more duplicated signals and the other having more single-copy signals (Table 8.1.). Ambiguous interphase FISH results such as these are possible if the clone used as a FISH probe is only partially duplicated, so it was considered that the distal breakpoint might be within this clone.

8.4. Further mapping of duplication breakpoints in family 4 using UPQFM-PCR.

As UPQFM-PCR had been successful in localising breakpoint regions in PMD families, this technique was also used to further map the location of duplication endpoints in Xq27.1 for family 4. Pairs of tagged primers were designed to amplify short regions throughout the genomic sequence corresponding to those clones that were thought to be near the ends of the duplicated region. UPQFM-PCR was carried out on the two affected males and their carrier mother, as well as the maternal grandmother. UPQFM-PCR dosage quotients for the maternal grandmother were consistent with normal copy number of the target sequences throughout this region (data not shown).

8.4.1. UPQFM-PCR mapping of proximal duplication breakpoint

At the proximal end of the duplicated region, the UPQFM-PCR data was in concordance with what had been deduced from the interphase FISH results. All four UPQFM-PCR primer pairs mapping within bA364B14 were shown to be duplicated in 4:4, 4:6 and 4:7 (Figure 8.2., Table 8.3). A single primer pair mapping to the telomeric end of bA35F15 produced dosage quotient ratios consistent with this region being single-copy in this family (Figure 8.2., Table 8.3). These data localised the proximal duplication breakpoint in this family to a region of just over 20Kb, including the overlap between the sequences from these two genomic clones.

8.4.2. UPQFM-PCR mapping of distal duplication breakpoint

Around the distal end of the duplicated region most UPQFM-PCR results were similar in all three individuals carrying a duplication of Xq27.1. Four different primer pairs were used throughout the sequence from clone dJ595A18, and all had mean dosage quotients consistent with a duplication of the region (Figure 8.2., Table 8.3.). Two primer pairs were also used that were located within the adjacent distal clone XXyac-291B3. In the two affected males both of these primer pairs produced ratios consistent with these regions being single copy (Table 8.3.). For 4:4, the dosage quotient for the more proximal primer within this clone appeared duplicated when compared to the X-linked PLP1 primer pair, but did not give a high dosage quotient when compared to the autosomal primer pair CF (Table 8.3.). As this was only one elevated dosage quotient ratio, it is most likely that it is just a false positive, perhaps as a result of poor amplification of the PLP1 primer pair. Thus it seems most likely that the two sequences within XXyac-291B3 assayed for changes in dosage by UPQFM-PCR are not duplicated in this family, and the distal breakpoint is located somewhere within a 27.7Kb region including the overlap in sequence between the genomic clones dJ595A18 and XXyac-291B3 (Figure 8.2., Table 8.3.).

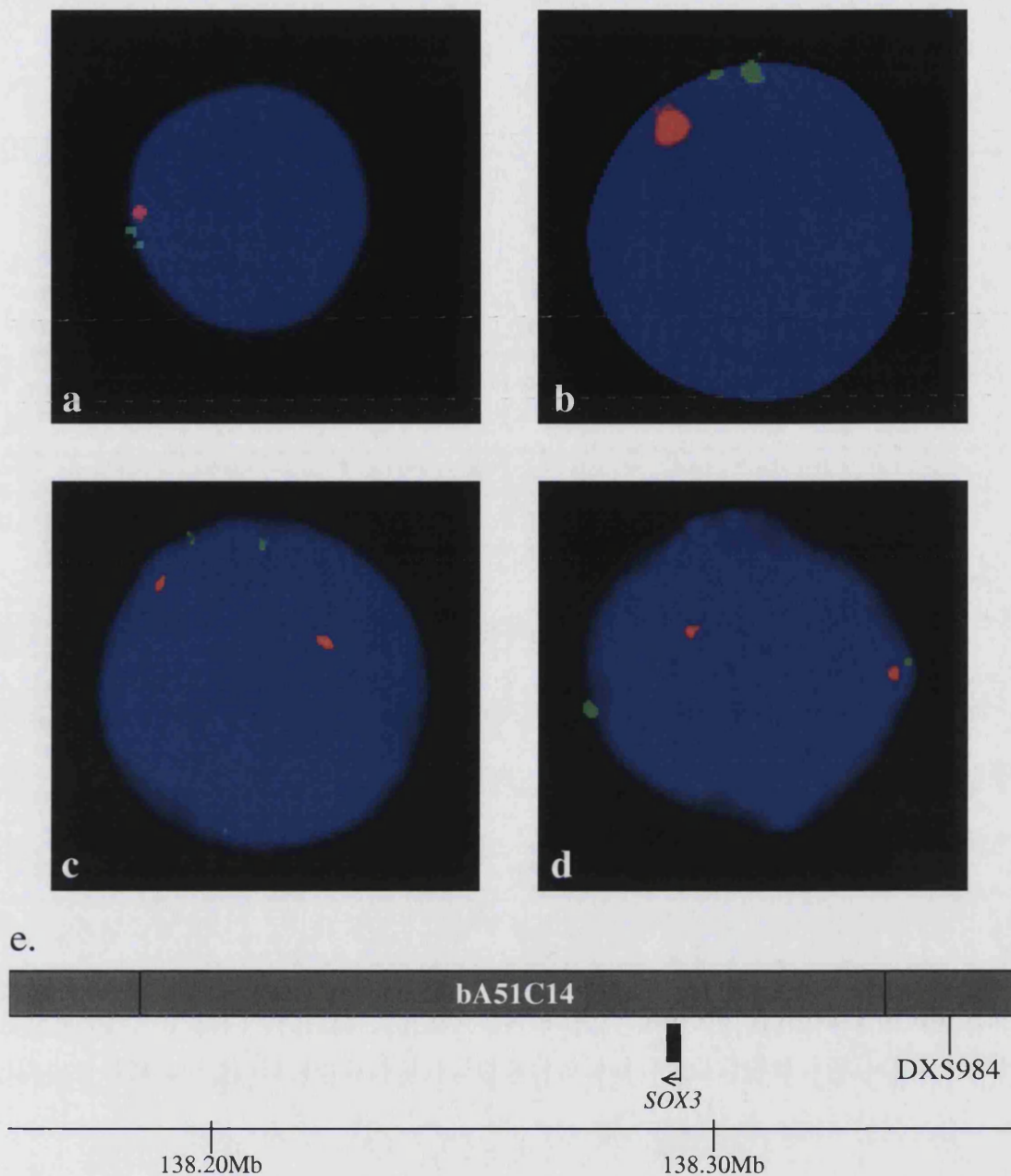


Figure 8.1. A duplication including the *SOX3* gene was detected in family 4 by interphase FISH. Nuclei from cultured peripheral blood lymphocytes are shown in a-d above, in all cases the *SOX3*-containing BAC, bA51C14 is labelled in green, while the X centromere is detected by a red fluorescent probe. (a.) is the younger affected boy, (4:7), (b.) is the older affected sibling (4:6), their carrier mother is in (c.) and an interphase nucleus from the maternal grandmother (4:1) is shown in (d.). Underneath, (e.) shows the location of bA51C14 in Xq27.1, position on the X chromosome is taken from the Ensembl genome browser, release 22.34d.1. The position of the *SOX3* gene within this clone, and a nearby polymorphic CA repeat microsatellite marker, DXS984 is also shown.





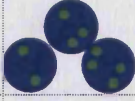
Clone	Individual	0,1	0,2	0,3	0,4	Other	Conclusions
							
bA35F15	4:6	78.43%	18.63%	0.98%	0	1.96%	Not duplicated
	4:7	71.58%	27.37%	1.05%	0	0	Not duplicated
bA364B14	4:6	35.24%	57.14%	6.67%	0	0.95%	Duplicated
bA51C14	4:6	30.77%	52.88%	10.58%	2.88%	2.88%	Duplicated
	4:7	36.12%	56.28%	5.61%	11.63%	1.99%	Duplicated
dJ595A18	4:6	41.82%	47.27%	5.45%	0	5.45%	Breakpoint?
	4:7	61.90%	34.29%	1.90%	0	1.90%	Not duplicated?
dJ177G6	4:6	75.00%	19.23%	0.96%	0	4.81%	Not duplicated
	4:7	84.52%	10.71%	0	0	4.76%	Not duplicated

Table 8.1. Interphase FISH scores, expressed as percentages, from the affected males in family 4, for clone bA51C14 and other clones that were close to the boundaries of the duplicated region. All scores are from a single slide. An average of 101 nuclei were scored per slide.









Clone	Individual	1,1 	1,2 	1,3 	1,4 	2,2 	2,3 	2,4 	Other 	Conclusions
bA51C14	(4:4)	11.00%	38.18%	8.57%	0.49%	9.52%	6.23%	1.91%	24.11%	Duplicated
	(4:1)	36.24%	35.81%	2.61%	0.44%	4.42%	0.58%	0.44%	17.24%	Not duplicated

Table 8.2. Interphase FISH scores, expressed as percentages, from females in family 4, for clone bA51C14. An average of 78 nuclei were scored per slide, the percentages shown are an average of two experiments for 4:4 and the average of three experiments for 4:1.

Clone	Position in clone of UPQFM primer pairs	4:7			4:6			4:4		
		Mean ratio of UPQFM PCR product compared to...		Number of experiments	Mean ratio of UPQFM PCR product compared to...		Number of experiments	Mean ratio of UPQFM PCR product compared to...		Number of experiments
		PLP1	CF		PLP1	CF		PLP1	CF	
bA35F15	66750-66949	1.01	1.19	4	0.95	1.14	1	0.99	1.09	2
bA364B14	16714-16959	3.00	2.93	1	2.79	2.81	1	1.40	1.29	1
bA364B14	25512-25756	2.22	2.28	5	2.46	2.87	1	1.36	1.56	2
bA364B14	48225-48392	2.21	2.20	5	1.95	2.21	1	1.36	1.53	2
bA364B14	74126-74369	2.00	2.36	4	1.92	2.31	1	1.47	1.63	2
dJ595A18	26757-27039	2.18	2.27	5	3.26	3.70	1	1.15	1.33	2
dJ595A18	54671-54825	2.15	2.15	5	2.04	2.31	1	1.31	1.48	2
dJ595A18	101906-102140	2.29	2.07	3	1.74	1.94	3	2.03	1.97	3
dJ595A18	114741-114935	2.21	1.98	3	1.65	1.77	3	1.87	1.80	3
XXyac-291B3	26828-27044	1.06	1.06	2	0.90	0.93	2	1.82	0.91	2
XXyac-291B3	64796-65011	1.43	1.07	1	0.76	-	1	0.67	1.03	1

Table 8.3. UPQFM ratios for the duplication breakpoints in family 4 (as compared against both *PLP1* exon 6 and CF control primers). Average ratios were taken from all experiments carried out that included that primer pair, and the number of experiments is indicated. Ratios consistent with duplication are highlighted in bold, and the zigzag lines show the assumed location of the breakpoint regions. The position of each target sequence within the relevant genomic clone is indicated.

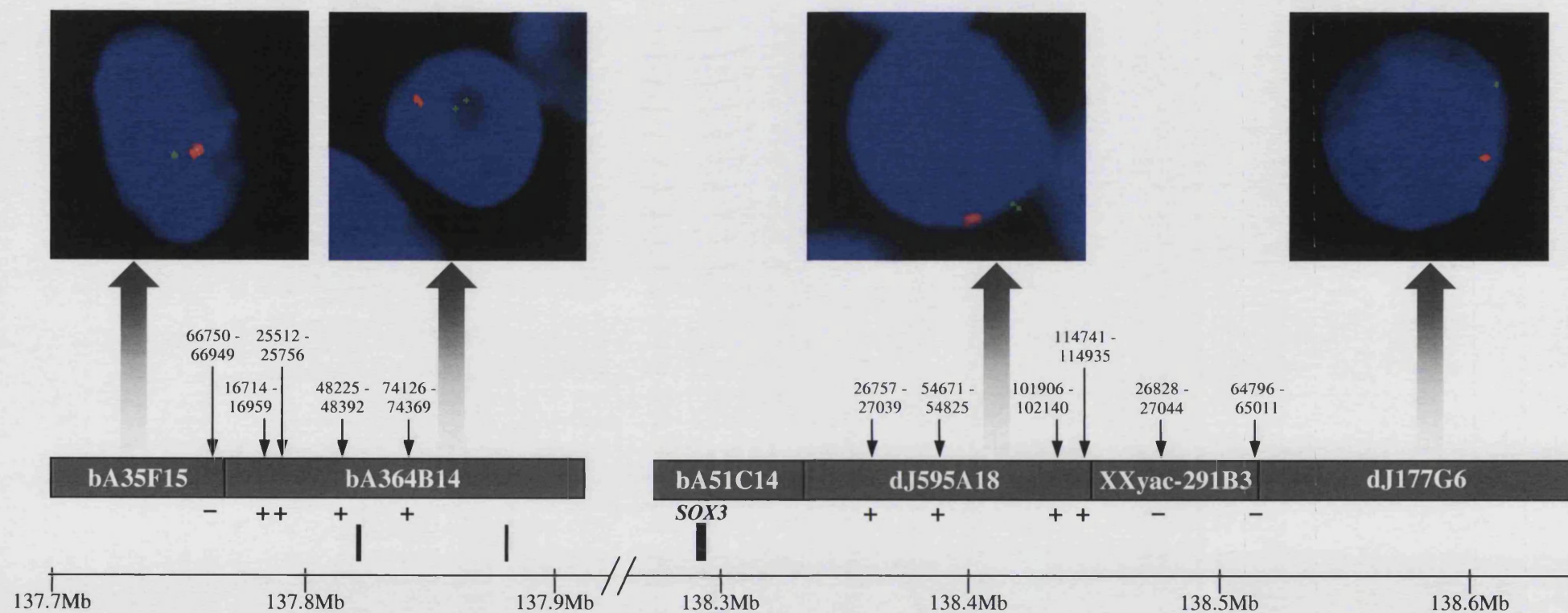


Figure 8.2. Diagram showing interphase FISH and UPQFM-PCR results around the proximal and distal duplication breakpoints in family 4.

Legend on next page.

Figure 8.2. Legend.

The genomic clones in the region are shown as grey boxes, and position on the X chromosome, from the Ensembl genome browser version 22.34d.1 is shown underneath the contig in megabases. Locations of UPQFM-PCR primers in the region are indicated by arrows, and the position within each clone for each pair of primers is shown. (+) indicates a sequence that appeared to be duplicated, (-) a sequence with normal copy number. Genes in the region are shown as black boxes underneath the contig. Representative nuclei from interphase FISH experiments with the various breakpoint region clones are shown above the contig, these are taken from experiments using either of the two affected males (364B14 with 4:6, others with 4:7). Genomic clones are labelled green and the X centromere probe is labelled red, as before.

8.5. Determining the nature of the Xq27.1 duplication by dual probe interphase FISH

The orientation of the duplicated region with respect to the original Xq27.1 sequence was investigated. Two different genomic clones that had both been shown to be within the duplicated region by interphase FISH (bA51C14, and a more proximal clone, bA189F12) were labelled by nick translation with either biotin (bA51C14) or digoxigenin (bA189F12) and detected by FITC- (bA51C14) or Texas Red-labelled (bA189F12) antibodies. The distribution of red and green signals within over 100 interphase nuclei were scored on a single slide (Table 8.4.). In more than one-third of nuclei scored, the signals were too close to resolve (Table 8.4). Three categories accounted for most of the other nuclei; those with an alternating red/green/red/green (or green/red/green/red) pattern were slightly more prevalent, followed by those with

a red/green/red pattern, then those with a green/red/green distribution of signals (Figure 8.3. and Table 8.4.). The remaining nuclei were scored as having a red/red/green distribution of signals (Table 8.4.). Based on these data, it was decided that the most likely orientation of the duplication would be in tandem. Although the red/green/red/green distribution of signals was only detected in a few more nuclei than some of the other patterns of signals that were seen, the red/green/red and green/red/green patterns were seen in similar numbers of nuclei (Table 8.4.). The other, relatively simple possible orientation for the duplication in family 4 is an inverted duplication, for which patterns corresponding to either red/(green/green)/red or green/(red/red)/green would be seen depending on the orientation of the inversion. The central double green or double red signals that would be expected for this type of rearrangement are given in brackets as they could well be too close to separate at this resolution. However, if the duplication truly was inverted with respect to the original copy of the sequence, one of the two inverted patterns would be expected to predominate over the other, depending upon the orientation of the inversion. A likely cause for the high number of pseudo-inverted signal patterns seen is failure of hybridisation or loss of signal of one of the outer red or green signals in the tandem array. Slightly more red/green/red signals were seen than the other way around, suggesting that the green signal was more likely to be lost in this experiment. The only other arrangement consistently seen on the slide was a red/red/green pattern, which could result from the loss of the internal green (bA189F12) signal in the tandem array (Table 8.4.).

8.6. Long-range PCR across the duplication breakpoint junction

As the duplication in family 4 was most likely in a tandem, head to tail orientation, a PCR strategy using several different primers around the breakpoint regions to amplify the junction sequence was attempted. 25mer oligonucleotides were used as long-range PCR primers, and were selected from the genomic sequence between the positions of the UPQFM-PCR primers that defined the two breakpoint regions (Figure 8.2.). Four primers were designed within the 20Kb proximal duplication breakpoint regions, and five primers were designed within the 27.7Kb breakpoint region at the distal end of the duplication. All possible combinations of one proximal and one distal primer were then tried, and two of the UPQFM-PCR primers were also used as long-range primers in these reactions. The UPQFM-PCR tagged primers that were used were one from each of the primer pairs nearest the breakpoint region that had been found to be duplicated, and which would prime DNA synthesis heading towards the breakpoint region, and were used in case the duplication junction was located relatively close to either of the UPQFM-PCR primers (16959R from clone bA364B14, and 114741F from clone dJ595A18) (Figures 8.2. and 8.4.). Just one of the 30 possible combinations of long-range primers was successful in amplifying a product from duplication-carrying individuals in family 4 that was not present when normal control DNA was used. A 4-5Kb product resulted when the combination of the proximal long-range PCR primer 71430R (bA35F15) and the distal UPQFM-PCR primer 114741F (dJ595A18) was used (Figures 8.2. and 8.4.). This product was only seen when using genomic DNA from the affected boys and their obligate carrier mother was used as a template, and not when genomic DNA from the grandmother (thought not to carry the Xq27.1 duplication) or unrelated normal controls was used (Figure 8.4.). To check that this PCR product did contain the duplication junction in this family, the bands were excised from the gel and the bands were extracted from

the gel (see section 2.2.1.4.4.). The eluted DNA (from 4:7) was then sequenced using either the proximal or distal primer (see section 2.2.4.). The sequence obtained using the proximal primer 71430R (mapping within bA35F15), matched sequence from this clone and similarly sequence produced using the distal primer matched sequence from the appropriate region of dJ595A18. Further sequencing of this junction fragment is ongoing.

A BLASTz comparison of 10Kb from around the approximate locations of the two duplication breakpoints in family 4 showed no large regions of sequence similarity between the two regions (data not shown). The only regions of similarity corresponded to the positions of *Alu* sequences near both ends of the duplication, as has been found with similar comparisons of breakpoints on PMD families (see Figures 4.11., 5.11. and 6.9.). At this stage, it seems likely that the sequences at the *SOX3* duplication breakpoint do not share much homology; although *Alu-Alu* mediated rearrangement is still a possibility.

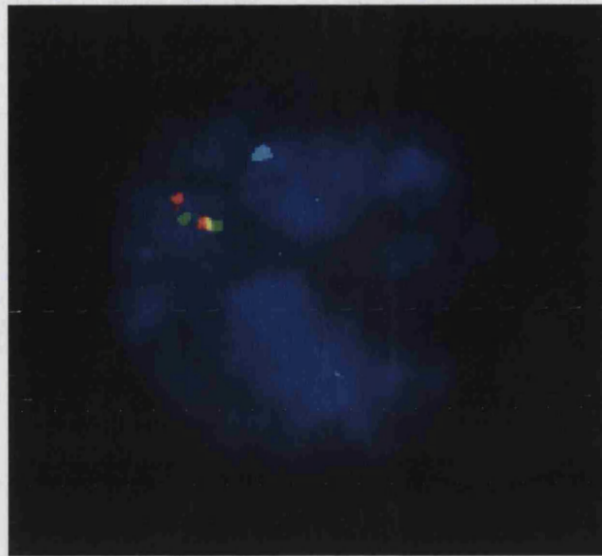


Figure 8.3. Interphase FISH using two genomic clones from within the duplication as probes indicates that the Xq27.1 duplication in family 4 is probably in a tandem orientation. An interphase nucleus from one of the affected males (4:7) is shown. The X centromere is detected by an Aqua-labelled probe, bA51C14 is red, and bA189F12 (a clone within the duplicated region, proximal to bA51C14) is green. A yellow signal is seen where the red and green signals overlap slightly.






Appearance of nuclei	Overlapping	R/G/R/G	G/R/G	R/G/R	R/R/G
					
% seen	34.91%	20.75%	16.98%	18.87%	8.49%

Table 8.4. Appearance of signals in interphase nuclei following hybridisation of two probes from within the duplication region in family 4. 106 nuclei were scored on one slide, and the percentage of nuclei within each category is shown. R = red signal, G = green signal.

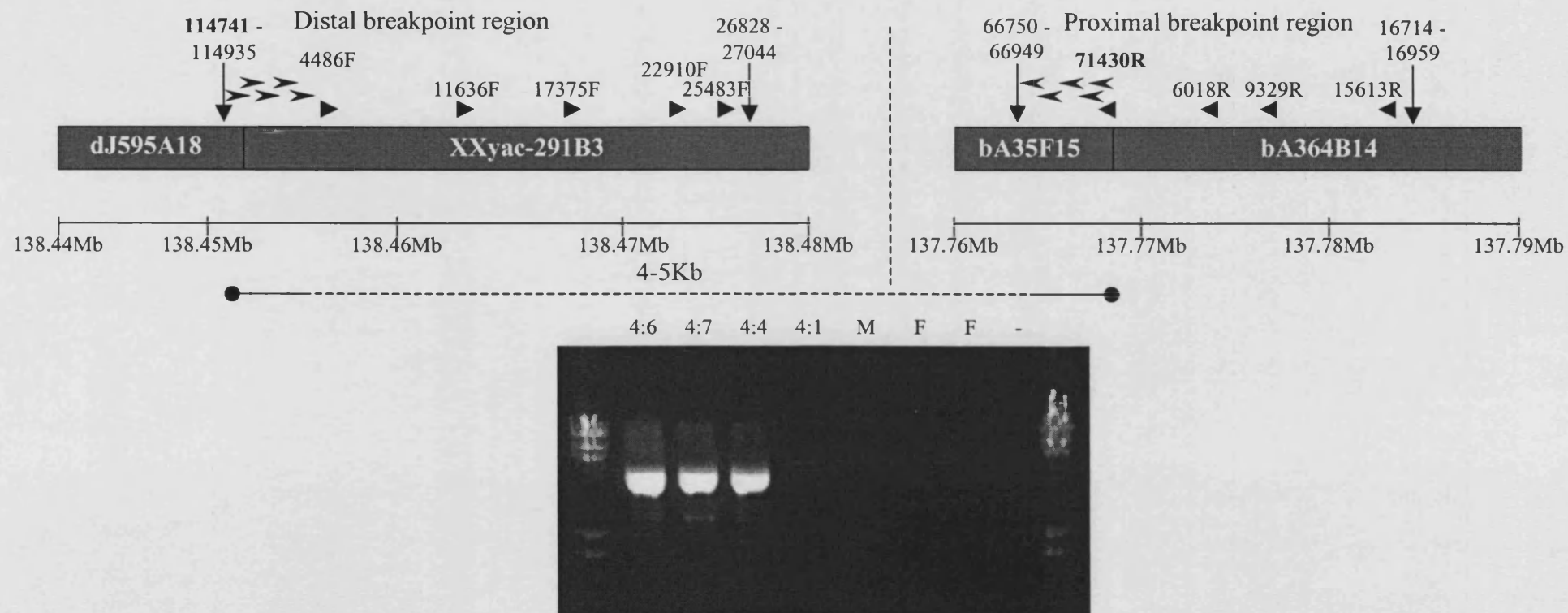


Figure 8.4. Long-range PCR across the tandem duplication breakpoint in family 4. Legend on next page.

Figure 8.4. Legend

40Kb from around the distal duplication breakpoint and 30Kb from around the proximal breakpoint is shown. Genomic clones in the region are represented by grey boxes, and the scale and location on the X chromosome is shown underneath the contig (taken from the Ensembl genome browser, version 22.34d.1.). Positions of various primers used to map the breakpoint are shown by arrows above the contig. Vertical arrows show the position of UPQFM-PCR primers, as previously shown in Figure 8.2., the location of long-range PCR primers are indicated by the larger labelled horizontal arrows. The smaller arrows show the positions of nested primers used for sequencing of the breakpoint fragment. An agarose gel is shown underneath the contig, showing the presence of a 4-5Kb band, using the combination of oligonucleotide primers highlighted in bold, seen in only the members of family 4 who had the duplication, and not in the maternal grandmother or unrelated male (M) and female (F) controls. λ HindIII was used as a size standard.

8.7. Determining the inheritance of the duplication in family 4

Microsatellite markers along the length of the X chromosome were typed in family 4 to determine the origin of the duplication in family 4. Genotyping was carried out using fluorescently labelled primers (FAM, TET or HEX) in a PCR, then the reactions were run on an ABI 377 automated DNA sequencer and results analysed using GENOTYPER software (Applied Biosystems). Three markers mapping close to the *SOX3* gene, within the duplicated region were typed; DXS1232, 305Kb proximal to *SOX3*, DXS8013, which is 99Kb proximal to the gene and DXS984, which is just 44Kb distal from *SOX3* (Figures 8.1. and 8.5.). Additional markers were chosen throughout the long arm of the X chromosome, and two markers near the Xp telomere were also used (Figure 8.5.). The two affected males shared the same alleles for

11/13 of the X chromosome markers examined, including the three markers within the duplicated region (Figure 8.5.). There was no evidence for heterozygosity for the three duplicated markers in either male. Only one marker from the three within the duplicated region, DXS984, was informative regarding the inheritance of this region of the X chromosome in this family. The allele present in the two affected males (176bp) was carried by their mother, but the maternal grandmother did not carry this allele of DXS984 (Figure 8.5.). It is likely that this allele was inherited from the maternal grandfather. This corresponds with the finding that the maternal grandmother does not carry the duplication. Although DNA from the maternal grandfather is not available, it is unlikely that he had a duplication of this region, as he is believed to be asymptomatic, and it is most probable that the Xq27.1 duplication in this family has arisen *de novo* in the mother of the affected boys. This duplication event could have occurred at several different stages, including during spermatogenesis in the maternal grandfather, early during development of the mother, or another possibility is that the maternal grandfather was a gonadal mosaic for this duplication mutation. Several recombination events were found to have occurred during the transmission of the X chromosome from the mother (4:4) to her sons. There has been a recombination event between DXS6797 (Xq22.3) and DXS984 (Xq27.1) during transmission of the X chromosome from the mother to both of her sons (Figure 8.5.). The location of this recombination event could not be localised further from these data, as the intervening markers used were not informative in this family. There had been an additional double recombination event in 4:7, as the two alleles amplified for the androgen receptor CAG repeat marker (Xq12) and DXS6800 (Xq21.1) appeared to have originated from the grandpaternal chromosome, whereas surrounding markers had alleles found on the grandmaternal X chromosome (Figure 8.5.).

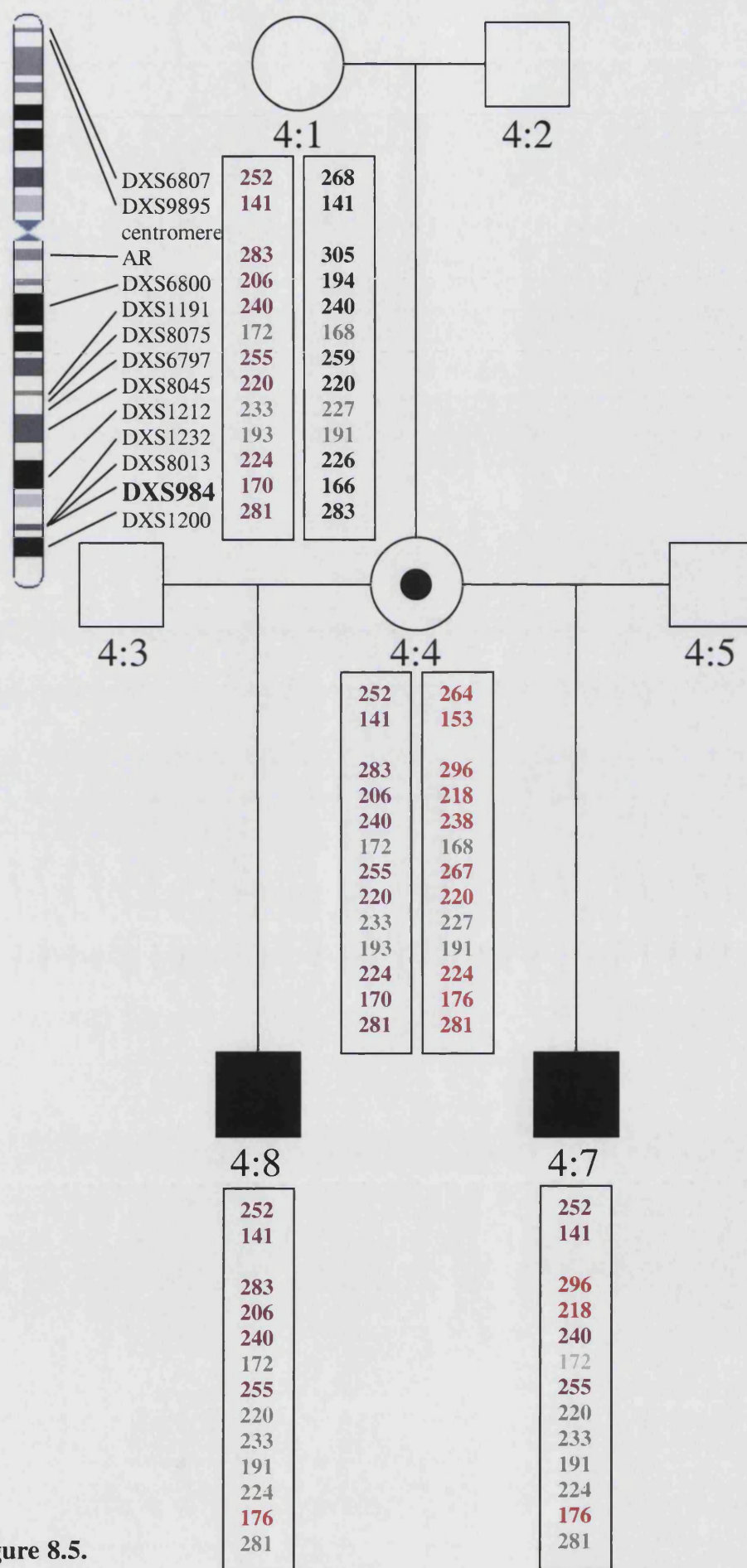


Figure 8.5.

Figure 8.5. Legend. Results of genotyping various X-linked polymorphic microsatellite markers in family 4. The pedigree is shown as in Figure 2.2. previously, and the allele sizes of the various microsatellites typed are shown underneath each individual, in the same order, Xp telomere going down to the Xq telomere, as they are found on the X chromosome. The ideogram in the top left-hand corner shows the approximate location of all the markers used on the X chromosome. The different alleles are colour coded according to which chromosome they were deduced to have originated from. The origin of alleles shaded in grey was unable to be determined due to lack of information about inheritance, mainly due to the lack of information on the genotype of the maternal grandfather, and the positioning of these alleles on one haplotype or the other is arbitrary in some cases. The DXS984 microsatellite marker is located just distal to *SOX3* (Figure 8.1.). AR is the polymorphic CAG repeat in the androgen receptor gene (Allen *et al.*, 1992).

8.8. Discussion

Increased dosage of a region including Xq27.1 and the *SOX3* gene has previously been associated with hypopituitary phenotypes (see section 1.6) (Lagerstrom-Fermer *et al.*, 1997; Hol *et al.*, 2000; Solomon *et al.*, 2002). During this study, another family with a duplication involving the *SOX3* gene has been identified. The duplication in family 4 has been shown in this study to include approximately 690Kb of genomic sequence, which is a much smaller region than the duplications of Xq that were previously reported to be associated with hypopituitarism (13Mb and 9Mb) (Lagerstrom-Fermer *et al.*, 1997; Hol *et al.*, 2000; Solomon *et al.*, 2002).

The duplicated region in family 4 only includes 3 genes, *SOX3* and two novel uncharacterised genes (Figure 8.2.). As a polyalanine tract expansion within *SOX3* has been associated with a phenotype including pituitary manifestations, and *SOX3* null mice have been shown to have abnormal pituitary development, increased dosage of *SOX3* during development is the most likely cause for the pituitary phenotype in this family (Laumonnier *et al.*, 2002; Rizzoti *et al.*, 2004). One of the other two genes in the region, the one closest to the proximal duplication breakpoint, contains an RNA binding domain and belongs to a family of heterogeneous nuclear riboproteins, and ESTs from this transcript have been isolated from a wide range of tissues, including brain and ovary libraries. The remaining gene in the duplicated region is also as yet uncharacterised and contains no recognised protein motifs, and ESTs for this gene have been found from tissues including brain and testis.

The phenotypes seen in the two affected males in this family are quite discordant, with the older sibling exhibiting isolated growth hormone deficiency, pituitary hypoplasia and dysgenesis of the corpus callosum, whereas his younger half-brother

has combined pituitary hormone deficiency and hypoplasia of the anterior pituitary, with no abnormalities of the corpus callosum. As they have different fathers, the divergence of the phenotypes is good evidence for the importance of genetic background effects in the development of pituitary disorders, and modifier loci may exist. However, this variability in phenotype may also be due to incomplete penetrance of the effects of *SOX3* duplication, as studies using *SOX3* null mice have reported a variable phenotype, with up to one-third of males showing no apparent phenotype, and a range of defects seen in those mice with a phenotype (Rizzoti *et al.*, 2004).

The duplication characterised in this study in Xq27.1 does not appear to be flanked by LCRs, as has been found to be the case with many genomic disorders, and in this aspect this duplication is similar to the duplications studied in PMD families (see section 1.4.1.). The Xq27.1 duplication may not share any homology between sequences at the duplication breakpoints, is tandem in orientation and is likely to have arisen on the grandpaternal X chromosome. It is interesting to note that duplications of *PLP1* in PMD families are thought to predominantly arise in male germ cells, and it is tempting to speculate that duplications involving Xq22.2 or Xq27.1 could arise by a similar mechanism during spermatogenesis, although many more families with duplications of this region would have to be studied to determine this (Mimault *et al.*, 1999).

8.9. Conclusions

This study has provided further evidence that *SOX3* is part of the small but increasing group of genes recognised to be sensitive to an increase in gene dosage. Interphase FISH has proved to be a reliable technique for screening for *SOX3* gene duplications. Although the tandem duplication breakpoint in family 4 has not yet been characterised at the sequence level, initial investigations suggest that the Xq26.2 duplication may have occurred by a similar mechanism to *PLP1* duplications in Xq22

9.0. DISCUSSION

9.1. Gene dosage

Gene dosage is increasingly being recognised as an important factor in human genetic disease (see section 1.4.). *PLP1* has been known to be dosage sensitive for the past ten years, and now more genes, such as *SOX3*, are also being recognised as such (Ellis and Malcolm, 1994). Testing for changes in gene dosage may sometimes be neglected during the process of mutation screening, which is often based on PCR and direct sequencing. The advent of new methodologies for measuring gene dosage, such as MAPH, may ensure that changes in sequence copy number are more easily detected. Technologies that can be used to assess genome-wide gene dosage, such as array CGH, will be particularly useful for detecting previously unknown and possibly pathogenic changes in gene dosage throughout the genome (Ishkanian *et al.*, 2004).

9.2. MAPH for detecting gene dosage

MAPH is potentially a very powerful technique for the rapid detection of copy number changes, and this study has shown that it can detect dosage changes in and around *PLP1*. However, the probe set that has been developed so far has proved to be too unreliable for diagnostic use, with many probes not producing consistent results. Further refinements to the experimental procedure, the redesigning of some probes and the addition of other probes (particularly for the other *PLP1* exons and extra control probes) may make the technique more suited for diagnostic use (see section 7.5.). However, as an MLPA kit including all the exons of the *PLP1* gene is now commercially available, it may be more cost effective to use this method to assay dosage of the gene, instead of making further modifications to the MAPH probe set.

9.3. The importance of duplications in evolution

As well as making a substantial contribution to genetic disorders in the present, ancient duplication events have been important in shaping the evolution of all genomes. Duplication events can range between that of whole genomes to duplication of a single nucleotide. Segmental duplications (>90% sequence identity, >10Kb in length) account for approximately 5% of the human genome sequence (Lander *et al.*, 2001). These segmental duplications represent relatively recent duplication events, and remnants of older duplication events can also be detected in the genomic sequence (Venter *et al.*, 2001). Segmental duplications are also associated with the process of chromosome and karyotype evolution and are frequently found at the boundaries of syntenic blocks in the genome (Bailey *et al.*, 2004). Chromosome rearrangements, possibly associated with segmental duplications, can be important in driving speciation events (O'Brien and Stanyon, 1999; Samonte and Eichler, 2002; Capanna and Castiglia, 2004). Gene duplication is important in the evolution of new proteins, and may be a primary source of innovation in the proteome. Members of functionally redundant pairs of duplicated genes may thus be freed from selective constraints to evolve new functions, unless one member of the pair accumulates deleterious mutations and becomes a pseudogene (Ohno, 1970, reviewed in Prince and Pickett, 2002). It has been shown that some genes within recent segmental duplications exhibit an accelerated pattern of evolution, probably driven by positive selection (Samonte and Eichler, 2002).

9.4. Duplications involving the *PLP1* genomic region

From examining the human genomic sequence around *PLP1* it appears likely that duplications of various parts of this region have occurred in the past, leaving behind a legacy of repeated sequences. Some of these repetitive sequences may also be contributing to the generation of novel duplications in the region today, such as the duplications involving *PLP1*, which are detected because they cause a recognisable

phenotype. However, it is quite possible that these local repetitive regions near *PLP1* are not directly involved in the generation of duplications in the region, and other factors may be influencing these rearrangements. In all species where the syntenic region of the genome was examined, each had a unique organisation of repetitive sequences within the region (see section 3.6.). Some of the repeated sequences distal to *PLP1* are probably unique to the primate lineage and may not even be present in our closest relatives, the chimpanzee (see section 3.6.1.). Duplications of a size comparable to the *PLP1* duplications may be occurring throughout the genome at a similar rate by the same mechanisms. It is only when a dosage-sensitive gene is involved that these events come to our attention, and studying rearrangements including *PLP1* can provide a valuable system to increase our understanding of these types of rearrangements.

9.5. Possible clustering of breakpoints within clone dJ1055C14

Two duplication breakpoints characterised during the course of this study were found to lie relatively close to each other within the same genomic clone dJ1055C14, and both are actually located within intronic sequences of the same gene, *MORF4L2* (Figure 9.1.). Both these duplications, in family 2 and family 3, are unusual rearrangements, not the tandem duplications that are found in most PMD families with *PLP1* gene duplication (Woodward *et al.*, in preparation). A few other breakpoints in PMD families have been sequenced within this clone, including 3 tandem duplications and one deletion (Woodward *et al.*, in preparation; Inoue *et al.*, 2002). Tandem duplication breakpoints proximal to *PLP1* are scattered throughout Xq22.1-2 (Woodward *et al.*, in preparation). The closest proximal duplication breakpoints to *PLP1* are all mapped to within sequence from clone dJ1055C4 (Woodward *et al.*, in preparation). As over 100Kb of sequence originates from this clone, it may not be surprising to find a substantial number of breakpoints scattered within this sequence, if breakpoint locations are assumed to be random. However, these six breakpoints are all located in a 28Kb region in the distal half

of dJ1055C14, and if just the duplication breakpoints are considered, these cluster in a ~15Kb region centred on the *MORF4L2* gene (Figure 9.1.). The 5Kb region around the proximal duplication breakpoint in family 3 contains a human replication origin consensus and this breakpoint is roughly in the middle of all 6 breakpoints (Figure 9.1.). The possible presence of an origin of replication in this region could be predisposing to replication-related rearrangements, creating a hotspot for duplications and deletions, triggered by DSBs (see section 6.4.2.). The *MORF4L2* gene has been reported to be widely expressed in different tissues, and another possibility is that transcription of this gene could be indirectly involved in the clustering of some breakpoints in this region, by creating a more open chromatin structure that may be more susceptible to DSBs (Bertram and Pereira-Smith, 2001).

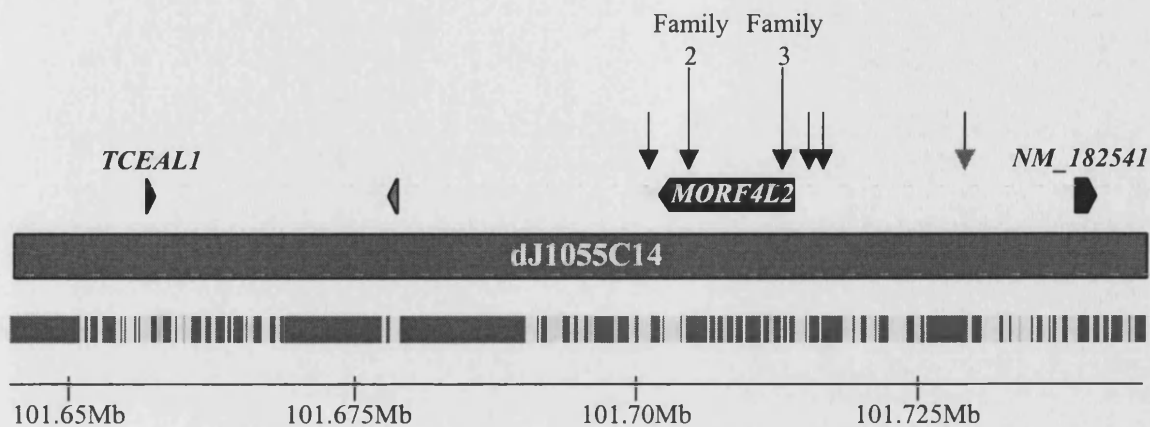


Figure 9.1. Location of sequenced breakpoints within clone dJ1055C14. The whole of clone dJ1055C14 is shown, adapted from the Ensembl genome browser. The scale bar shows the distance from the start of the X chromosome genomic sequence at the Xp telomere (Ensembl version 22.34d.1). Genes within this clone are shown by the black horizontal arrows, and a pseudogene in the region is shown as a grey arrow. Interspersed repeat content (as shown in Ensembl) is represented by the grey areas above the scale bar. The vertical arrows show the relative positions of sequenced breakpoints, duplication breakpoints are black, and a deletion breakpoint is shaded grey. The two breakpoints that have been sequenced in families 2 and 3 are labelled, the other three duplication breakpoints are tandem duplications and taken from Woodward *et al.* (in preparation). The deletion breakpoint sequence was reported by Inoue *et al.* (2002).

9.6. Mechanisms for gene duplication

Various different mechanisms that could be causing the genomic duplications described in this thesis have been postulated (see sections 4.12.2.2., 4.12.4-5., 5.8.2.5., 5.8.2.6. and 6.4.5.2.). The duplications described in PMD cases are definitely not caused by NAHR, as has been shown to be the case with many genomic disorders, but appear to have been mediated by non-homologous end-joining processes. Perhaps the most promising mechanism involves the repair of a double strand break by a homologous strand invasion event priming replication, followed by non-homologous rejoining of the newly synthesised strand to the other end of the DSB (Figure 4.16.). A mechanism of this sort can be used to explain all the breakpoints sequenced in this study, but cannot be proven at this point. Experimental evidence increasingly supports a substantial role for coupled homologous and non-homologous repair of DSBs in many systems (Morrow *et al.*, 1997; Richardson and Jasin, 2000; Michel, 2000; Johnson and Jasin, 2000; Kraus *et al.*, 2001). This type of aberrant repair process can easily be invoked in the generation of tandem duplications, such as described in family 1, but for more complicated rearrangements, such as have been partially characterised in families 2 and 3, this model may not be sufficient. A process involving capture of sequences into DSBs during the repair process may be important for generating these more complicated rearrangements, (see sections 5.8.2.6. and 6.4.5.7.).

Although rearrangements involving Xq22 are not simply caused by NAHR between LCRs, the association of these regions, particularly the highly homologous distal LCRs, with some of the breakpoints sequenced in this and other studies shows that they may be involved in the rearrangement in some other way (Inoue *et al.*, 2002; Woodward *et al.*, in preparation). The presence of the distal LCRs, may contribute to genomic instability in the region, and stimulate rearrangements, possibly by interactions between the repeat sequences, or by an unusual chromatin organisation. Alternatively, a sequence element

within the LCRs could initiate duplication events in this region, possibly by being prone to DSBs, and this could have also created the duplicated copies of the LCRs themselves. A complicated local repeat structure was also noted in Xq26.2 near the site of integration of an ectopic *PLP1*-containing duplicated region in family 3, and the presence of these repeats may be related to genomic instability in this region (see section 6.4.5.1.)

Studies involving *PLP1*, *SOX3* and some other dosage-sensitive genes in the human genome, and investigations of mechanisms of DNA rearrangement/DSB repair, are starting to produce an understanding of the mechanisms and significance of non-recurrent genomic duplication events. In contrast to the well-studied mechanisms of NAHR that mediate “genomic” disorders, non-homologous mechanisms of rearrangement are less well understood. Genomic disorders that are caused by NAHR between LCRs, such as CMT1A, VCFS and SMS, are more commonly seen than PMD by at least an order of magnitude. However, the highly homologous LCRs that predispose to particular genomic rearrangements may have originated by a similar non-homologous mechanism to the PMD duplications, particularly the duplication and insertion event described in family 3.

Although duplication and other rearrangement events involving non-homologous sequences at a one locus may be rarer than those caused by NAHR in a susceptible region, when the incidence of all such events at all loci in the genome is considered, such events may have a substantial but largely unrecognised effect on human health, disease, normal variation and evolution. Many genetic and congenital conditions have an as yet unknown aetiology and genomic copy number changes may be important in a number of these. Increased awareness of the importance of copy number changes and the availability of several methods to assess such changes should lead to more dosage-sensitive genes being recognised in the future.

9.7. Future work

Further study of the *in vivo* characteristics of the *PLP1* genomic region will provide more insight into the processes that may trigger rearrangements involving Xq22. Potentially interesting aspects of the organisation of this region, considered during the course of this study, and which it would be of interest to verify and investigate experimentally include matrix attachment regions, origins of replication and recombination hotspots.

The rearrangements in families 2 and 3 are not yet fully characterised, and elucidation of the sequences involved in these other breakpoints will enable a fuller understanding of the processes leading to these unusual rearrangements.

Completion of the sequencing of the duplication breakpoint in family 4 will be necessary to understand the exact nature of the rearrangement in this family. Screening of further patients with pituitary dysfunction may identify more families with *SOX3* duplications, and further our knowledge of the mechanisms and prevalence of duplications in this region. The development of an animal model of *Sox3* overdosage will be important in advancing understanding of the pathogenesis of increased copy number at this locus.

Pituitary development is sensitive to changes in gene dosage of *SOX3*, which is just one of numerous transcription factors and signalling molecules involved in this complex process. It is possible that increased dosage of other genes involved in pituitary development may also be pathogenic, and investigation of other genes known to be active during pituitary development may prove to be a productive line of investigation.

REFERENCES

- Aarskog,N.K., Vedeler,C.A. (2000). Real-time quantitative polymerase chain reaction. A new method that detects both the peripheral myelin protein 22 duplication in Charcot-Marie-Tooth type 1A disease and the peripheral myelin protein 22 deletion in hereditary neuropathy with liability to pressure palsies. *Hum.Genet.* 107, 494-498.
- Abeyasinghe,S.S., Chuzhanova,N., Krawczak,M., Ball,E.V., Cooper,D.N. (2003). Translocation and gross deletion breakpoints in human inherited disease and cancer I: Nucleotide composition and recombination-associated motifs. *Hum.Mutat.* 22, 229-244.
- Agundez,J.A., Ledesma,M.C., Ladero,J.M., Benitez,J. (1995). Prevalence of CYP2D6 gene duplication and its repercussion on the oxidative phenotype in a white population. *Clin.Pharmacol.Ther.* 57, 265-269.
- Akira,S., Okazaki,K., Sakano,H. (1987). Two pairs of recombination signals are sufficient to cause immunoglobulin V-(D)-J joining. *Science* 238, 1134-1138.
- Aklillu,E., Persson,I., Bertilsson,L., Johansson,I., Rodrigues,F., Ingelman-Sundberg,M. (1996). Frequent distribution of ultrarapid metabolizers of debrisoquine in an ethiopian population carrying duplicated and multiduplicated functional CYP2D6 alleles. *J.Pharmacol.Exp.Ther.* 278, 441-446.
- Akrami,S.M., Rowland,J.S., Taylor,G.R., Armour,J.A. (2003). Diagnosis of gene dosage alterations at the PMP22 gene using MAPH. *J.Med.Genet.* 40, e123.
- Akrami,S.M., Winter,R.M., Brook,J.D., Armour,J.A. (2001). Detection of a large TBX5 deletion in a family with Holt-Oram syndrome. *J Med Genet* 38, E44.
- Alitalo,K., Schwab,M., Lin,C.C., Varmus,H.E., Bishop,J.M. (1983). Homogeneously staining chromosomal regions contain amplified copies of an abundantly expressed cellular oncogene (c-myc) in malignant neuroendocrine cells from a human colon carcinoma. *Proc.Natl.Acad.Sci.U.S.A* 80, 1707-1711.
- Allen,C., Miller,C.A., Nickoloff,J.A. (2003). The mutagenic potential of a single DNA double-strand break in a mammalian chromosome is not influenced by transcription. *DNA Repair (Amst)* 2, 1147-1156.
- Altschul,S.F., Gish,W., Miller,W., Myers,E.W., Lipman,D.J. (1990). Basic local alignment search tool. *J Mol.Biol.* 215, 403-410.
- Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389-3402.
- Amos-Landgraf,J.M., Ji,Y., Gottlieb,W., Depinet,T., Wandstrat,A.E., Cassidy,S.B., Driscoll,D.J., Rogan,P.K., Schwartz,S., Nicholls,R.D. (1999). Chromosome breakage in the Prader-Willi and Angelman syndromes involves recombination between large, transcribed repeats at proximal and distal breakpoints. *Am.J.Hum.Genet.* 65, 370-386.
- Anderson,T.J., Klugmann,M., Thomson,C.E., Schneider,A., Readhead,C., Nave,K.A., Griffiths,I.R. (1999). Distinct phenotypes associated with increasing dosage of the PLP

gene: implications for CMT1A due to PMP22 gene duplication. *Ann N Y Acad Sci* 883, 234-46.

Aoki,K., Suzuki,K., Sugano,T., Tasaka,T., Nakahara,K., Kuge,O., Omori,A., Kasai,M. (1995). A novel gene, Translin, encodes a recombination hotspot binding protein associated with chromosomal translocations. *Nat.Genet* 10, 167-174.

Aoyagi,Y., Kobayashi,H., Tanaka,K., Ozawa,T., Nitta,H., Tsuji,S. (1999). A de novo splice donor site mutation causes in-frame deletion of 14 amino acids in the proteolipid protein in Pelizaeus-Merzbacher disease. *Ann Neurol* 46, 112-5.

Arcot,S.S., Wang,Z., Weber,J.L., Deininger,P.L., Batzer,M.A. (1995). Alu repeats: a source for the genesis of primate microsatellites. *Genomics* 29, 136-144.

Armour,J.A., Sismani,C., Patsalis,P.C., Cross,G. (2000). Measurement of locus copy number by hybridisation with amplifiable probes. *Nucleic Acids Res* 28, 605-9.

Arnott,S., Chandrasekaran,R., Hall,I.H., Puigjaner,L.C. (1983). Heteronomous DNA. *Nucleic Acids Res.* 11, 4141-4155.

Ashley,T., Plug,A.W., Xu,J., Solari,A.J., Reddy,G., Golub,E.I., Ward,D.C. (1995). Dynamic changes in Rad51 distribution on chromatin during meiosis in male and female vertebrates. *Chromosoma* 104, 19-28.

Babcock,M., Pavlicek,A., Spiteri,E., Kashork,C.D., Ioshikhes,I., Shaffer,L.G., Jurka,J., Morrow,B.E. (2003). Shuffling of genes within low-copy repeats on 22q11 (LCR22) by Alu-mediated recombination events during evolution. *Genome Res.* 13, 2519-2532.

Babenko,V.N., Kosarev,P.S., Bazin,V.V., Frolov,A.S. (1999). [Repeating sequences in promoter regions of eukaryotic genes]. *Biofizika* 44, 664-667.

Bailey,J.A., Baertsch,R., Kent,W.J., Haussler,D., Eichler,E.E. (2004). Hotspots of mammalian chromosomal evolution. *Genome Biol.* 5, R23.

Bailey,J.A., Gu,Z., Clark,R.A., Reinert,K., Samonte,R.V., Schwartz,S., Adams,M.D., Myers,E.W., Li,P.W., Eichler,E.E. (2002). Recent segmental duplications in the human genome. *Science* 297, 1003-7.

Bailey,J.A., Liu,G., Eichler,E.E. (2003). An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet* 73, 823-834.

Balmain,A. (2002). Cancer: new-age tumour suppressors. *Nature* 417, 235-237.

Balmain,A., Gray,J., Ponder,B. (2003). The genetics and genomics of cancer. *Nat.Genet* 33 *Suppl*, 238-244.

Baudat,F., Nicolas,A. (1997). Clustering of meiotic double-strand breaks on yeast chromosome III. *Proc.Natl.Acad.Sci.U.S.A* 94, 5213-5218.

Baumann,G. (2001). Growth Hormone and Its Disorders. In: *Principles and Practice of Endocrinology and Metabolism*, ed. K.L.Becker Lippincott, Williams & Wilkins, 129-145.

Baumann,N., Pham-Dinh,D. (2001). Biology of oligodendrocyte and myelin in the mammalian central nervous system. *Physiol Rev* 81, 871-927.

Bayes,M., Magano,L.F., Rivera,N., Flores,R., Perez Jurado,L.A. (2003). Mutational mechanisms of Williams-Beuren syndrome deletions. *Am J Hum Genet* 73, 131-51.

Been,M.D., Burgess,R.R., Champoux,J.J. (1984). Nucleotide sequence preference at rat liver and wheat germ type 1 DNA topoisomerase breakage sites in duplex SV40 DNA. *Nucleic Acids Res.* 12, 3097-3114.

Bensasson,D., Feldman,M.W., Petrov,D.A. (2003). Rates of DNA duplication and mitochondrial DNA insertion in the human genome. *J Mol.Evol.* 57, 343-354.

Benson,G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573-580.

Bergemann,A.D., Johnson,E.M. (1992). The HeLa Pur factor binds single-stranded DNA at a specific element conserved in gene flanking regions and origins of DNA replication. *Mol.Cell Biol.* 12, 1257-1265.

Bergemann,A.D., Ma,Z.W., Johnson,E.M. (1992). Sequence of cDNA comprising the human pur gene and sequence-specific single-stranded-DNA-binding properties of the encoded protein. *Mol.Cell Biol.* 12, 5673-5682.

Bernard,P., Gabant,P., Bahassi,E.M., Couturier,M. (1994). Positive-selection vectors using the F plasmid ccdB killer gene. *Gene* 148, 71-74.

Bernues,J., Beltran,R., Casasnovas,J.M., Azorin,F. (1990). DNA-sequence and metal-ion specificity of the formation of ³H-DNA. *Nucleic Acids Res.* 18, 4067-4073.

Bertram,M.J., Pereira-Smith,O.M. (2001). Conservation of the MORF4 related gene family: identification of a new chromo domain subfamily and novel protein motif. *Gene* 266, 111-21.

Birney,E., Andrews,D., Bevan,P., Caccamo,M., Cameron,G., Chen,Y., Clarke,L., Coates,G., Cox,T., Cuff,J., Curwen,V., Cutts,T., Down,T., Durbin,R., Eyraes,E., Fernandez-Suarez,X.M., Gane,P., Gibbins,B., Gilbert,J., Hammond,M., Hotz,H., Iyer,V., Kahari,A., Jekosch,K., Kasprzyk,A., Keefe,D., Keenan,S., Lehtsalaiho,H., McVicker,G., Melsopp,C., Meidl,P., Mongin,E., Pettett,R., Potter,S., Proctor,G., Rae,M., Searle,S., Slater,G., Smedley,D., Smith,J., Spooner,W., Stabenau,A., Stalker,J., Storey,R., Ureta-Vidal,A., Woodwark,C., Clamp,M., Hubbard,T. (2004). Ensembl 2004. *Nucleic Acids Res.* 32 Database issue, D468-D470.

Bizzozero,O.A., Malkoski,S.P., Mobarak,C., Bixler,H.A., Evans,J.E. (2002). Mass-spectrometric analysis of myelin proteolipids reveals new features of this family of palmitoylated membrane proteins. *J Neurochem* 81, 636-45.

Bock,J.B., Matern,H.T., Peden,A.A., Scheller,R.H. (2001). A genomic perspective on membrane compartment organization. *Nature* 409, 839-841.

Boehm,T., Mengle-Gaw,L., Kees,U.R., Spurr,N., Lavenir,I., Forster,A., Rabbitts,T.H. (1989). Alternating purine-pyrimidine tracts may promote chromosomal translocations seen in a variety of human lymphoid tumours. *EMBO J* 8, 2621-2631.

Boison,D., Stoffel,W. (1989). Myelin-deficient rat: a point mutation in exon III (A----C, Thr75----Pro) of the myelin proteolipid protein causes dysmyelination and oligodendrocyte death. *EMBO J* 8, 3295-302.

Boison,D., Stoffel,W. (1994). Disruption of the compacted myelin sheath of axons of the central nervous system in proteolipid protein-deficient mice. *Proc Natl Acad Sci U S A* 91, 11709-13.

Bongarzone,E.R., Campagnoni,C.W., Kampf,K., Jacobs,E.C., Handley,V.W., Schonmann,V., Campagnoni,A.T. (1999). Identification of a new exon in the myelin proteolipid protein gene encoding novel protein isoforms that are restricted to the somata of oligodendrocytes and neurons. *J Neurosci* 19, 8349-57.

Bongarzone,E.R., Jacobs,E., Schonmann,V., Campagnoni,A.T. (2001). Classic and soma-restricted proteolipids are targeted to different subcellular compartments in oligodendrocytes. *J.Neurosci.Res.* 65, 477-484.

Bowles,J., Schepers,G., Koopman,P. (2000). Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators. *Dev.Biol.* 227, 239-255.

Bremner,W.J., Huhtaniemi,I., Amory,J.K. (2001). Pituitary Gonadotropins and Their Disorders. In: *Principles and Practice of Endocrinology and Metabolism*, ed. K.L.Becker Lippincott, Williams & Wilkins, 170-177.

Broach,J.R., Li,Y.Y., Feldman,J., Jayaram,M., Abraham,J., Nasmyth,K.A., Hicks,J.B. (1983). Localization and sequence analysis of yeast origins of DNA replication. *Cold Spring Harb.Symp.Quant.Biol.* 47 Pt 2, 1165-1173.

Brown,A.L., Kay,G.F. (1999). Bex1, a gene with increased expression in parthenogenetic embryos, is a member of a novel gene family on the mouse X chromosome. *Hum Mol.Genet* 8, 611-619.

Bullock,P., Miller,J., Botchan,M. (1986). Effects of poly[d(pGpT).d(pApC)] and poly[d(pCpG).d(pCpG)] repeats on homologous recombination in somatic cells. *Mol.Cell Biol.* 6, 3948-3953.

Cailloux,F., Gauthier-Barichard,F., Mimault,C., Isabelle,V., Courtois,V., Giraud,G., Dastugue,B., Boespflug-Tanguy,O. (2000). Genotype-phenotype correlation in inherited brain myelination defects due to proteolipid protein gene mutations. *Clinical European Network on Brain Dysmyelinating Disease. Eur J Hum Genet* 8, 837-45.

Campagnoni,C.W., Garbay,B., Micevych,P., Pribyl,T., Kampf,K., Handley,V.W., Campagnoni,A.T. (1992). DM20 mRNA splice product of the myelin proteolipid protein gene is expressed in the murine heart. *J.Neurosci.Res.* 33, 148-155.

Capanna,E., Castiglia,R. (2004). Chromosomes and speciation in *Mus musculus domesticus*. *Cytogenet.Genome Res.* 105, 375-384.

Carango,P., Funanage,V.L., Quiros,R.E., Debruyne,C.S., Marks,H.G. (1995). Overexpression of DM20 messenger RNA in two brothers with Pelizaeus- Merzbacher disease. *Ann Neurol* 38, 610-7.

Chance,P.F., Abbas,N., Lensch,M.W., Pentao,L., Roa,B.B., Patel,P.I., Lupski,J.R. (1994). Two autosomal dominant neuropathies result from reciprocal DNA duplication/deletion of a region on chromosome 17. *Hum Mol Genet* 3, 223-8.

Chandley,A.C. (1991). On the parental origin of de novo mutation in man. *J Med.Genet* 28, 217-223.

Chen,K.S., Manian,P., Koeuth,T., Potocki,L., Zhao,Q., Chinault,A.C., Lee,C.C., Lupski,J.R. (1997). Homologous recombination of a flanking repeat gene cluster is a mechanism for a common contiguous gene deletion syndrome. *Nat.Genet.* 17, 154-163.

Chishti,A.H., Kim,A.C., Marfatia,S.M., Lutchman,M., Hanspal,M., Jindal,H., Liu,S.C., Low,P.S., Rouleau,G.A., Mohandas,N., Chasis,J.A., Conboy,J.G., Gascard,P., Takakuwa,Y., Huang,S.C., Benz,E.J., Jr., Bretscher,A., Fehon,R.G., Gusella,J.F., Ramesh,V., Solomon,F., Marchesi,V.T., Tsukita,S., Tsukita,S., Hoover,K.B., . (1998). The FERM domain: a unique module involved in the linkage of cytoplasmic proteins to the membrane. *Trends Biochem.Sci.* 23, 281-282.

Chuzhanova,N., Abeysinghe,S.S., Krawczak,M., Cooper,D.N. (2003). Translocation and gross deletion breakpoints in human inherited disease and cancer II: Potential involvement of repetitive sequence elements in secondary structure formation between DNA ends. *Hum.Mutat.* 22, 245-251.

Cohen,J.L. (2001). Thyroid-Stimulating Hormone and Its Disorders. In: *Principles and Practice of Endocrinology and Metabolism*, ed. K.L.Becker Lippincott, Williams & Wilkins, 159-170.

Cohen,L.E., Radovick,S. (2002). Molecular basis of combined pituitary hormone deficiencies. *Endocr.Rev.* 23, 431-442.

Collignon,J., Sockanathan,S., Hacker,A., Cohen-Tannoudji,M., Norris,D., Rastan,S., Stevanovic,M., Goodfellow,P.N., Lovell-Badge,R. (1996). A comparison of the properties of Sox-3 with Sry and two related genes, Sox-1 and Sox-2. *Development* 122, 509-520.

Cook,P.R. (1999). The organization of replication and transcription. *Science* 284, 1790-1795.

Cremers,F.P., Pfeiffer,R.A., van de Pol,T.J., Hofker,M.H., Kruse,T.A., Wieringa,B., Ropers,H.H. (1987). An interstitial duplication of the X chromosome in a male allows physical fine mapping of probes from the Xq13-q22 region. *Hum Genet* 77, 23-7.

Crothers,D.M., Haran,T.E., Nadeau,J.G. (1990). Intrinsically bent DNA. *J Biol.Chem.* 265, 7093-7096.

Darbinian,N., Gallia,G.L., Khalili,K. (2001). Helix-destabilizing properties of the human single-stranded DNA- and RNA-binding protein Puralpha. *J Cell Biochem.* 80, 589-595.

Dasen,J.S., Rosenfeld,M.G. (1999). Signaling mechanisms in pituitary morphogenesis and cell fate determination. *Curr.Opin.Cell Biol.* 11, 669-677.

Dattani,M.T., Martinez-Barbera,J.P., Thomas,P.Q., Brickman,J.M., Gupta,R., Martensson,I.L., Toresson,H., Fox,M., Wales,J.K., Hindmarsh,P.C., Krauss,S., Beddington,R.S., Robinson,I.C. (1998). Mutations in the homeobox gene HESX1/Hesx1 associated with septo-optic dysplasia in human and mouse. *Nat.Genet* 19, 125-133.

Day,G.R., Blake,R.D. (1982). Statistical significance of symmetrical and repetitive segments in DNA. *Nucleic Acids Res.* 10, 8323-8339.

Dayn,A., Samadashwily,G.M., Mirkin,S.M. (1992). Intramolecular DNA triplexes: unusual sequence requirements and influence on DNA polymerization. *Proc.Natl.Acad.Sci.U.S.A* 89, 11406-11410.

de Mollerat,X.J., Gurrieri,F., Morgan,C.T., Sangiorgi,E., Everman,D.B., Gaspari,P., Amiel,J., Bamshad,M.J., Lyle,R., Blouin,J.L., Allanson,J.E., Le Marec,B., Wilson,M., Braverman,N.E., Radhakrishna,U., Delozier-Blanchet,C., Abbott,A., Elghouzzi,V., Antonarakis,S., Stevenson,R.E., Munnich,A., Neri,G., Schwartz,C.E. (2003). A genomic rearrangement resulting in a tandem duplication is associated with split hand-split foot malformation 3 (SHFM3) at 10q24. *Hum.Mol.Genet.* 12, 1959-1971.

Dechering,K.J., Cuelenaere,K., Konings,R.N., Leunissen,J.A. (1998). Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. *Nucleic Acids Res.* 26, 4056-4062.

den Dunnen,J.T., Grootsholten,P.M., Bakker,E., Blonden,L.A., Ginjaar,H.B., Wapenaar,M.C., van Paassen,H.M., van Broeckhoven,C., Pearson,P.L., van Ommen,G.J. (1989). Topography of the Duchenne muscular dystrophy (DMD) gene: FIGE and cDNA analysis of 194 cases reveals 115 deletions and 13 duplications. *Am J Hum Genet* 45, 835-847.

Diehl,H.J., Schaich,M., Budzinski,R.M., Stoffel,W. (1986). Individual exons encode the integral membrane domains of human myelin proteolipid protein. *Proc Natl Acad Sci U S A* 83, 9807-11.

Dobbs,D.L., Shaiu,W.L., Benbow,R.M. (1994). Modular sequence elements associated with origin regions in eukaryotic chromosomal DNA. *Nucleic Acids Res.* 22, 2479-2489.

Duesberg,P., Fabarius,A., Hehlmann,R. (2004). Aneuploidy, the primary cause of the multilateral genomic instability of neoplastic and preneoplastic cells. *IUBMB.Life* 56, 65-81.

Dunham,A., *et al.* (2004). The DNA sequence and analysis of human chromosome 13. *Nature* 428, 522-528.

Early,P., Huang,H., Davis,M., Calame,K., Hood,L. (1980). An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: VH, D and JH. *Cell* 19, 981-992.

Edelmann,L., Pandita,R.K., Morrow,B.E. (1999a). Low-copy repeats mediate the common 3-Mb deletion in patients with velo-cardio-facial syndrome. *Am.J.Hum.Genet.* 64, 1076-1086.

Edelmann,L., Pandita,R.K., Spiteri,E., Funke,B., Goldberg,R., Palanisamy,N., Chaganti,R.S., Magenis,E., Shprintzen,R.J., Morrow,B.E. (1999b). A common molecular basis for rearrangement disorders on chromosome 22q11. *Hum.Mol.Genet.* 8, 1157-1167.

Edelmann,W., Kroger,B., Goller,M., Horak,I. (1989). A recombination hotspot in the LTR of a mouse retrotransposon identified in an in vitro system. *Cell* 57, 937-946.

Eisenbarth,I., Vogel,G., Krone,W., Vogel,W., Assum,G. (2000). An isochore transition in the NF1 gene region coincides with a switch in the extent of linkage disequilibrium. *Am J Hum Genet* 67, 873-880.

Ejima,Y., Yang,L. (2003). Trans mobilization of genomic DNA as a mechanism for retrotransposon-mediated exon shuffling. *Hum Mol Genet* 12, 1321-8.

Ellis,D., Malcolm,S. (1994). Proteolipid protein gene dosage effect in Pelizaeus-Merzbacher disease [letter]. *Nat Genet* 6, 333-4.

Ensenauer,R.E., Adeyinka,A., Flynn,H.C., Michels,V.V., Lindor,N.M., Dawson,D.B., Thorland,E.C., Lorentz,C.P., Goldstein,J.L., McDonald,M.T., Smith,W.E., Simon-Fayard,E., Alexander,A.A., Kulharya,A.S., Ketterling,R.P., Clark,R.D., Jalal,S.M. (2003). Microduplication 22q11.2, an emerging syndrome: clinical, cytogenetic, and molecular analysis of thirteen patients. *Am.J.Hum.Genet.* 73, 1027-1040.

Ewart,A.K., Morris,C.A., Atkinson,D., Jin,W., Sternes,K., Spallone,P., Stock,A.D., Leppert,M., Keating,M.T. (1993). Hemizyosity at the elastin locus in a developmental disorder, Williams syndrome. *Nat.Genet.* 5, 11-16.

Fardaei,M., Rogers,M.T., Thorpe,H.M., Larkin,K., Hamshire,M.G., Harper,P.S., Brook,J.D. (2002). Three proteins, MBNL, MBLL and MBXL, co-localize in vivo with nuclear foci of expanded-repeat transcripts in DM1 and DM2 cells. *Hum Mol.Genet* 11, 805-814.

Fiegler,H., Carr,P., Douglas,E.J., Burford,D.C., Hunt,S., Scott,C.E., Smith,J., Vetrie,D., Gorman,P., Tomlinson,I.P., Carter,N.P. (2003). DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. *Genes Chromosomes.Cancer* 36, 361-374.

Froelich-Ammon,S.J., Gale,K.C., Osherooff,N. (1994). Site-specific cleavage of a DNA hairpin by topoisomerase II. DNA secondary structure as a determinant of enzyme recognition/cleavage. *J Biol Chem* 269, 7719-25.

Fuchs,T., Malecova,B., Linhart,C., Sharan,R., Khen,M., Herwig,R., Shmulevich,D., Elkon,R., Steinfath,M., O'Brien,J.K., Radelof,U., Lehrach,H., Lancet,D., Shamir,R. (2002). DEFOG: a practical scheme for deciphering families of genes. *Genomics* 80, 295-302.

Gaedigk,A., Blum,M., Gaedigk,R., Eichelbaum,M., Meyer,U.A. (1991). Deletion of the entire cytochrome P450 CYP2D6 gene as a cause of impaired drug metabolism in poor metabolizers of the debrisoquine/sparteine polymorphism. *Am.J.Hum.Genet.* 48, 943-950.

Gale,J.M., Tobey,R.A., D'Anna,J.A. (1992). Localization and DNA sequence of a replication origin in the rhodopsin gene locus of Chinese hamster cells. *J Mol.Biol.* 224, 343-358.

Gallia,G.L., Johnson,E.M., Khalili,K. (2000). Puralpha: a multifunctional single-stranded DNA- and RNA-binding protein. *Nucleic Acids Res.* 28, 3197-3205.

Garbern,J., Cambi,F., Shy,M., Kamholz,J. (1999). The molecular pathogenesis of Pelizaeus-Merzbacher disease. *Arch Neurol* 56, 1210-4.

Garbern,J.Y., Cambi,F., Tang,X.M., Sima,A.A., Vallat,J.M., Bosch,E.P., Lewis,R., Shy,M., Sohi,J., Kraft,G., Chen,K.L., Joshi,I., Leonard,D.G., Johnson,W., Raskind,W., Dlouhy,S.R., Pratt,V., Hodes,M.E., Bird,T., Kamholz,J. (1997). Proteolipid protein is necessary in peripheral as well as central myelin. *Neuron* 19, 205-18.

Garbern,J.Y., Yool,D.A., Moore,G.J., Wilds,I.B., Faulk,M.W., Klugmann,M., Nave,K.A., Siermans,E.A., van der Knaap,M.S., Bird,T.D., Shy,M.E., Kamholz,J.A., Griffiths,I.R. (2002). Patients lacking the major CNS myelin protein, proteolipid protein 1, develop length-dependent axonal degeneration in the absence of demyelination and inflammation. *Brain* 125, 551-561.

Gencic,S., Abuelo,D., Ambler,M., Hudson,L.D. (1989). Pelizaeus-Merzbacher disease: an X-linked neurologic disorder of myelin metabolism with a novel mutation in the gene encoding proteolipid protein. *Am J Hum Genet* 45, 435-42.

Gencic,S., Hudson,L.D. (1990). Conservative amino acid substitution in the myelin proteolipid protein of jimpymd mice. *J Neurosci* 10, 117-24.

Giglio,S., Broman,K.W., Matsumoto,N., Calvari,V., Gimelli,G., Neumann,T., Ohashi,H., Voullaire,L., Larizza,D., Giorda,R., Weber,J.L., Ledbetter,D.H., Zuffardi,O. (2001). Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *Am J Hum Genet* 68, 874-83.

Giglio,S., Calvari,V., Gregato,G., Gimelli,G., Camanini,S., Giorda,R., Ragusa,A., Gueneri,S., Selicorni,A., Stumm,M., Tonnies,H., Ventura,M., Zollino,M., Neri,G., Barber,J., Wiczorek,D., Rocchi,M., Zuffardi,O. (2002). Heterozygous submicroscopic inversions involving olfactory receptor-gene clusters mediate the recurrent t(4;8)(p16;p23) translocation. *Am J Hum Genet* 71, 276-85.

Gilbert,D.E., Feigon,J. (1999). Multistranded DNA structures. *Curr.Opin.Struct.Biol.* 9, 305-314.

Ginzinger,D.G. (2002). Gene quantification using real-time quantitative PCR: an emerging technology hits the mainstream. *Exp Hematol* 30, 503-12.

Girard-Reydet,C., Gregoire,D., Vassetzky,Y., Mechali,M. (2004). DNA replication initiates at domains overlapping with nuclear matrix attachment regions in the xenopus and mouse c-myc promoter. *Gene* 332, 129-138.

Goedecke,W., Eijpe,M., Offenbergh,H.H., van Aalderen,M., Heyting,C. (1999). Mre11 and Ku70 interact in somatic cells, but are differentially expressed in early meiosis. *Nat.Genet* 23, 194-198.

Gonzalez,F.J., Skoda,R.C., Kimura,S., Umeno,M., Zanger,U.M., Nebert,D.W., Gelboin,H.V., Hardwick,J.P., Meyer,U.A. (1988). Characterization of the common genetic defect in humans deficient in debrisoquine metabolism. *Nature* 331, 442-446.

Gow,A. (1997). Redefining the lipophilin family of proteolipid proteins. *J Neurosci Res* 50, 659-64.

Gow,A., Lazzarini,R.A. (1996). A cellular mechanism governing the severity of Pelizaeus-Merzbacher disease. *Nat Genet* 13, 422-8.

Gow,A., Southwood,C.M., Lazzarini,R.A. (1998). Disrupted proteolipid protein trafficking results in oligodendrocyte apoptosis in an animal model of Pelizaeus-Merzbacher disease. *J Cell Biol* 140, 925-34.

Graves,J.A. (1998). Evolution of the mammalian Y chromosome and sex-determining genes. *J.Exp.Zool.* 281, 472-481.

Graw,S.L., Sample,T., Bleskan,J., Sujansky,E., Patterson,D. (2000). Cloning, sequencing, and analysis of inv8 chromosome breakpoints associated with recombinant 8 syndrome. *Am J Hum Genet* 66, 1138-44.

Greer,J.M., Lees,M.B. (2002). Myelin proteolipid protein--the first 50 years. *Int J Biochem Cell Biol* 34, 211-5.

Griffiths,I., Klugmann,M., Anderson,T., Thomson,C., Vouyiouklis,D., Nave,K.A. (1998a). Current concepts of PLP and its role in the nervous system. *Microsc Res Tech* 41, 344-58.

Griffiths,I., Klugmann,M., Anderson,T., Yool,D., Thomson,C., Schwab,M.H., Schneider,A., Zimmermann,F., McCulloch,M., Nadon,N., Nave,K.A. (1998b). Axonal swellings and degeneration in mice lacking the major proteolipid of myelin. *Science* 280, 1610-3.

Gudz,T.I., Schneider,T.E., Haas,T.A., Macklin,W.B. (2002). Myelin proteolipid protein forms a complex with integrins and may participate in integrin receptor signaling in oligodendrocytes. *J Neurosci* 22, 7398-407.

Hassed,S., Hopcus-Niccum,D., Zhang,L., Li,S., Mulvihill,J. (2004). A new genomic duplication syndrome complementary to the velocardiofacial (22q11 deletion) syndrome. *Clin.Genet* 65, 400-404.

Hattori,M. *et al.* (2000). The DNA sequence of human chromosome 21. *Nature* 405, 311-319.

Heath,K.E., Day,I.N., Humphries,S.E. (2000). Universal primer quantitative fluorescent multiplex (UPQFM) PCR: a method to detect major and minor rearrangements of the low density lipoprotein receptor gene. *J Med Genet* 37, 272-80.

Hermesz,E., Mackem,S., Mahon,K.A. (1996). Rpx: a novel anterior-restricted homeobox gene progressively activated in the prechordal plate, anterior neural plate and Rathke's pouch of the mouse embryo. *Development* 122, 41-52.

Hilton,D.J., Richardson,R.T., Alexander,W.S., Viney,E.M., Willson,T.A., Sprigg,N.S., Starr,R., Nicholson,S.E., Metcalf,D., Nicola,N.A. (1998). Twenty proteins containing a C-terminal SOCS box form five structural classes. *Proc.Natl.Acad.Sci.U.S.A* 95, 114-119.

Hobson, G., Cundall, M., Sperle, K., Kamholz, J., Garbern, J., Heng, H, Sistermans, E., Malcolm, S., and Woodward, K. Fine mapping of duplication endpoints in Pelizaeus-Merzbacher disease. (2003) *Am J Hum Genet* 73 [5], 342.

Hodes,M.E., Blank,C.A., Pratt,V.M., Morales,J., Napier,J., Dlouhy,S.R. (1997). Nonsense mutation in exon 3 of the proteolipid protein gene (PLP) in a family with an unusual form of Pelizaeus-Merzbacher disease. *Am J Med Genet* 69, 121-5.

Hodes,M.E., DeMyer,W.E., Pratt,V.M., Edwards,M.K., Dlouhy,S.R. (1995). Girl with signs of Pelizaeus-Merzbacher disease heterozygous for a mutation in exon 2 of the proteolipid protein gene. *Am J Med Genet* 55, 397-401.

Hodes,M.E., Pratt,V.M., Dlouhy,S.R. (1993). Genetics of Pelizaeus-Merzbacher disease. *Dev Neurosci* 15, 383-94.

Hodes,M.E., Woodward,K., Spinner,N.B., Emanuel,B.S., Enrico-Simon,A., Kamholz,J., Stambolian,D., Zackai,E.H., Pratt,V.M., Thomas,I.T., Crandall,K., Dlouhy,S.R., Malcolm,S. (2000). Additional copies of the proteolipid protein gene causing Pelizaeus-Merzbacher disease arise by separate integration into the X chromosome. *Am J Hum Genet* 67, 14-22.

Hogan,A., Faust,E.A. (1986). Nonhomologous recombination in the parvovirus chromosome: role for a CTATTCT motif. *Mol.Cell Biol.* 6, 3005-3009.

Hogervorst,F.B., Nederlof,P.M., Gille,J.J., McElgunn,C.J., Grippeling,M., Pruntel,R., Regnerus,R., van Welsem,T., van Spaendonk,R., Menko,F.H., Kluijdt,I., Dommering,C., Verhoef,S., Schouten,J.P., van't Veer,L.J., Pals,G. (2003). Large genomic deletions and duplications in the BRCA1 gene identified by a novel quantitative method. *Cancer Res* 63, 1449-53.

Hol,F.A., Schepens,M.T., van Beersum,S.E., Redolfi,E., Affer,M., Vezzoni,P., Hamel,B.C., Karnes,P.S., Mariman,E.C., Zucchi,I. (2000). Identification and characterization of an Xq26-q27 duplication in a family with spina bifida and panhypopituitarism suggests the involvement of two distinct genes. *Genomics* 69, 174-81.

Hollox,E.J., Atia,T., Cross,G., Parkin,T., Armour,J.A. (2002). High throughput screening of human subtelomeric DNA for copy number changes using multiplex amplifiable probe hybridisation (MAPH). *J Med Genet* 39, 790-5.

Hu,X.Y., Burghes,A.H., Ray,P.N., Thompson,M.W., Murphy,E.G., Worton,R.G. (1988). Partial gene duplication in Duchenne and Becker muscular dystrophies. *J.Med.Genet.* 25, 369-376.

Hudson,L.D., Puckett,C., Berndt,J., Chan,J., Gencic,S. (1989). Mutation of the proteolipid protein gene PLP in a human X chromosome- linked myelin disorder. *Proc Natl Acad Sci U S A* 86, 8128-31.

Inoue,K., Osaka,H., Imaizumi,K., Nezu,A., Takanashi,J., Arii,J., Murayama,K., Ono,J., Kikawa,Y., Mito,T., Shaffer,L.G., Lupski,J.R. (1999). Proteolipid protein gene duplications causing Pelizaeus-Merzbacher disease: molecular mechanism and phenotypic manifestations. *Ann Neurol* 45, 624-32.

Inoue,K., Osaka,H., Sugiyama,N., Kawanishi,C., Onishi,H., Nezu,A., Kimura,K., Yamada,Y., Kosaka,K. (1996a). A duplicated PLP gene causing Pelizaeus-Merzbacher disease detected by comparative multiplex PCR. *Am J Hum Genet* 59, 32-9.

Inoue,K., Osaka,H., Thurston,V.C., Clarke,J.T., Yoneyama,A., Rosenbarker,L., Bird,T.D., Hodes,M.E., Shaffer,L.G., Lupski,J.R. (2002). Genomic rearrangements resulting in PLP1 deletion occur by nonhomologous end joining and cause different dysmyelinating phenotypes in males and females. *Am J Hum Genet* 71, 838-53.

Inoue,K., Tanaka,H., Scaglia,F., Araki,A., Shaffer,L.G., Lupski,J.R. (2001). Compensating for central nervous system dysmyelination: females with a proteolipid protein gene duplication and sustained clinical improvement. *Ann Neurol* 50, 747-54.

Inoue,Y., Kagawa,T., Matsumura,Y., Ikenaka,K., Mikoshiba,K. (1996b). Cell death of oligodendrocytes or demyelination induced by overexpression of proteolipid protein depending on expressed gene dosage. *Neurosci Res* 25, 161-72.

Ishkanian,A.S., Malloff,C.A., Watson,S.K., DeLeeuw,R.J., Chi,B., Coe,B.P., Snijders,A., Albertson,D.G., Pinkel,D., Marra,M.A., Ling,V., MacAulay,C., Lam,W.L. (2004). A tiling resolution DNA microarray with complete coverage of the human genome. *Nat.Genet.* 36, 299-303.

Iwaki, A., Kondo, J., Ototsuji, M., Kurosawa, K., and Fukumaki, Y. Characterisation of the breakpoints of PLP1 duplication in three cases of Pelizaeus-Merzbacher disease. (2003) *Am J Hum Genet* 73 [5], 549.

Jackson,D.A., Pombo,A. (1998). Replicon clusters are stable units of chromosome structure: evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells. *J Cell Biol.* 140, 1285-1295.

Jeffreys,A.J., Wilson,V., Thein,S.L. (1985). Hypervariable 'minisatellite' regions in human DNA. *Nature* 314, 67-73.

Ji,Y., Walkowicz,M.J., Buiting,K., Johnson,D.K., Tarvin,R.E., Rinchik,E.M., Horsthemke,B., Stubbs,L., Nicholls,R.D. (1999). The ancestral gene for transcribed, low-copy repeats in the Prader-Willi/Angelman region encodes a large protein implicated in protein trafficking, which is deficient in mice with neuromuscular and spermiogenic abnormalities. *Hum.Mol.Genet.* 8, 533-542.

Johansson,I., Lundqvist,E., Bertilsson,L., Dahl,M.L., Sjoqvist,F., Ingelman-Sundberg,M. (1993). Inherited amplification of an active gene in the cytochrome P450 CYP2D locus as a cause of ultrarapid metabolism of debrisoquine. *Proc.Natl.Acad.Sci.U.S.A* 90, 11825-11829.

Johnson,R.D., Jasin,M. (2000). Sister chromatid gene conversion is a prominent double-strand break repair pathway in mammalian cells. *EMBO J* 19, 3398-3407.

Jun,L., Frints,S., Duhamel,H., Herold,A., Abad-Rodrigues,J., Dotti,C., Izaurralde,E., Marynen,P., Froyen,G. (2001). NXF5, a novel member of the nuclear RNA export factor family, is lost in a male patient with a syndromic form of mental retardation. *Curr Biol* 11, 1381-91.

Jurka,J. (2000). Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* 16, 418-420.

Jurka,J., Kohany,O., Pavlicek,A., Kapitonov,V.V., Jurka,M.V. (2004). Duplication, coclustering, and selection of human Alu retrotransposons. *Proc.Natl.Acad.Sci.U.S.A* 101, 1268-1272.

Kagawa,T., Ikenaka,K., Inoue,Y., Kuriyama,S., Tsujii,T., Nakao,J., Nakajima,K., Aruga,J., Okano,H., Mikoshiba,K. (1994). Glial cell degeneration and hypomyelination caused by overexpression of myelin proteolipid protein gene. *Neuron* 13, 427-42.

Kallioniemi,A., Kallioniemi,O.P., Sudar,D., Rutovitz,D., Gray,J.W., Waldman,F., Pinkel,D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* 258, 818-821.

Kamachi,Y., Uchikawa,M., Collignon,J., Lovell-Badge,R., Kondoh,H. (1998). Involvement of Sox1, 2 and 3 in the early and subsequent molecular events of lens induction. *Development* 125, 2521-2532.

Karanjawala,Z.E., Grawunder,U., Hsieh,C.L., Lieber,M.R. (1999). The nonhomologous DNA end joining pathway is important for chromosome stability in primary fibroblasts. *Curr.Biol.* 9, 1501-1504.

Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J., Weber,R.J., Haussler,D., Kent,W.J. (2003). The UCSC Genome Browser Database. *Nucleic Acids Res.* 31, 51-54.

Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D., Kent,W.J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 32 *Database issue*, D493-D496.

Katoh,K., Miyata,T. (1999). A heuristic approach of maximum likelihood method for inferring phylogenetic tree and an application to the mammalian SOX-3 origin of the testis-determining gene SRY. *FEBS Lett.* 463, 129-132.

Katznelson,L., Klibanski,A. (2001). Prolactin and Its Disorders. In: *Principles and Practice of Endocrinology and Metabolism*, ed. K.L.Becker Lippincott, Williams & Wilkins, 145-153.

Kauppi,L., Jeffreys,A.J., Keeney,S. (2004). Where the crossovers are: recombination distributions in mammals. *Nat.Rev.Genet* 5, 413-424.

Kile,B.T., Schulman,B.A., Alexander,W.S., Nicola,N.A., Martin,H.M., Hilton,D.J. (2002). The SOCS box: a tale of destruction and degradation. *Trends Biochem.Sci.* 27, 235-241.

Kitagawa,K., Sinoway,M.P., Yang,C., Gould,R.M., Colman,D.R. (1993). A proteolipid protein gene family: expression in sharks and rays and possible evolution from an ancestral gene encoding a pore-forming polypeptide. *Neuron* 11, 433-48.

Kleinjan,D.J., van Heyningen,V. (1998). Position effect in human genetic disease. *Hum Mol.Genet* 7, 1611-1618.

Klugmann,M., Schwab,M.H., Puhlhofer,A., Schneider,A., Zimmermann,F., Griffiths,I.R., Nave,K.A. (1997). Assembly of CNS myelin in the absence of proteolipid protein. *Neuron* 18, 59-70.

Knuutila,S., Autio,K., Aalto,Y. (2000). Online access to CGH data of DNA sequence copy number changes. *Am.J.Pathol.* 157, 689.

Kobori,J.A., Strauss,E., Minard,K., Hood,L. (1986). Molecular analysis of the hotspot of recombination in the murine major histocompatibility complex. *Science* 234, 173-179.

Koenig,M., Beggs,A.H., Moyer,M., Scherpf,S., Heindrich,K., Bettecken,T., Meng,G., Muller,C.R., Lindlof,M., Kaariainen,H., . (1989). The molecular basis for Duchenne versus Becker muscular dystrophy: correlation of severity with type of deletion. *Am J Hum Genet* 45, 498-506.

Koenig,M., Hoffman,E.P., Bertelson,C.J., Monaco,A.P., Feener,C., Kunkel,L.M. (1987). Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. *Cell* 50, 509-517.

- Kohwi,Y. (1989). Cationic metal-specific structures adopted by the poly(dG) region and the direct repeats in the chicken adult beta A globin gene promoter. *Nucleic Acids Res.* *17*, 4493-4502.
- Kohwi,Y., Kohwi-Shigematsu,T. (1988). Magnesium ion-dependent triple-helix structure formed by homopurine-homopyrimidine sequences in supercoiled plasmid DNA. *Proc.Natl.Acad.Sci.U.S.A* *85*, 3781-3785.
- Kohwi,Y., Kohwi-Shigematsu,T. (1993). Structural polymorphism of homopurine-homopyrimidine sequences at neutral pH. *J Mol.Biol.* *231*, 1090-1101.
- Kohwi,Y., Panchenko,Y. (1993). Transcription-dependent recombination induced by triple-helix formation. *Genes Dev.* *7*, 1766-1778.
- Kong,A., Gudbjartsson,D.F., Sainz,J., Jonsdottir,G.M., Gudjonsson,S.A., Richardsson,B., Sigurdardottir,S., Barnard,J., Hallbeck,B., Masson,G., Shlien,A., Palsson,S.T., Frigge,M.L., Thorgeirsson,T.E., Gulcher,J.R., Stefansson,K. (2002). A high-resolution recombination map of the human genome. *Nat.Genet* *31*, 241-247.
- Konopka,A.K. (1988). Compilation of DNA strand exchange sites for non-homologous recombination in somatic cells. *Nucleic Acids Res.* *16*, 1739-1758.
- Kraus,E., Leung,W.Y., Haber,J.E. (2001). Break-induced replication: a review and an example in budding yeast. *Proc.Natl.Acad.Sci.U.S.A* *98*, 8255-8262.
- Krawczak,M., Cooper,D.N. (1991). Gene deletions causing human genetic disease: mechanisms of mutagenesis and the role of the local DNA sequence environment. *Hum Genet* *86*, 425-441.
- Krowczynska,A.M., Rudders,R.A., Krontiris,T.G. (1990). The human minisatellite consensus at breakpoints of oncogene translocations. *Nucleic Acids Res.* *18*, 1121-1127.
- Kumar,S., Subramanian,S. (2002). Mutation rates in mammalian genomes. *Proc.Natl.Acad.Sci.U.S.A* *99*, 803-808.
- Kunkel,T.A. (1985a). The mutational specificity of DNA polymerase-beta during in vitro DNA synthesis. Production of frameshift, base substitution, and deletion mutations. *J Biol.Chem.* *260*, 5787-5796.
- Kunkel,T.A. (1985b). The mutational specificity of DNA polymerases-alpha and -gamma during in vitro DNA synthesis. *J Biol.Chem.* *260*, 12866-12874.
- Kuzminov,A. (2001). Single-strand interruptions in replicating chromosomes cause double-strand breaks. *Proc.Natl.Acad.Sci.U.S.A* *98*, 8241-8246.
- Lagarkova,M.A., Svetlova,E., Giacca,M., Falaschi,A., Razin,S.V. (1998). DNA loop anchorage region colocalizes with the replication origin located downstream to the human gene encoding lamin B2. *J Cell Biochem.* *69*, 13-18.
- Lagerstrom-Fermer,M., Sundvall,M., Johnsen,E., Warne,G.L., Forrest,S.M., Zajac,J.D., Rickards,A., Ravine,D., Landegren,U., Pettersson,U. (1997). X-linked recessive panhypopituitarism associated with a regional duplication in Xq25-q26. *Am.J.Hum.Genet.* *60*, 910-916.

Lakich,D., Kazazian,H.H.Jr., Antonarakis,S.E., Gitschier,J. (1993). Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nat Genet* 5, 236-241.

Lander,E.S., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860-921.

Lapidot,A., Baran,N., Manor,H. (1989). (dT-dC)_n and (dG-dA)_n tracts arrest single stranded DNA replication in vitro. *Nucleic Acids Res.* 17, 883-900.

Laumonnier,F., Ronce,N., Hamel,B.C., Thomas,P., Lespinasse,J., Raynaud,M., Paringaux,C., Van Bokhoven,H., Kalscheuer,V., Fryns,J.P., Chelly,J., Moraine,C., Briault,S. (2002). Transcription Factor SOX3 Is Involved in X-Linked Mental Retardation with Growth Hormone Deficiency. *Am J Hum Genet* 71, 1450-5.

Lehmann,O.J., Ebenezer,N.D., Ekong,R., Ocaka,L., Mungall,A.J., Fraser,S., McGill,J.I., Hitchings,R.A., Khaw,P.T., Sowden,J.C., Povey,S., Walter,M.A., Bhattacharya,S.S., Jordan,T. (2002). Ocular developmental abnormalities and glaucoma associated with interstitial 6p25 duplications and deletions. *Invest Ophthalmol.Vis.Sci.* 43, 1843-1849.

Lehmann,O.J., Ebenezer,N.D., Jordan,T., Fox,M., Ocaka,L., Payne,A., Leroy,B.P., Clark,B.J., Hitchings,R.A., Povey,S., Khaw,P.T., Bhattacharya,S.S. (2000). Chromosomal duplication involving the forkhead transcription factor gene FOXC1 causes iris hypoplasia and glaucoma. *Am.J.Hum.Genet.* 67, 1129-1135.

Lemmers,R.J., Van Overveld,P.G., Sandkuijl,L.A., Vrieling,H., Padberg,G.W., Frants,R.R., van der Maarel,S.M. (2004). Mechanism and timing of mitotic rearrangements in the subtelomeric D4Z4 repeat involved in facioscapulohumeral muscular dystrophy. *Am J Hum Genet* 75, 44-53.

Lengauer,C., Kinzler,K.W., Vogelstein,B. (1998). Genetic instabilities in human cancers. *Nature* 396, 643-649.

Liao,D. (1999). Concerted evolution: molecular mechanism and biological implications. *Am J Hum Genet* 64, 24-30.

Lieber,M.R., Ma,Y., Pannicke,U., Schwarz,K. (2003). Mechanism and regulation of human non-homologous DNA end-joining. *Nat.Rev.Mol.Cell Biol.* 4, 712-720.

Lin,Y., Waldman,A.S. (2001a). Capture of DNA sequences at double-strand breaks in mammalian chromosomes. *Genetics* 158, 1665-1674.

Lin,Y., Waldman,A.S. (2001b). Promiscuous patching of broken chromosomes in mammalian cells with extrachromosomal DNA. *Nucleic Acids Res.* 29, 3975-3981.

Little,K.C., Chartrand,P. (2004). Genomic DNA is captured and amplified during double-strand break (DSB) repair in human cells. *Oncogene* 23, 4166-4172.

Lopes,J., Ravise,N., Vandenberghe,A., Palau,F., Ionasescu,V., Mayer,M., Levy,N., Wood,N., Tachi,N., Bouche,P., Latour,P., Ruberg,M., Brice,A., LeGuern,E. (1998). Fine mapping of de novo CMT1A and HNPP rearrangements within CMT1A-REPs evidences two distinct sex-dependent mechanisms and candidate sequences involved in recombination. *Hum Mol.Genet* 7, 141-148.

Lopes,J., Vandenberghe,A., Tardieu,S., Ionasescu,V., Levy,N., Wood,N., Tachi,N., Bouche,P., Latour,P., Brice,A., LeGuern,E. (1997). Sex-dependent rearrangements resulting in CMT1A and HNPP. *Nat.Genet* 17, 136-137.

Lupski,J.R. (1998). Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet* 14, 417-22.

Lupski,J.R., de Oca-Luna,R.M., Slaugenhaupt,S., Pentao,L., Guzzetta,V., Trask,B.J., Saucedo-Cardenas,O., Barker,D.F., Killian,J.M., Garcia,C.A., al,e. (1991). DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell* 66, 219-32.

Lynch,M., Conery,J.S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151-1155.

Machinis,K., Pantel,J., Netchine,I., Leger,J., Camand,O.J., Sobrier,M.L., Dastot-Le Moal,F., Duquesnoy,P., Abitbol,M., Czernichow,P., Amselem,S. (2001). Syndromic short stature in patients with a germline mutation in the LIM homeobox LHX4. *Am J Hum Genet* 69, 961-968.

MacNeil,D.J., Howard,A.D., Guan,X., Fong,T.M., Nargund,R.P., Bednarek,M.A., Goulet,M.T., Weinberg,D.H., Strack,A.M., Marsh,D.J., Chen,H.Y., Shen,C.P., Chen,A.S., Rosenblum,C.I., MacNeil,T., Tota,M., MacIntyre,E.D., Van Der Ploeg,L.H. (2002). The role of melanocortins in body weight regulation: opportunities for the treatment of obesity. *Eur.J Pharmacol.* 440, 141-157.

Magenis,R.E., Toth-Fejel,S., Allen,L.J., Black,M., Brown,M.G., Budden,S., Cohen,R., Friedman,J.M., Kalousek,D., Zonana,J., . (1990). Comparison of the 15q deletions in Prader-Willi and Angelman syndromes: specific regions, extent of deletions, parental origin, and clinical consequences. *Am.J.Med.Genet.* 35, 333-349.

Mahadevaiah,S.K., Turner,J.M., Baudat,F., Rogakou,E.P., de Boer,P., Blanco-Rodriguez,J., Jasin,M., Keeney,S., Bonner,W.M., Burgoyne,P.S. (2001). Recombinational DNA double-strand breaks in mice precede synapsis. *Nat Genet* 27, 271-6.

Majewski,J., Ott,J. (2000). GT repeats are associated with recombination on human chromosome 22. *Genome Res.* 10, 1108-1114.

Malnic,B., Godfrey,P.A., Buck,L.B. (2004). The human olfactory receptor gene family. *Proc.Natl.Acad.Sci.U.S.A* 101, 2584-2589.

Marx,J. (2002). Debate surges over the origins of genomic defects in cancer. *Science* 297, 544-546.

Marzluff,W.F., Gongidi,P., Woods,K.R., Jin,J., Maltais,L.J. (2002). The human and mouse replication-dependent histone genes. *Genomics* 80, 487-498.

Maundrell,K., Hutchison,A., Shall,S. (1988). Sequence analysis of ARS elements in fission yeast. *EMBO J* 7, 2203-2209.

Max,E.E., Seidman,J.G., Leder,P. (1979). Sequences of five potential recombination sites encoded close to an immunoglobulin kappa constant region gene. *Proc.Natl.Acad.Sci.U.S.A* 76, 3450-3454.

- McLaughlin,M., Hunter,D.J., Thomson,C.E., Yool,D., Kirkham,D., Freer,A.A., Griffiths,I.R. (2002). Evidence for possible interactions between PLP and DM20 within the myelin sheath. *Glia* 39, 31-6.
- McNaughton,J.C., Cockburn,D.J., Hughes,G., Jones,W.A., Laing,N.G., Ray,P.N., Stockwell,P.A., Petersen,G.B. (1998). Is gene deletion in eukaryotes sequence-dependent? A study of nine deletion junctions and nineteen other deletion breakpoints in intron 7 of the human dystrophin gene. *Gene* 222, 41-51.
- McTaggart,K.E., Budarf,M.L., Driscoll,D.A., Emanuel,B.S., Ferreira,P., McDermid,H.E. (1998). Cat eye syndrome chromosome breakpoint clustering: identification of two intervals also associated with 22q11 deletion syndrome breakpoints. *Cytogenet.Cell Genet.* 81, 222-228.
- Merrihew,R.V., Marburger,K., Pennington,S.L., Roth,D.B., Wilson,J.H. (1996). High-frequency illegitimate integration of transfected DNA at preintegrated target sites in a mammalian genome. *Mol.Cell Biol.* 16, 10-18.
- Michel,B. (2000). Replication fork arrest and DNA recombination. *Trends Biochem.Sci.* 25, 173-178.
- Milner,R.J., Lai,C., Nave,K.A., Lenoir,D., Ogata,J., Sutcliffe,J.G. (1985). Nucleotide sequences of two mRNAs for rat brain myelin proteolipid protein. *Cell* 42, 931-939.
- Milunsky,J.M., Huang,X.L. (2003). Unmasking Kabuki syndrome: chromosome 8p22-8p23.1 duplication revealed by comparative genomic hybridization and BAC-FISH. *Clin.Genet.* 64, 509-516.
- Mimault,C., Giraud,G., Courtois,V., Cailloux,F., Boire,J.Y., Dastugue,B., Boespflug-Tanguy,O. (1999). Proteolipoprotein gene analysis in 82 patients with sporadic Pelizaeus-Merzbacher Disease: duplications, the major cause of the disease, originate more frequently in male germ cells, but point mutations do not. The Clinical European Network on Brain Dysmyelinating Disease. *Am J Hum Genet* 65, 360-9.
- Mirkin,S.M., Lyamichev,V.I., Drushlyak,K.N., Dobrynin,V.N., Filippov,S.A., Frank-Kamenetskii,M.D. (1987). DNA H form requires a homopurine-homopyrimidine mirror repeat. *Nature* 330, 495-497.
- Morrish,T.A., Gilbert,N., Myers,J.S., Vincent,B.J., Stamato,T.D., Taccioli,G.E., Batzer,M.A., Moran,J.V. (2002). DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat.Genet* 31, 159-165.
- Morrow,D.M., Connelly,C., Hieter,P. (1997). "Break copy" duplication: a model for chromosome fragment formation in *Saccharomyces cerevisiae*. *Genetics* 147, 371-382.
- Moser,H.E., Dervan,P.B. (1987). Sequence-specific cleavage of double helical DNA by triple helix formation. *Science* 238, 645-650.
- Mukai,J., Hachiya,T., Shoji-Hoshino,S., Kimura,M.T., Nadano,D., Suvanto,P., Hanaoka,T., Li,Y., Irie,S., Greene,L.A., Sato,T.A. (2000). NADE, a p75NTR-associated cell death executor, is involved in signal transduction mediated by the common neurotrophin receptor p75NTR. *J Biol.Chem.* 275, 17566-17570.

- Mukai,J., Suvant,P., Sato,T.A. (2003). Nerve growth factor-dependent regulation of NADE-induced apoptosis. *Vitam.Horm.* 66, 385-402.
- Nachman,M.W., Crowell,S.L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics* 156, 297-304.
- Nadon,N.L., Arnheiter,H., Hudson,L.D. (1994). A combination of PLP and DM20 transgenes promotes partial myelination in the jimpy mouse. *J.Neurochem.* 63, 822-833.
- Nadon,N.L., Duncan,I.D., Hudson,L.D. (1990). A point mutation in the proteolipid protein gene of the 'shaking pup' interrupts oligodendrocyte development. *Development* 110, 529-37.
- Namciu,S.J., Friedman,R.D., Marsden,M.D., Sarausad,L.M., Jasoni,C.L., Fournier,R.E. (2004). Sequence organization and matrix attachment regions of the human serine protease inhibitor gene cluster at 14q32.1. *Mamm.Genome* 15, 162-178.
- Nance,M.A., Boyadjiev,S., Pratt,V.M., Taylor,S., Hodes,M.E., Dlouhy,S.R. (1996). Adult-onset neurodegenerative disorder due to proteolipid protein gene mutation in the mother of a man with Pelizaeus-Merzbacher disease. *Neurology* 47, 1333-5.
- Napierala,M., Dere,R., Vetcher,A., Wells,R.D. (2004). Structure-dependent recombination hot spot activity of GAA.TTC sequences from intron 1 of the Friedreich's ataxia gene. *J Biol.Chem.* 279, 6444-6454.
- Nave,K.A., Lai,C., Bloom,F.E., Milner,R.J. (1986). Jimpy mutant mouse: a 74-base deletion in the mRNA for myelin proteolipid protein and evidence for a primary defect in RNA splicing. *Proc Natl Acad Sci U S A* 83, 9264-8.
- Naylor,J., Brinke,A., Hassock,S., Green,P.M., Giannelli,F. (1993). Characteristic mRNA abnormality found in half the patients with severe haemophilia A is due to large DNA inversions. *Hum.Mol.Genet.* 2, 1773-1778.
- Netchine,I., Sobrier,M.L., Krude,H., Schnabel,D., Maghnie,M., Marcos,E., Duriez,B., Cacheux,V., Moers,A., Goossens,M., Gruters,A., Amselem,S. (2000). Mutations in LHX3 result in a new syndrome revealed by combined pituitary hormone deficiency. *Nat.Genet* 25, 182-186.
- Nishimura,D.Y., Searby,C.C., Alward,W.L., Walton,D., Craig,J.E., Mackey,D.A., Kawase,K., Kanis,A.B., Patil,S.R., Stone,E.M., Sheffield,V.C. (2001). A spectrum of FOXC1 mutations suggests gene dosage as a mechanism for developmental defects of the anterior chamber of the eye. *Am.J.Hum.Genet.* 68, 364-372.
- O'Brien,S.J., Stanyon,R. (1999). Phylogenomics. Ancestral primate viewed. *Nature* 402, 365-366.
- Ochman,H., Gerber,A.S., Hartl,D.L. (1988). Genetic applications of an inverse polymerase chain reaction. *Genetics* 120, 621-3.
- Ohno,M., Fukagawa,T., Lee,J.S., Ikemura,T. (2002). Triplex-forming DNAs in the human interphase nucleus visualized in situ by polypurine/polypyrimidine DNA probes and antitriplex antibodies. *Chromosoma* 111, 201-213.
- Ohno,S. (1970). *Evolution by Gene Duplication*. Berlin: Springer-Verlag.

Ohno,S. (1981). (AGCTG) (AGCTG) (AGCTG) (GGGTG) as the primordial sequence of intergenic spacers: the role in immunoglobulin class switch. *Differentiation* 18, 65-74.

Oliner,J.D., Kinzler,K.W., Meltzer,P.S., George,D.L., Vogelstein,B. (1992). Amplification of a gene encoding a p53-associated protein in human sarcomas. *Nature* 358, 80-83.

Osborne,L.R., Li,M., Pober,B., Chitayat,D., Bodurtha,J., Mandel,A., Costa,T., Grebe,T., Cox,S., Tsui,L.C., Scherer,S.W. (2001). A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. *Nat Genet* 29, 321-5.

Ovcharenko,I., Loots,G.G., Hardison,R.C., Miller,W., Stubbs,L. (2004). zPicture: dynamic alignment and visualization tool for analyzing conservation profiles. *Genome Res.* 14, 472-477.

Patel,H.P., Lu,L., Blaszak,R.T., Bissler,J.J. (2004). PKD1 intron 21: triplex DNA formation and effect on replication. *Nucleic Acids Res.* 32, 1460-1468.

Pavelitz,T., Liao,D., Weiner,A.M. (1999). Concerted evolution of the tandem array encoding primate U2 snRNA (the RNU2 locus) is accompanied by dramatic remodeling of the junctions with flanking chromosomal sequences. *EMBO J* 18, 3783-3792.

Pereira-Leal,J.B., Hume,A.N., Seabra,M.C. (2001). Prenylation of Rab GTPases: molecular mechanisms and involvement in genetic disease. *FEBS Lett.* 498, 197-200.

Pereira-Leal,J.B., Seabra,M.C. (2000). The mammalian Rab family of small GTPases: definition of family and subfamily sequence motifs suggests a mechanism for functional specificity in the Ras superfamily. *J Mol.Biol.* 301, 1077-1087.

Pereira-Leal,J.B., Seabra,M.C. (2001). Evolution of the Rab family of small GTP-binding proteins. *J Mol.Biol.* 313, 889-901.

Petrij-Bosch,A., Peelen,T., van Vliet,M., van Eijk,R., Olmer,R., Drusedau,M., Hogervorst,F.B., Hageman,S., Arts,P.J., Ligtenberg,M.J., Meijers-Heijboer,H., Klijn,J.G., Vasen,H.F., Cornelisse,C.J., 't Veer,L.J., Bakker,E., van Ommen,G.J., Devilee,P. (1997). BRCA1 genomic deletions are major founder mutations in Dutch breast cancer patients. *Nat.Genet.* 17, 341-345.

Pfaffle,R.W., DiMattia,G.E., Parks,J.S., Brown,M.R., Wit,J.M., Jansen,M., Van der,N.H., Van den Brande,J.L., Rosenfeld,M.G., Ingraham,H.A. (1992). Mutation of the POU-specific domain of Pit-1 and hypopituitarism without pituitary hypoplasia. *Science* 257, 1118-1121.

Philipp-Staheli,J., Payne,S.R., Kemp,C.J. (2001). p27(Kip1): regulation and function of a haploinsufficient tumor suppressor and its misregulation in cancer. *Exp.Cell Res.* 264, 148-168.

Pihan,G., Doxsey,S.J. (2003). Mutations and aneuploidy: co-conspirators in cancer? *Cancer Cell* 4, 89-94.

Pillutla,R.C., Shimamoto,A., Furuichi,Y., Shatkin,A.J. (1999). Genomic structure and chromosomal localization of TCEAL1, a human gene encoding the nuclear phosphoprotein p21/SIIR. *Genomics* 56, 217-220.

- Pinkel,D., Segreaves,R., Sudar,D., Clark,S., Poole,I., Kowbel,D., Collins,C., Kuo,W.L., Chen,C., Zhai,Y., Dairkee,S.H., Ljung,B.M., Gray,J.W., Albertson,D.G. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat.Genet.* 20, 207-211.
- Pinzone,J.J. (2001). Hypopituitarism. In: *Principles and Practice of Endocrinology and Metabolism*, ed. K.L.Becker Lippincott, Williams & Wilkins, 177-192.
- Plug,A.W., Xu,J., Reddy,G., Golub,E.I., Ashley,T. (1996). Presynaptic association of Rad51 protein with selected sites in meiotic chromatin. *Proc.Natl.Acad.Sci.U.S.A* 93, 5920-5924.
- Potocki,L., Chen,K.S., Park,S.S., Osterholm,D.E., Withers,M.A., Kimonis,V., Summers,A.M., Meschino,W.S., Anyane-Yeboah,K., Kashork,C.D., Shaffer,L.G., Lupski,J.R. (2000). Molecular mechanism for duplication 17p11.2- the homologous recombination reciprocal of the Smith-Magenis microdeletion. *Nat Genet* 24, 84-87.
- Pribyl,T.M., Campagnoni,C., Kampf,K., Handley,V.W., Campagnoni,A.T. (1996). The major myelin protein genes are expressed in the human thymus. *J.Neurosci.Res.* 45, 812-819.
- Prince,V.E., Pickett,F.B. (2002). Splitting pairs: the diverging fates of duplicated genes. *Nat.Rev.Genet* 3, 827-837.
- Puget,N., Stoppa-Lyonnet,D., Sinilnikova,O.M., Pages,S., Lynch,H.T., Lenoir,G.M., Mazoyer,S. (1999). Screening for germ-line rearrangements and regulatory mutations in BRCA1 led to the identification of four new deletions. *Cancer Res.* 59, 455-461.
- Qian,Z., Lin,C., Espinosa,R., LeBeau,M., Rosner,M.R. (2001). Cloning and characterization of MST4, a novel Ste20-like kinase. *J Biol.Chem.* 276, 22439-22445.
- Rabbitts,T.H., Forster,A., Milstein,C.P. (1981). Human immunoglobulin heavy chain genes: evolutionary comparisons of C mu, C delta and C gamma genes and associated switch sequences. *Nucleic Acids Res.* 9, 4509-4524.
- Radovick,S., Nations,M., Du,Y., Berg,L.A., Weintraub,B.D., Wondisford,F.E. (1992). A mutation in the POU-homeodomain of Pit-1 responsible for combined pituitary hormone deficiency. *Science* 257, 1115-1118.
- Raeymaekers,P., Timmerman,V., Nelis,E., De Jonghe,P., Hoogendijk,J.E., Baas,F., Barker,D.F., Martin,J.J., de Visser,M., Bolhuis,P.A., . (1991). Duplication in chromosome 17p11.2 in Charcot-Marie-Tooth neuropathy type 1a (CMT 1a). The HMSN Collaborative Research Group. *Neuromuscul.Disord.* 1, 93-97.
- Raskind,W.H., Williams,C.A., Hudson,L.D., Bird,T.D. (1991). Complete deletion of the proteolipid protein gene (PLP) in a family with X-linked Pelizaeus-Merzbacher disease. *Am J Hum Genet* 49, 1355-60.
- Rassool,F.V., McKeithan,T.W., Neilly,M.E., van Melle,E., Espinosa,R., III, Le Beau,M.M. (1991). Preferential integration of marker DNA into the chromosomal fragile site at 3p14: an approach to cloning fragile sites. *Proc.Natl.Acad.Sci.U.S.A* 88, 6657-6661.

Razin,S.V., Vassetzky,Y.S., Hancock,R. (1991). Nuclear matrix attachment regions and topoisomerase II binding and reaction sites in the vicinity of a chicken DNA replication origin. *Biochem.Biophys.Res.Commun.* 177, 265-270.

Readhead,C., Schneider,A., Griffiths,I., Nave,K.A. (1994). Premature arrest of myelin formation in transgenic mice with increased proteolipid protein gene dosage. *Neuron* 12, 583-95.

Reich, D. E., Lander, E. S., Waterson, R., Paabo, S, Ruvolo, M., and Varki, A. Sequencing the Chimpanzee Genome. (2002)

<http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/ChimpGenome2.pdf>.

Reiter,L.T., Murakami,T., Koeuth,T., Pentao,L., Muzny,D.M., Gibbs,R.A., Lupski,J.R. (1996). A recombination hotspot responsible for two inherited peripheral neuropathies is located near a mariner transposon-like element. *Nat.Genet* 12, 288-297.

Remenyi,A., Lins,K., Nissen,L.J., Reinbold,R., Scholer,H.R., Wilmanns,M. (2003). Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. *Genes Dev.* 17, 2048-2059.

Richardson,C., Jasin,M. (2000). Coupled homologous and nonhomologous repair of a double-strand break preserves genomic integrity in mammalian cells. *Mol Cell Biol* 20, 9068-75.

Riley,D.E., Reeves,R., Gartler,S.M. (1986). Xrep, a plasmid-stimulating X chromosomal sequence bearing similarities to the BK virus replication origin and viral enhancers. *Nucleic Acids Res.* 14, 9407-9423.

Rizzoti,K., Brunelli,S., Carmignac,D., Thomas,P.Q., Robinson,I.C., Lovell-Badge,R. (2004). SOX3 is required during the formation of the hypothalamo-pituitary axis. *Nat.Genet.* 36, 247-255.

Robertson,G.L. (2001). Physiology of Vasopressin, Oxytocin and Thirst. In: Principles and Practice of Endocrinology and Metabolism, ed. K.L.Becker Lippincott, Williams & Wilkins, 276-285.

Robinson,W.P., Waslynka,J., Bernasconi,F., Wang,M., Clark,S., Kotzot,D., Schinzel,A. (1996). Delineation of 7q11.2 deletions associated with Williams-Beuren syndrome and mapping of a repetitive sequence to within and to either side of the common deletion. *Genomics* 34, 17-23.

Rocher,C., Letellier,T., Copeland,W.C., Lestienne,P. (2002). Base composition at mtDNA boundaries suggests a DNA triple helix model for human mitochondrial DNA large-scale rearrangements. *Mol.Genet Metab* 76, 123-132.

Rockwood,L.D., Felix,K., Janz,S. (2004). Elevated presence of retrotransposons at sites of DNA double strand break repair in mouse models of metabolic oxidative stress and MYC-induced lymphoma. *Mutat.Res.* 548, 117-125.

Rooney,S.M., Moore,P.D. (1995). Antiparallel, intramolecular triplex DNA stimulates homologous recombination in human cells. *Proc.Natl.Acad.Sci.U.S.A* 92, 2141-2144.

Rosenthal,N., Kress,M., Gruss,P., Khoury,G. (1983). BK viral enhancer element and a human cellular homolog. *Science* 222, 749-755.

- Roth,D.B., Chang,X.B., Wilson,J.H. (1989). Comparison of filler DNA at immune, nonimmune, and oncogenic rearrangements suggests multiple mechanisms of formation. *Mol.Cell Biol.* 9, 3049-3057.
- Roth,D.B., Proctor,G.N., Stewart,L.K., Wilson,J.H. (1991). Oligonucleotide capture during end joining in mammalian cells. *Nucleic Acids Res.* 19, 7201-7205.
- Roth,D.B., Wilson,J.H. (1986). Nonhomologous recombination in mammalian cells: role for short sequence homologies in the joining reaction. *Mol.Cell Biol.* 6, 4295-4304.
- Saito-Ohara,F., Fukuda,Y., Ito,M., Agarwala,K.L., Hayashi,M., Matsuo,M., Imoto,I., Yamakawa,K., Nakamura,Y., Inazawa,J. (2002). The Xq22 Inversion Breakpoint Interrupted a Novel Ras-Like GTPase Gene in a Patient with Duchenne Muscular Dystrophy and Profound Mental Retardation. *Am J Hum Genet* 71, 637-45.
- Saland,L.C. (2001). The mammalian pituitary intermediate lobe: an update on innervation and regulation. *Brain Res.Bull.* 54, 587-593.
- Samonte,R.V., Eichler,E.E. (2002). Segmental duplications and the evolution of the primate genome. *Nat.Rev.Genet* 3, 65-72.
- Sander,M., Hsieh,T.S. (1985). Drosophila topoisomerase II double-strand DNA cleavage: analysis of DNA sequence homology at the cleavage site. *Nucleic Acids Res.* 13, 1057-1072.
- Sargent,R.G., Brenneman,M.A., Wilson,J.H. (1997). Repair of site-specific double-strand breaks in a mammalian chromosome by homologous and illegitimate recombination. *Mol.Cell Biol.* 17, 267-277.
- Saugier-Veber,P., Munnich,A., Bonneau,D., Rozet,J.M., Le Merrer,M., Gil,R., Boespflug-Tanguy,O. (1994). X-linked spastic paraplegia and Pelizaeus-Merzbacher disease are allelic disorders at the proteolipid protein locus. *Nat Genet* 6, 257-62.
- Savage,J.J., Yaden,B.C., Kiratipranon,P., Rhodes,S.J. (2003). Transcriptional control during mammalian anterior pituitary development. *Gene* 319, 1-19.
- Schepers,G.E., Teasdale,R.D., Koopman,P. (2002). Twenty pairs of sox: extent, homology, and nomenclature of the mouse and human sox transcription factor gene families. *Dev.Cell* 3, 167-170.
- Schneider,A., Montague,P., Griffiths,I., Fanarraga,M., Kennedy,P., Brophy,P., Nave,K.A. (1992). Uncoupling of hypomyelination and glial cell death by a mutation in the proteolipid protein gene. *Nature* 358, 758-61.
- Schouten,J.P., McElgunn,C.J., Waaijer,R., Zwiijnenburg,D., Diepvens,F., Pals,G. (2002). Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res* 30, e57.
- Schuchert,P., Langsford,M., Kaslin,E., Kohli,J. (1991). A specific DNA sequence is required for high frequency of recombination in the ade6 gene of fission yeast. *EMBO J* 10, 2157-2163.
- Schwab,M. (1999). Oncogene amplification in solid tumors. *Semin.Cancer Biol.* 9, 319-325.

Schwab,M., Alitalo,K., Klempnauer,K.H., Varmus,H.E., Bishop,J.M., Gilbert,F., Brodeur,G., Goldstein,M., Trent,J. (1983). Amplified DNA with limited homology to myc cellular oncogene is shared by human neuroblastoma cell lines and a neuroblastoma tumour. *Nature* 305, 245-248.

Schwartz,S., Kent,W.J., Smit,A., Zhang,Z., Baertsch,R., Hardison,R.C., Haussler,D., Miller,W. (2003). Human-mouse alignments with BLASTZ. *Genome Res.* 13, 103-107.

Schwartz,S., Zhang,Z., Frazer,K.A., Smit,A., Riemer,C., Bouck,J., Gibbs,R., Hardison,R., Miller,W. (2000). PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res* 10, 577-86.

Shaw,C.J., Lupski,J.R. (2004). Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Hum Mol.Genet* 13 Spec No 1, R57-R64.

Shiroishi,T., Koide,T., Yoshino,M., Sagai,T., Moriwaki,K. (1995). Hotspots of homologous recombination in mouse meiosis. *Adv.Biophys.* 31, 119-132.

Shuman,S. (1991). Site-specific DNA cleavage by vaccinia virus DNA topoisomerase I. Role of nucleotide sequence and DNA secondary structure. *J Biol.Chem.* 266, 1796-1803.

Simons,M., Kramer,E.M., Macchi,P., Rathke-Hartlieb,S., Trotter,J., Nave,K.A., Schulz,J.B. (2002). Overexpression of the myelin proteolipid protein leads to accumulation of cholesterol and proteolipid protein in endosomes/lysosomes: implications for Pelizaeus-Merzbacher disease. *J Cell Biol* 157, 327-36.

Simons,M., Kramer,E.M., Thiele,C., Stoffel,W., Trotter,J. (2000). Assembly of myelin by association of proteolipid protein with cholesterol- and galactosylceramide-rich membrane domains. *J.Cell Biol.* 151, 143-154.

Singh,G.B., Kramer,J.A., Krawetz,S.A. (1997). Mathematical model to predict regions of chromatin attachment to the nuclear matrix. *Nucleic Acids Res.* 25, 1419-1425.

Singleton,A.B., Farrer,M., Johnson,J., Singleton,A., Hague,S., Kachergus,J., Hulihan,M., Peuralinna,T., Dutra,A., Nussbaum,R., Lincoln,S., Crawley,A., Hanson,M., Maraganore,D., Adler,C., Cookson,M.R., Muenter,M., Baptista,M., Miller,D., Blancato,J., Hardy,J., Gwinn-Hardy,K. (2003). alpha-Synuclein locus triplication causes Parkinson's disease. *Science* 302, 841.

Sismani,C., Armour,J.A., Flint,J., Girgalli,C., Regan,R., Patsalis,P.C. (2001). Screening for subtelomeric chromosome abnormalities in children with idiopathic mental retardation using multiprobe telomeric FISH and the new MAPH telomeric assay. *Eur J Hum Genet* 9, 527-32.

Sistmans,E.A., de Coe,R.F., De Wijs,I.J., Van Oost,B.A. (1998). Duplication of the proteolipid protein gene is the major cause of Pelizaeus-Merzbacher disease. *Neurology* 50, 1749-54.

Sistmans,E.A., De Wijs,I.J., de Coe,R.F., Smit,L.M., Menko,F.H., Van Oost,B.A. (1996). A (G-to-A) mutation in the initiation codon of the proteolipid protein gene causing a relatively mild form of Pelizaeus-Merzbacher disease in a Dutch family. *Hum Genet* 97, 337-9.

Slater,H.R., Bruno,D.L., Ren,H., Pertile,M., Schouten,J.P., Choo,K.H. (2003). Rapid, high throughput prenatal detection of aneuploidy using a novel quantitative method (MLPA). *J.Med.Genet.* 40, 907-912.

Smahi,A., Courtois,G., Vabres,P., Yamaoka,S., Heuertz,S., Munnich,A., Israel,A., Heiss,N.S., Klauck,S.M., Kioschis,P., Wiemann,S., Poustka,A., Esposito,T., Bardaro,T., Gianfrancesco,F., Ciccodicola,A., D'Urso,M., Woffendin,H., Jakins,T., Donnai,D., Stewart,H., Kenwrick,S.J., Aradhya,S., Yamagata,T., Levy,M., Lewis,R.A., Nelson,D.L. (2000). Genomic rearrangement in NEMO impairs NF-kappaB activation and is a cause of incontinentia pigmenti. The International Incontinentia Pigmenti (IP) Consortium. *Nature* 405, 466-472.

Small,K., Iber,J., Warren,S.T. (1997). Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats. *Nat Genet* 16, 96-99.

Small,K., Warren,S.T. (1998). Emerin deletions occurring on both Xq28 inversion backgrounds. *Hum Mol Genet* 7, 135-9.

Smit,A.F. (1993). Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucleic Acids Res.* 21, 1863-1872.

Smit,A.F., Riggs,A.D. (1995). MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Res.* 23, 98-102.

Smit,A.F., Toth,G., Riggs,A.D., Jurka,J. (1995). Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J Mol.Biol.* 246, 401-417.

Smith,A.C., McGavran,L., Robinson,J., Waldstein,G., Macfarlane,J., Zonona,J., Reiss,J., Lahr,M., Allen,L., Magenis,E. (1986). Interstitial deletion of (17)(p11.2p11.2) in nine patients. *Am.J.Med.Genet.* 24, 393-414.

Smith,G.P. (1976). Evolution of repeated DNA sequences by unequal crossover. *Science* 191, 528-535.

Smith,G.R., Kunes,S.M., Schultz,D.W., Taylor,A., Triman,K.L. (1981). Structure of chi hotspots of generalized recombination. *Cell* 24, 429-436.

Smith,R.A., Ho,P.J., Clegg,J.B., Kidd,J.R., Thein,S.L. (1998). Recombination breakpoints in the human beta-globin gene cluster. *Blood* 92, 4415-4421.

Solomon,N.M., Nouri,S., Warne,G.L., Lagerstrom-Fermer,M., Forrest,S.M., Thomas,P.Q. (2002). Increased gene dosage at Xq26-q27 is associated with X-linked hypopituitarism. *Genomics* 79, 553-9.

Southwood,C., Gow,A. (2001). Molecular pathways of oligodendrocyte apoptosis revealed by mutations in the proteolipid protein gene. *Microsc.Res.Tech.* 52, 700-708.

Southwood,C.M., Garbern,J., Jiang,W., Gow,A. (2002). The unfolded protein response modulates disease severity in Pelizaeus-Merzbacher disease. *Neuron* 36, 585-96.

Sperry,A.O., Blasquez,V.C., Garrard,W.T. (1989). Dysfunction of chromosomal loop attachment sites: illegitimate recombination linked to matrix association regions and topoisomerase II. *Proc.Natl.Acad.Sci.U.S.A* 86, 5497-5501.

Spitzner,J.R., Chung,I.K., Muller,M.T. (1990). Eukaryotic topoisomerase II preferentially cleaves alternating purine-pyrimidine repeats. *Nucleic Acids Res.* 18, 1-11.

Spitzner,J.R., Muller,M.T. (1988). A consensus sequence for cleavage by vertebrate DNA topoisomerase II. *Nucleic Acids Res.* 16, 5533-5556.

Sporkel,O., Uschkureit,T., Bussow,H., Stoffel,W. (2002). Oligodendrocytes expressing exclusively the DM20 isoform of the proteolipid protein gene: myelination and development. *Glia* 37, 19-30.

Squillace,R.M., Chenault,D.M., Wang,E.H. (2002). Inhibition of muscle differentiation by the novel muscleblind-related protein CHCR. *Dev.Biol.* 250, 218-230.

Stankiewicz,P., Lupski,J.R. (2002). Genome architecture, rearrangements and genomic disorders. *Trends Genet* 18, 74-82.

Stary,A., Sarasin,A. (1992). Molecular analysis of DNA junctions produced by illegitimate recombination in human cells. *Nucleic Acids Res.* 20, 4269-4274.

Stecca,B., Southwood,C.M., Gragerov,A., Kelley,K.A., Friedrich,V.L.Jr., Gow,A. (2000). The evolution of lipophilin genes from invertebrates to tetrapods: DM-20 cannot replace proteolipid protein in CNS myelin. *J Neurosci* 20, 4002-10.

Steinmetz,M., Stephan,D., Fischer,L.K. (1986). Gene organization and recombinational hotspots in the murine major histocompatibility complex. *Cell* 44, 895-904.

Stevanovic,M., Lovell-Badge,R., Collignon,J., Goodfellow,P.N. (1993). SOX3 is an X-linked gene related to SRY. *Hum.Mol.Genet.* 2, 2013-2018.

Stork,P.J. (2003). Does Rap1 deserve a bad Rap? *Trends Biochem.Sci.* 28, 267-275.

Sugawara,H., Harada,N., Ida,T., Ishida,T., Ledbetter,D.H., Yoshiura,K., Ohta,T., Kishino,T., Niikawa,N., Matsumoto,N. (2003). Complex low-copy repeats associated with a common polymorphic inversion at human chromosome 8p23. *Genomics* 82, 238-244.

Takata,M., Sasaki,M.S., Sonoda,E., Morrison,C., Hashimoto,M., Utsumi,H., Yamaguchi-Iwai,Y., Shinohara,A., Takeda,S. (1998). Homologous recombination and non-homologous end-joining pathways of DNA double-strand break repair have overlapping roles in the maintenance of chromosomal integrity in vertebrate cells. *EMBO J* 17, 5497-5508.

Takuma,N., Sheng,H.Z., Furuta,Y., Ward,J.M., Sharma,K., Hogan,B.L., Pfaff,S.L., Westphal,H., Kimura,S., Mahon,K.A. (1998). Formation of Rathke's pouch requires dual induction from the diencephalon. *Development* 125, 4835-4840.

Tatusova,T.A., Madden,T.L. (1999). BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol.Lett.* 174, 247-250.

Telenius,H., Carter,N.P., Bebb,C.E., Nordenskjold,M., Ponder,B.A., Tunnacliffe,A. (1992). Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* 13, 718-25.

Ten Hagen,K.G., Gilbert,D.M., Willard,H.F., Cohen,S.N. (1990). Replication timing of DNA sequences associated with human centromeres and telomeres. *Mol.Cell Biol.* 10, 6348-6355.

Thapar,H., Kovacs,K., Horvath,E. (2001). Morphology of the Pituitary in Health and Disease. In: Principles and Practice of Endocrinology and Metabolism, ed. K.L.Becker Lippincott, Williams & Wilkins, 103-129.

The BRCA1 Exon 13 Duplication Screening Group (2000). The exon 13 duplication in the BRCA1 gene is a founder mutation present in geographically diverse populations. *Am.J.Hum.Genet.* 67, 207-212.

Thiel,C.T., Kraus,C., Rauch,A., Ekici,A.B., Rautenstrauss,B., Reis,A. (2003). A new quantitative PCR multiplex assay for rapid analysis of chromosome 17p11.2-12 duplications and deletions leading to HMSN/HNPP. *Eur.J.Hum.Genet.* 11, 170-178.

Thompson,J.D., Higgins,D.G., Gibson,T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673-4680.

Toffolatti,L., Cardazzo,B., Nobile,C., Danieli,G.A., Gualandi,F., Muntoni,F., Abbs,S., Zanetti,P., Angelini,C., Ferlini,A., Fanin,M., Patarnello,T. (2002). Investigating the mechanism of chromosomal deletion: characterization of 39 deletion breakpoints in introns 47 and 48 of the human dystrophin gene. *Genomics* 80, 523-530.

Tominaga,K., Leung,J.K., Rookard,P., Echigo,J., Smith,J.R., Pereira-Smith,O.M. (2003). MRGX is a novel transcriptional regulator that exhibits activation or repression of the B-myb promoter in a cell type-dependent manner. *J.Biol.Chem.* 278, 49618-49624.

Toriello,H.V., Glover,T.W., Takahara,K., Byers,P.H., Miller,D.E., Higgins,J.V., Greenspan,D.S. (1996). A translocation interrupts the COL5A1 gene in a patient with Ehlers-Danlos syndrome and hypomelanosis of Ito. *Nat.Genet* 13, 361-365.

Torpy,D.J., Jackson,R.V. (2001). Adrenocorticotropin: Physiology and Clinical Aspects. In: Principles and Practice of Endocrinology and Metabolism, ed. K.L.Becker Lippincott, Williams & Wilkins, 153-159.

Tosic M, Dolivo M, Domanska-Janik K, M,M.J. (1994). Paralytic tremor (pt): a new allele of the proteolipid protein gene in rabbits. *J Neurochem* 63, 2210-6.

Tosic,M., Matthey,B., Gow,A., Lazzarini,R.A., Matthieu,J.M. (1997). Intracellular transport of the DM-20 bearing shaking pup (shp) mutation and its possible phenotypic consequences. *J Neurosci Res* 50, 844-52.

Treier,M., Gleiberman,A.S., O'Connell,S.M., Szeto,D.P., McMahon,J.A., McMahon,A.P., Rosenfeld,M.G. (1998). Multistep signaling requirements for pituitary organogenesis in vivo. *Genes Dev.* 12, 1691-1704.

Uematsu,A., Yorifuji,T., Muroi,J., Kawai,M., Mamada,M., Kaji,M., Yamanaka,C., Momoi,T., Nakahata,T. (2002). Parental origin of normal X chromosomes in Turner syndrome patients with various karyotypes: implications for the mechanism leading to generation of a 45,X karyotype. *Am J Med.Genet* 111, 134-139.

Uematsu,Y., Kiefer,H., Schulze,R., Fischer-Lindahl,K., Steinmetz,M. (1986). Molecular characterization of a meiotic recombinational hotspot enhancing homologous equal crossing-over. *EMBO J* 5, 2123-2129.

Uhlenberg,B., Schuelke,M., Ruschendorf,F., Ruf,N., Kaindle,A.M., Henneke,M., Thiele,H., Stoltenburg-Didinger,G., Aksu,F., Topaloglu,H., Nurnberg,P., Hubner,C., Weschke,B., Gartner,J. (2004). Mutations in the Gene Encoding Gap Junction Protein alpha 12 (Connexin 46.6) Cause Pelizaeus-Merzbacher-Like Disease. *Am J Hum Genet* 75.

Ussery,D., Soumpasis,D.M., Brunak,S., Staerfeldt,H.H., Worning,P., Krogh,A. (2002). Bias of purine stretches in sequenced chromosomes. *Comput.Chem.* 26, 531-541.

Van Arsdell,S.W., Weiner,A.M. (1984). Human genes for U2 small nuclear RNA are tandemly repeated. *Mol.Cell Biol.* 4, 492-499.

Vaurs-Barriere,C., Wong,K., Weibel,T.D., Abu-Asab,M., Weiss,M.D., Kaneski,C.R., Mixon,T.H., Bonavita,S., Creveaux,I., Heiss,J.D., Tsokos,M., Goldin,E., Quarles,R.H., Boespflug-Tanguy,O., Schiffmann,R. (2003). Insertion of mutant proteolipid protein results in missorting of myelin proteins. *Ann.Neurol.* 54, 769-780.

Veltman,J.A., Yntema,H.G., Lugtenberg,D., Arts,H., Briault,S., Huys,E.H., Osoegawa,K., de Jong,P., Brunner,H.G., Geurts,v.K., Van Bokhoven,H., Schoenmakers,E.F. (2004). High resolution profiling of X chromosomal aberrations by array comparative genomic hybridisation. *J Med.Genet* 41, 425-432.

Venter,J.C., *et al.* (2001). The sequence of the human genome. *Science* 291, 1304-1351.

Vetcher,A.A., Napierala,M., Wells,R.D. (2002). Sticky DNA: effect of the polypurine.polypyrimidine sequence. *J Biol.Chem.* 277, 39228-39234.

Viguera,E., Canceill,D., Ehrlich,S.D. (2001). Replication slippage involves DNA polymerase pausing and dissociation. *EMBO J* 20, 2587-2595.

Vissers,L.E., de Vries,B.B., Osoegawa,K., Janssen,I.M., Feuth,T., Choy,C.O., Straatman,H., van,d., V, Huys,E.H., van Rijk,A., Smeets,D., Ravenswaaij-Arts,C.M., Knoers,N.V., van,d.B., I, de Jong,P.J., Brunner,H.G., van Kessel,A.G., Schoenmakers,E.F., Veltman,J.A. (2003). Array-based comparative genomic hybridization for the genomewide detection of submicroscopic chromosomal abnormalities. *Am.J.Hum.Genet.* 73, 1261-1270.

Vogelstein,B., Pardoll,D.M., Coffey,D.S. (1980). Supercoiled loops and eucaryotic DNA replicaton. *Cell* 22, 79-85.

Wahls,W.P., Wallace,L.J., Moore,P.D. (1990). Hypervariable minisatellite DNA is a hotspot for homologous recombination in human cells. *Cell* 60, 95-103.

Wang,A.H., Quigley,G.J., Kolpak,F.J., Crawford,J.L., van Boom,J.H., van der,M.G., Rich,A. (1979). Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature* 282, 680-686.

Wang,B., Dickinson,L.A., Koivunen,E., Ruoslahti,E., Kohwi-Shigematsu,T. (1995). A novel matrix attachment region DNA binding motif identified using a random phage peptide library. *J Biol.Chem.* 270, 23239-23242.

Wang,P.J., Hwu,W.L., Lee,W.T., Wang,T.R., Shen,Y.Z. (1997). Duplication of proteolipid protein gene: a possible major cause of Pelizaeus-Merzbacher disease. *Pediatr Neurol* 17, 125-8.

Weaver,D.T., DePamphilis,M.L. (1982). Specific sequences in native DNA that arrest synthesis by DNA polymerase alpha. *J Biol.Chem.* 257, 2075-2086.

Weimbs,T., Stoffel,W. (1992). Proteolipid protein (PLP) of CNS myelin: positions of free, disulfide-bonded, and fatty acid thioester-linked cysteine residues and implications for the membrane topology of PLP. *Biochemistry* 31, 12289-96.

Weinreb,A., Katzenberg,D.R., Gilmore,G.L., Birshtein,B.K. (1988). Site of unequal sister chromatid exchange contains a potential Z-DNA-forming tract. *Proc.Natl.Acad.Sci.U.S.A* 85, 529-533.

Westin,G., Zabielski,J., Hammarstrom,K., Monstein,H.J., Bark,C., Pettersson,U. (1984). Clustered genes for human U2 RNA. *Proc.Natl.Acad.Sci.U.S.A* 81, 3811-3815.

White,S., Kalf,M., Liu,Q., Villerius,M., Engelsma,D., Kriek,M., Vollebregt,E., Bakker,B., van Ommen,G.J., Breuning,M.H., den Dunnen,J.T. (2002). Comprehensive detection of genomic duplications and deletions in the DMD gene, by use of multiplex amplifiable probe hybridization. *Am J Hum Genet* 71, 365-74.

White,S.J., Sterrenburg,E., van Ommen,G.J., den Dunnen,J.T., Breuning,M.H. (2003). An alternative to FISH: detecting deletion and duplication carriers within 24 hours. *J.Med.Genet.* 40, e113.

Wight,P.A., Dobretsova,A. (2004). Where, when and how much: regulation of myelin proteolipid protein gene expression. *Cell Mol.Life Sci.* 61, 810-821.

Williams,T.J., Fried,M. (1986). Inverted duplication-transposition event in mammalian cells at an illegitimate recombination join. *Mol.Cell Biol.* 6, 2179-2184.

Wolf,N.I., Sistermans,E.A., Cundall,M., Palmer,R., Stubbs,P., Hobson,G.M., Davis-Williams,A.P., Sperle,K., Garbern,J., Baty,D., Davies,S., Endziniene,M., Chong,K., Malcolm,S., Woodward,K. (2004) Triplication of the proteolipid protein gene is associated with a severe form of Pelizaeus-Merzbacher disease. (In preparation).

Woodfine,K., Fiegler,H., Beare,D.M., Collins,J.E., McCann,O.T., Young,B.D., Debernardi,S., Mott,R., Dunham,I., Carter,N.P. (2004). Replication timing of the human genome. *Hum Mol.Genet* 13, 191-202.

Woodward,K., Cundall,M., Palmer,R., Surtees,R., Winter,R.M., Malcolm,S. (2003). Complex chromosomal rearrangement and associated counseling issues in a family with Pelizaeus-Merzbacher disease. *Am J Med Genet* 118A, 15-24.

Woodward,K., Cundall,M., Sperle,K., Sistermans,E.A., Garbern,J., Kamholtz,J., Ross,M., Howell,G., Carter,N., Gribble,S., Burford,D., Heng,H., Malcolm,S., Hobson,G.M. (2004). Heterogeneous Xq22 duplications in Pelizaeus-Merzbacher disease: breakpoint junctions suggest mechanism of nonhomologous end-joining. (In preparation)

Woodward,K., Kendall,E., Vetrie,D., Malcolm,S. (1998). Pelizaeus-Merzbacher disease: identification of Xq22 proteolipid- protein duplications and characterization of breakpoints by interphase FISH. *Am J Hum Genet* 63, 207-17.

- Woodward,K., Kirtland,K., Dlouhy,S., Raskind,W., Bird,T., Malcolm,S., Abeliovich,D. (2000). X inactivation phenotype in carriers of Pelizaeus-Merzbacher disease: skewed in carriers of a duplication and random in carriers of point mutations. *Eur J Hum Genet* 8, 449-54.
- Wu,W., Cogan,J.D., Pfaffle,R.W., Dasen,J.S., Frisch,H., O'Connell,S.M., Flynn,S.E., Brown,M.R., Mullis,P.E., Parks,J.S., Phillips,J.A., III, Rosenfeld,M.G. (1998). Mutations in PROP1 cause familial combined pituitary hormone deficiency. *Nat.Genet* 18, 147-149.
- Yamamoto,T., Nanba,E., Zhang,H., Sasaki,M., Komaki,H., Takeshita,K. (1998). Jimpy(msd) mouse mutation and connatal Pelizaeus-Merzbacher disease [letter]. *Am J Med Genet* 75, 439-40.
- Yan,Y., Lagenaur,C., Narayanan,V. (1993). Molecular cloning of M6: identification of a PLP/DM20 gene family. *Neuron* 11, 423-31.
- Yang,Q.S., Xia,F., Gu,S.H., Yuan,H.L., Chen,J.Z., Ying,K., Xie,Y., Mao,Y.M. (2002). Cloning and expression pattern of a spermatogenesis-related gene, BEX1, mapped to chromosome Xq22. *Biochem Genet* 40, 1-12.
- Yeh,C.H., Shatkin,A.J. (1994). A HeLa-cell-encoded p21 is homologous to transcription elongation factor SII. *Gene* 143, 285-287.
- Yoshida,M., Colman,D.R. (1996). Parallel evolution and coexpression of the proteolipid proteins and protein zero in vertebrate myelin. *Neuron* 16, 1115-26.
- Yu,A., Zhao,C., Fan,Y., Jang,W., Mungall,A.J., Deloukas,P., Olsen,A., Doggett,N.A., Ghebranious,N., Broman,K.W., Weber,J.L. (2001). Comparison of human genetic and sequence-based physical maps. *Nature* 409, 951-953.
- Zimmerer,E.J., Passmore,H.C. (1991). Structural and genetic properties of the Eb recombinational hotspot in the mouse. *Immunogenetics* 33, 132-140.

APPENDIX A

List of primers used for this thesis. The sequence of each primer is given, the accession number and position where the primer hybridises, the size of the amplification product and the basic reaction conditions.

Primer name	Primer sequence	Accession	Position within clone	Annealing temperature	Enzyme	Size of product (bp)	Other reaction conditions
Family 1 primers							
cU65A4F14986	TCTGCCTTAATTTTCATTATTTACCCAAGTG	Z81014	14986-15015	60	H + PS	-	LR-PCR
cU65A4F15564	CCTCAGAATTGCTTGTTGAACAATATCT	Z81014	15564-15593	60	H + PS	-	LR-PCR
dJ839M11F10232	CAAAATAGTAGCTGATGACCTACCCACAG	AL034485	10232-10260	60	H + PS	-	LR-PCR
cV362H12F41101	GATCTTAAAGTCTGACAGGAAACAGCATT	Z70227	41101-41133	60	H + PS	-	LR-PCR
dJ839M11F3171	AGTCAAACATTTTATCAGCTTTCCTGTAA	AL034485	3171-3200	60	H + PS	-	LR-PCR
dJ839M11F8857	CTCTGTGGAATACAGATCTAGAAGGTGGTT	AL034485	8857-8886	60	H + PS	-	LR-PCR
dJ839M11F10069	AGAGAAAGCTGTACGCAATAAAGACATAGG	AL034485	10069-10098	60	H + PS	-	LR-PCR
dJ839M11F12790	ATAAGTAATGATTCTTCTGAGGTGGCTCTG	AL034485	12790-12819	60	H + PS	-	LR-PCR
dJ839M11F13586	GAATCCGACTCCTAACCTCTCATCTCTATT	AL034485	13586-13615	60	H + PS	-	LR-PCR
dJ839M11F19825	AGCATCCCTAAAACTAAAAGTGCTTTCTC	AL034485	19825-19854	60	H + PS	-	LR-PCR
dJ839M11F25035	AACTTTCTCCCTGTAAACCTCTCTATGAAC	AL034485	25035-25064	60	H + PS	-	LR-PCR
dJ635G19R67455	CTTGTCAGTCTTTCTCCTCTTAAGTTCTGG	AL035494	67455-67426	60	H + PS	-	LR-PCR
cU65A4F11529	TCTTTTGGAGGTGATTAAAGTGTGCTAT	Z81014	11529-11558	60	H + PS	-	LR-PCR
cU65A4F10369	ATTTTGGGGGTTACTGTGAGTAATGTTATG	Z81014	10369-10398	60	H + PS	-	LR-PCR
cU65A4F7943	GCTTAGTATGCTGTAGAGGCAAGAACCTAA	Z81014	7943-7972	60	H + PS	-	LR-PCR
Family 2 primers							
dJ1055C14R60006	GGGTACCCAGTTAGCTGGACAAAAC	AL049610	60006-59982	55	H	-	-
177e8F37219	CCCATAGGCATGTCAATAACAATGA	Z68694	37219-37243	55	H	-	-
177e8R36049	TGTCTTCTCCCCCTCATCCTCT	Z68694	36050-36028	55	H	-	-
iPCR177e8R35970	CCTGAATAAGAGGAATTTACTCACACCTTG	Z68694	35970-35941	60	H	-	iPCR
iPCR177e8F36264	ACATAGACCTGGGGGTAGCTTGTAATAAT	Z68694	36264-36293	60	H	-	iPCR
Family 3 primers							
6MW	CCGACTCGAGNNNNNNATGTGG	-	-	55	H	-	DOP-PCR
102P23AF	GCCTTTCATTGCACATTTCA	AL357115	132693-132712	53	H	178	35 cycles
102P23AR	ACTTGGCATCCCACTTTCAT		132870-132851				
RH63626F	AAAAAGACAAGAAACATGGCC	AL357115	133192-133212	55	H	203	35 cycles
RH63626R	GCATAGCATCAAGTTTGAATCTCA		133394-133371				
346E8CF	CCTTTGGGCTGCTACCATTA	AL359885	952-971	55	H	191	35 cycles
346E8CR	TTCGGACTGTGTGTGTGTGA		1142-1123				

Primer name	Primer sequence	Accession	Position within clone	Annealing temperature	Enzyme	Size of product (bp)	Other reaction conditions
346E8F8386	TAAGAGAGACTGGACCTCTCTCCATATAGT	AL359885	8386-8415	60	H	377	40 cycles
346E8R8763	TGGGAGGGTACACAGTGATTAGAGAGTAAT		8763-8732				
346E8F17767	TATTTTCTGCTGTTTTATTTCTTTTCTCA	AL359885	17767-17796	60	H	204	35 cycles
346E8R17970	AAGGATAGACACAGACATATTCCTAAGGTT		17970-17941				
346E8AF	TCCCTGCCTGTCCTACACTT	AL359885	26616-26635	53	H	199	35 cycles
346E8AR	TCAATTTCCATTCCCCACAT		26814-26795				
346E8BF	AATTGGCATCTTGCATTTC	AL359885	69417-69436	55	B	205	35 cycles
346E8BR	GATACATCCTGGGAGCTGGA		69621-69602				
52K8AF	TGCATGGGGAATTAAGCATT	AL445213	23495-23514	55	B	205	35 cycles
52K8AR	TGTAGGAAAATGGCAAACA		23699-23680				
DXS8045F-TET	ACTGCGGTGCTGACTAGG	AL109654	159160-159177	55	B	189-223	-
DXS8045R	CAGGTAAATCTGAGAAATGTTCTGC		159378-159354				
183K14CF	GGTTTCATTTTTGCCCTCT	AL109913	66090-66109	55	B	222	-
183K14CR	AAGGGGAATAAATGGTGA		66311-66292				
183K14BF	TTTATGTGCTCCCCCTCAAC	AL109913	84019-84038	55	B	218	-
183K14BR	TTGTGGGAAATGGATGTGAA		84236-84217				
183K14DF	GAACGAAGGAAGCAAACAGG	AL109913	129708-129727	55	B	193	-
183K14DR	ACAAGCGGGTTCCTACCTCT		129900-129881				
79A21BF	CTGAAACGAAGTGGGGAGAG	AL513491	10058-10077	55	B	203	-
79A21BR	GCATCTTGCATTTTTCTCCA		10260-10241				
203P18CF	AAGGGGAGCAGGACTAGCTT	Z97180	126349-126368	56	B	196	-
203P18CR	GGGAAATCAGTCACAACAGGA		126544-126525				
203P18DF	GGCCCACATTAGCAATTCAC	Z97180	116365-116384	56	H	223	-
203P18DR	ACAAGATTCTGGCCCCTAC		116587-116568				
203P18F105705	GGTATTAGAATTGCTAGTGACCTCTTCTCT	Z97180	105705-105734	60	H	340	-
203P18R106044	ACATGTTTCGAGTACATTCTAACAGCTAAGT		106044-106015				
203P18F98450	TATCTGCCATGAGTAGGGTTATTGTGACTA	Z97180	98450-98479	60	H	378	40 cycles
203P18R98827	TATAGAGATCCAGGCCAGACAATAAAGAAA		98827-98798				
203P18F97768	TTAAAACTGAACCTATTGCCATC	Z97180	97768-97791	57	H	159	40 cycles
203P18R97926	GGCAAAATTGGGAAGATTGTTACTA		97926-97902				

Primer name	Primer sequence	Accession	Position within clone	Annealing temperature	Enzyme	Size of product (bp)	Other reaction conditions
203P18F96708	GGACGATATCAACTAGGATGTTGGA	Z97180	96708-96732	57	H	239	40 cycles
203P18R96946	GATTCTGTGCAAGAAGAGCTAAACG		96946-96922				
203P18F95518	ATTTGCCTCTTTTCAACTTTCTAAACTCTA	Z97180	95518-95547	60	H	240	40 cycles
203P18R95757	TCTAAGACATAATGAATTTAAAGCTGAGTTG		95757-95727				
203P18IF	TTAATGTGAGCCCTGGGAAC	Z97180	95368-95387	55	H	390	-
203P18IR	CAGAACAAGGGCAAGGTAG		95757-95738				
203P18GF	TTCAAAAGTCTGAGGCACCA	Z97180	87737-87758	55	H	198	-
203P18GR	AATGGGGTGTCTGAATCTCCT		87934-87915				
DXS1200F	TACACACCAAACAACAGAGCCT	Z97180	79512-79491	55	H	275-285	-
DXS1200R-TET	CTAGGGGGCACTTGAAAACAA		79234-79254				
203P18HF	TTGAGTGTGCAAGCCAGAT	Z97180	74979-74989	55	H	192	-
203P18HR	ACACTACCGTGGTGCCAAAT		75170-75151				
203P18FF	TGTGCTTTTGTGCCAAATC	Z97180	53464-53484	55	B	200	-
203P18FR	GGAATTCTGAAACCCTGCAC		53663-53644				
203P18BF	AAGAGCTTCATGCCATGTCC	Z97180	4745-4764	55	H	184	-
203P18BR	TGGTTTGTGACCCTTTTGG		4928-4909				
346E8F1446	TAGTTTCACAAAAATGCTCTAAGCTGTCAA	AL359885	1446-1475	63	H + PS	-	LR-PCR
346E8F2596	TCTTGCTATACAGGCTCTTTTGTAGTTCTG	AL359885	2596-2625	63	H + PS	-	LR-PCR
346E8F4250	GAGAGACTGCATCCTTTTCTTGTCTAGTC	AL359885	4250-4279	63	H + PS	-	LR-PCR
346E8F4815	GCTGTGAATTTCCCTGGTTC	AL359885	4815-4834	63	H + PS	-	LR-PCR
203P18F98454	TGCCATGAGTAGGGTTATTGTG	Z97180	98454-98475	63	H + PS	-	LR-PCR
1055c14ipcr68617	TTTACTATGGCGGTTGGAGG	AL049610	68617-68598	55	H	-	40 cycles
1055c14ipcr68830	ACCAATCCTCTTCCTCCGTT	AL049610	68830-68849	55	H	-	40 cycles
305b16R44033	GGAGCCTTCCCACTAAAATCA	AL049792	44053-44033	55	H	-	-
197o17F69387	GGTAGTGGCAAGCACAGCAT	AL008629	69387-69406	55	H	-	-
453f18R24991	TGCTCCAGGTTGCTGATACA	AL513243	24991-24672	55	H	-	-
1055c14R68328	TAATAGTGCGTATAAGGCACGTAATGTGAA	AL049610	68328-68299	65	H + PS	-	LR-PCR
453f18R25585	GCACATGACCAATATGTACTCCAACAGTAT	AL513243	25585-25556	65	H + PS	-	LR-PCR
453f18R26962	TTGATTTTGGGAATCCAGATTACTACAAG	AL513243	26962-26933	65	H + PS	-	LR-PCR
453f18R29754	GTCTCTGTAGAATCTATGCGTTTCCCTAA	AL513243	29754-29725	65	H + PS	-	LR-PCR

Primer name	Primer sequence	Accession	Position within clone	Annealing temperature	Enzyme	Size of product (bp)	Other reaction conditions
Family 4 primers							
DXS6807F	GAGCAATGATCTCATTTGCA	AC073617	56049-56030	55	B	250-273	-
DXS6807R	AAGTAAACATGTATAGGAAAAAGCT		55783-55807				
DXS9895F	TTGGGTGGGGACACAGAG	AC120338	39035-39018	55	B	140-161	-
DXS9895R	CCTGGCTCAAGGAATTACAA		38893-38908				
AR1 5'FAM	GCTGTGAAGGTTGCTGTTCTCAT	AL049564	104443-104420	60	B	269-306	-
AR2	TCCAGAATCTGTTCCAGAGCGTGC		104156-104179				
DXS6800F	GTGGGACCTTGTGATTGTGT	AL009028	59056-59037	55	B	197-221	-
DXS6800R	CTGGCTGACACTTAGGGAAA		58863-58882				
DXS1191F	AACAGCTATTGTGCCTGGCAGAGAA	Z73497	23844-23865	55	B	233-245	-
DXS1191R	GCCCCGTTTGATGCTTCTAAATTG		24081-24058				
DXS8075F	GGGCTACCAAAGGACTGT	AL050401	66985-67002	55	B	167-177	-
DXS8075R	CTGGGTTGTGACTGTTTCAT		67155-67136				
DXS6797F	TTCCCTCTCTCCCTCTGTCT	AL136080	67519-67538	55	B	250-270	-
DXS6797R	ACACACACCCAAAACCAGAT		67787-67768				
DXS1212R	TGGCATTACAAGCCCTCAAGTC	Z82899	30107-30128	55	B	230-238	-
DXS1212F-HEX	TGGAAGCATGAGAAATCACATCCT		30338-30315				
DXS1232F	ACCAACAGCCTAATAATGC	AC004070	58530-58512	55	B	163-199	-
DXS1232R-HEX	AGAGATGGGAGCAGCA		58362-58377				
DXS8013F	CCAACCCAACTGTCTATCAA	AL121875	6392-6373	55	B	193-229	-
DXS8013R-FAM	GTTTGGTTTTCCATTCTGA		6169-6188				
DXS984F-FAM	TTTCTGTCTGCCAAGTGTTT	AL137016	3278-3297	55	B	154-184	-
DXS984R	TACTGNGCCCTACTCCATTC		3445-3426				
291b3F17375	AGGCTGAGCTGATGAGGATAGGAGT	AL844175	17375-17399	65	H + PS	-	LR-PCR
291b3F22910	TAATCCAGAGAGGTGTCGTTTTCCA	AL844175	22910-22934	65	H + PS	-	LR-PCR
291b3F25483	TGTCCTGTTGAATTTCACTCACCTG	AL844175	25483-25507	65	H + PS	-	LR-PCR
291b3F4486	ATTAAAGTCATTGCCGAGGAATCGT	AL844175	4486-4510	65	H + PS	-	LR-PCR
291b3F11639	GAGCTGAAAGATCATGATGCAGGTT	AL844175	11639-11663	65	H + PS	-	LR-PCR
35f15R71430	TTGCTCTTCACAAACTCGACCATT	AL590077	71430-71406	65	H + PS	-	LR-PCR
364b14R15613	CCCTTGCTGTATAGTGCTCTTCT	AL589987	15613-15589	65	H + PS	-	LR-PCR
364b14R9239	TTTTGGAAGGAAAAAGGCCTCAGTA	AL589987	9239-9215	65	H + PS	-	LR-PCR
364b14R6018	GTCTGTGTGTCAGACCCGATCTAGGA	AL589987	6018-5994	65	H + PS	-	LR-PCR

Primer name	Primer sequence	Accession	Position within clone	Annealing temperature	Enzyme	Size of product (bp)	Other reaction conditions
UPQFM primers							
dJ635g19F8797	tccgtcttagctgagtggcgtaGTTGGGGCCTTCTGTGTTTA	AL035494	8797-8816	55	H	211	UPQFM-PCR
dJ635g19R8962	aggcagaatcgactcaccgctaACTCACCTGGATGGTTCAG		8962-8943				
dJ635g19F30902	tccgtcttagctgagtggcgtaTCCTTCCTGCCTGACTCCTA	AL035494	30902-30921	55	H	301	UPQFM-PCR
dJ635g19R31157	aggcagaatcgactcaccgctaTCGCAGATGATGTCCCACTA		31157-31138				
dJ635g19F48059	tccgtcttagctgagtggcgtaGGTGGTCCCTCCCATTTAGT	AL035494	48059-48078	55	H	315	UPQFM-PCR
dJ635g19R48309	aggcagaatcgactcaccgctaATTGGTGTGCCAGATGTGAA		48328-48309				
dJ635G19F60570	tccgtcttagctgagtggcgtaAGAGTGGGTCTCTGGCTCAA	AL035494	60570-60589	55	H	203	UPQFM-PCR
dJ635g16R60728	aggcagaatcgactcaccgctaGGTGCCTCCAACCTTTTGTGT		60728-60709				
dJ635G19F68207	tccgtcttagctgagtggcgtaAGGACAGCCAGCAAGAAAAA	AL035494	68207-68226	55	H	293	UPQFM-PCR
dJ635G19R68454	aggcagaatcgactcaccgctaCCAGTTTCCTCCCCCTTTAG		68454-68435				
dJ635G19F69048	tccgtcttagctgagtggcgtaCCACCCCAATATCACTTTG	AL035494	69048-69067	55	H	263	UPQFM-PCR
dJ635G19R69265	aggcagaatcgactcaccgctaGCTTAGTTTCGCACCTGGAG		69265-69246				
cU65A4F11590	tccgtcttagctgagtggcgtaATTTTGATTGCGGTCATGGT	Z81014	11590-11609	55	H	212	UPQFM-PCR
cU65A4R11756	aggcagaatcgactcaccgctaCCAAAACAGTTTGCCACTTTC		11756-11736				
cU65A4F9959	tccgtcttagctgagtggcgtaGAAGCTTGGGGGAAAGAATC	Z81014	9959-9938	55	H	256	UPQFM-PCR
cU65A4R9747	aggcagaatcgactcaccgctaCAATCAGGGCATGAATGAGA		9747-9766				
cU65A4F6833	tccgtcttagctgagtggcgtaGGGCCATATGGAAGACAGAA	Z81014	6833-6852	55	H	320	UPQFM-PCR
cU65A4R7109	aggcagaatcgactcaccgctaGCCATCCAGTATTGGCACTT		7109-7088				
cU65a4F4480	tccgtcttagctgagtggcgtaGCTTGGAAGCTCCACTTGAC	Z81014	4480-4499	55	H	270	UPQFM-PCR
cU65A4R4685	aggcagaatcgactcaccgctaCATGTGGAGGAAGTGTGTGG		4704-4685				
cU177E8F36491	tccgtcttagctgagtggcgtaGTGAAACAGGGATGGGAAGAG	Z68694	36491 – 36510	55	H	267	UPQFM-PCR
cU177E8R36712	aggcagaatcgactcaccgctaCTAGTCCAGAGCATGGCACA		36712 – 36693				
cU177E8F38464	tccgtcttagctgagtggcgtaCTTCTCCCAACCCAGTACA	Z68694	38464 – 38483	55	H	240	UPQFM-PCR
cU177E8R38658	aggcagaatcgactcaccgctaCAGCACATGGCAGTTCCTTA		38658 – 38639				
cU177E8F20484	tccgtcttagctgagtggcgtaTCTTGGTGCACAAAAACAGG	Z68694	20484-20503	55	H	306	UPQFM-PCR
cU177E8R20744	aggcagaatcgactcaccgctaAGGCCTTCACCTTCATCCTT		20744-20725				
cU177E8F32435	tccgtcttagctgagtggcgtaCTTTGCCAGCAATGAGAACA	Z68694	32435-32454	55	H	304	UPQFM-PCR
cU177E8R32693	aggcagaatcgactcaccgctaACCTCTGGCCCTTGTGTAAA		32693-32674				
cU177E8F5951	tccgtcttagctgagtggcgtaAGACTCTAGCCACCAGCAA	Z68694	5951-5932	55	H	299	UPQFM-PCR
cU177e8R5698	aggcagaatcgactcaccgctaTGACGCTTGAGTCTG TGACC		5698-5717				

Primer name	Primer sequence	Accession	Position within clone	Annealing temperature	Enzyme	Size of product (bp)	Other reaction conditions
dJ43H13F12256	tccgtcttagctgagtggcgtaAGATGTGGATTTAGGGTCAACAGC	AL035444	12256-12279	55	H	297	UPQFM-PCR
dJ43H13R12507	aggcagaatcgactcaccgctaCAAGCTCCTAACCTCCATCTGTC		12507-12485				
cV857G6F11085	tccgtcttagctgagtggcgtaCTGTGCTAAACTGCGTGGA	Z73965	11085-11104	55	H	250	UPQFM-PCR
cV857G6R11289	aggcagaatcgactcaccgctaTGCAAAGCCAAGCTACAAGA		11289-11270				
cV857G6F41850	tccgtcttagctgagtggcgtaCTTTCCTTGAGTCATTGAGGTTTT	Z73965	41850-41873	55	H	277	UPQFM-PCR
cV857G6R42081	aggcagaatcgactcaccgctaTTGTTGAGTGATTCTGATCTCCAT		42081-42058				
cV857G6F32969	tccgtcttagctgagtggcgtaCACAAATGTCCATCATTACAGGTT	Z73965	32969-32992	55	H	282	UPQFM-PCR
cV857G6R33205	aggcagaatcgactcaccgctaGCATCTCTAGAGGGCAATATCACT		33205-33182				
cU246D9F8349	tccgtcttagctgagtggcgtaGTTGGTATCAAAATGGGCATATCAC	AL021308	8349-8372	55	H	249	UPQFM-PCR
cU246D9R9530	aggcagaatcgactcaccgctaGAACTCATGTGGCTCCTGAACTG		8552-9530				
cU246D9F16048	tccgtcttagctgagtggcgtaTGTCGATTGCATTATTCCTGTTA	AL021308	16048-16070	55	H	275	UPQFM-PCR
cU246D9R16277	aggcagaatcgactcaccgctaCTGCTCAGTGTTGACATCCATTT		16277-16255				
dJ1055C14F1	tccgtcttagctgagtggcgtaCATGCCCACTGTTATAGGG	AL049610	18844-18863	55	H	270	UPQFM-PCR
dJ1055C14R1	aggcagaatcgactcaccgctaGCCCCATTTCTATTGGGTGA		19070-19051				
dJ1055C14F3	tccgtcttagctgagtggcgtaGTAAACACACGGGCCTCACT	AL049610	46009-46028	55	H	274	UPQFM-PCR
dJ1055C14R3	aggcagaatcgactcaccgctaTTGCATGCGCTATGATCTTC		46237-46218				
dJ1055C14F4	tccgtcttagctgagtggcgtaGCATTTTAGGGCAGCAAGAG	AL049610	54449-54468	55	H	279	UPQFM-PCR
dJ1055C14R4	aggcagaatcgactcaccgctaTGTTTAGGGATGAGGTGCTC		54682-54663				
dJ1055C14F5	tccgtcttagctgagtggcgtaAGAGCCAAGCTGGACAAAA	AL049610	62946-62965	55	H	205	UPQFM-PCR
dJ1055C14R5	aggcagaatcgactcaccgctaGTAGGGAACAGGTGGCAAAA		63086-63105				
dJ1055C14F8	tccgtcttagctgagtggcgtaGGAAGGTTCTGCAATCAGGA	AL049610	67104-67123	55	H	273	UPQFM-PCR
dJ1055C14R8	aggcagaatcgactcaccgctaGTGGCCACAAGAAAGGAAAA		67331-67312				
dJ1055C14F10	tccgtcttagctgagtggcgtaCCTCCAACCGCCATAGTAAA	AL049610	68598-68617	55	H	297	UPQFM-PCR
dJ1055C14R10	aggcagaatcgactcaccgctaAACGGAGGAAGAGGATTGGT		68849-68830				
dJ1055C14F7	tccgtcttagctgagtggcgtaGCTCTTACCTTTGCGTCCTG	AL049610	69896-69915	55	H	306	UPQFM-PCR
dJ1055C14R7	aggcagaatcgactcaccgctaACCAAACAGCCAATCCTTTG		69955-69936				
dJ1055C14F6	tccgtcttagctgagtggcgtaGGTTTGAAACTGGGGACAGA	AL049610	75773-75792	55	H	275	UPQFM-PCR
dJ1055C14R6	aggcagaatcgactcaccgctaAAGGGCCAGGACTTTTGTTT		76033-76014				
dJ1055C14F2	tccgtcttagctgagtggcgtaCGCAATAGGAAAAGGGATGA	AL049610	88473-88492	55	H	220	UPQFM-PCR
dJ1055C14R2	aggcagaatcgactcaccgctaTCCTGCTGCAGATTCAAATG		88702-88683				
cU35F15F5631	tccgtcttagctgagtggcgtaTCCATGGGGAAGTTCTTGAG	Z93848	5631-5650	55	H	220	UPQFM-PCR
cU35F15R5805	aggcagaatcgactcaccgctaAGCCAGGAGAGTTTGCCTTA		5805-5786				

cV362H12F34770	tccgtcttagctgagtgccgtaGTTGCATCTGGTAGCCATCC	Z70227	34770-34963	55	H	239	UPQFM-PCR
cV362H12R34963	aggcagaatcgactcaccgctaAACGTGGGGCAAACTACTG		34963-34944				
dJ839M11F37922	tccgtcttagctgagtgccgtaGCTGTTCTCAGAGGTCGTC	AL034485	37922-37941	55	H	205	UPQFM-PCR
dJ839M11R38082	aggcagaatcgactcaccgctaTCCAATCAGACGTGAAGCTG		38082-38063				
cU240C4F29788	tccgtcttagctgagtgccgtaGAGGTGCCAACAGGGAATAG	Z73497	29788-29807	55	H	240	UPQFM-PCR
cU240C2R29983	aggcagaatcgactcaccgctaGGTCACGTGCCCTTACTTA		29983-29964				
cU46H11F32229	tccgtcttagctgagtgccgtaATTCCAGAGGCTCCATTCTT	Z82254	32229-32248	55	H	221	UPQFM-PCR
cU46H11R32404	aggcagaatcgactcaccgctaTCCTTGCCAGTTTTGACCTC		32404-32385				
35F15tag66750F	tccgtcttagctgagtgccgtaTAACCACCGTCCACTCACAA	AL590077	66750-66769	55	H	244	UPQFM-PCR
35F15tag66949R	aggcagaatcgactcaccgctaGAGCCTTGTTCTGTGGAAG		66949-66930				
364B14F16714	tccgtcttagctgagtgccgtaAGGATCTGGACCAACACAGG	AL589987	16714-16733	55	H	290	UPQFM-PCR
364B14R16959	aggcagaatcgactcaccgctaTGAAGGCAGGGAACCTCTTA		16959-16940				
364B14F25512	tccgtcttagctgagtgccgtaATGATCCAAGGGTTCATGA	AL589987	25512-25531	55	H	289	UPQFM-PCR
364B14R25756	aggcagaatcgactcaccgctaACTGGCAGGAATTGGCTAAC		25756-25737				
364B14F48225	tccgtcttagctgagtgccgtaGATTTCCACCTACTGCTGGA	AL589987	48225-48244	55	H	212	UPQFM-PCR
364b14R48392	aggcagaatcgactcaccgctaAATGCTGCAGGAGCCTAAGA		48392-48373				
364B14tag74126F	tccgtcttagctgagtgccgtaTTTCCAGTGGGCTTGTTAG	AL589987	74126-74145	55	H	288	UPQFM-PCR
364B14tag74369R	aggcagaatcgactcaccgctaTAGAGGGCTCTGCGGAATTA		74369-74350				
595A18F26757	tccgtcttagctgagtgccgtaTTTGTITGCTGCTTCAGTGG	AL137016	26757-26776	55	H	327	UPQFM-PCR
595A18R27039	aggcagaatcgactcaccgctaTTGGAAGGCGGATACAATTC		27039-27020				
595A18F54671	tccgtcttagctgagtgccgtaAGGAAGGCCCATCAACTTC	AL137016	54671-54689	55	H	199	UPQFM-PCR
595A18R54825	aggcagaatcgactcaccgctaGAGGCATCATAGGGGCAGTA		54825-54806				
595a18f101906	tccgtcttagctgagtgccgtaGATGCCCCAGATTGTACCAC	AL137016	101906-101925	55	H	279	UPQFM-PCR
595a18r102140	aggcagaatcgactcaccgctaTATTTGCCGAATTTCAACCA		102140-102121				
595a18f114741	tccgtcttagctgagtgccgtaACTCTATGCGGTGGATGACC	AL137016	114741-114760	55	H	239	UPQFM-PCR
595a18r114935	aggcagaatcgactcaccgctaGAAAGGTCTGAGCCAGTTGC		114935-114916				
291b3f26828	tccgtcttagctgagtgccgtaTCCTACAAACGCGATTAGCC	AL844175	26828-26847	55	H	261	UPQFM-PCR
291b3r27044	aggcagaatcgactcaccgctaACCCTGGTCCCAGAGTAGT		27044-27025				
291B3F64796	tccgtcttagctgagtgccgtaACTTGCCCCACTCTGTATGG	AL844175	64796-64815	55	H	260	UPQFM-PCR
291B3R65011	aggcagaatcgactcaccgctaTGGGCTCATATGCATGTTGT		65011-64092				
UNIVF	TCCGTCTTAGCTGAGTGGCGTA	-	-	55	H	-	UPQFM-PCR
UNIVR	AGGCAGAATCGACTCACCCTA	-	-	-	-	-	-
MAPH primers							
PZA	AGTAACGGCCGCCAGTGTGCTG	-	-		H	-	MAPH
PZB	CGAGCGGCCGCCAGTGTGATG	-	-				

Primer name	Primer sequence	Accession	Position within clone	Annealing temperature	Enzyme	Size of product (bp)	Other reaction conditions
PZAX	AGTAACGGCCGCCAGTGTGCTGGAATTCTGCAGAT	-	-	65	-	-	-
PZBX	CGAGCGGCCGCCAGTGTGATGGAT						
ch7q31.2F	CGAGGCTACAGCTTTGGAAC	AC000111	51188-51207	56	B	185	-
ch7q31.2R	CATCACACTTGTGCCATTCC		51372-51353				
XLNKXF	AAACAAGCAGCAGAAAAGGGA	AF283102	1245-1264	58	B	193	-
XLNKXR	AGGGAGAGCTGCCATACAGA		1437-1418				
plp5F	GGCCCCGTAACCTCCATAAAG	Z73964	17947-17966	58	B	199	-
plp5R	AATTGAAGGCCATGGGTGTA		18145-18126				
144a10F	TTCCAGGTATAACACCCCCA	Z70224	9207-9226	56	B	205	-
144a10R	GGGGTTGATAAGGGGAAAAA		9411-9392				
ch17p13.1F	AGCTTCAAGCAGCAACCACT	AC004706	4720-4739	56	H	209	-
ch17p13.1R	GGGCAGGGTAAAAGATGGAT		4928-4909				
240c2F	TAAAACCCCTCGCAGGAGAGA	Z73497	37375-37394	56	B	214	-
240c2R	CCCAGCACAGTTAGTCAGCA		37588-37569				
79p11F	GCATGTGGCCTCAGTATT	AL133348	12909-12928	58	B	223	-
79p11R	GGCAAAACCAGACAACACCT		13131-13112				
ch1q24.2F	ACCTGCTTATGGGGACCTCT	Z97200	10517-10536	58	B	230	-
ch1q24.2R	CAGCAGGGGCAGTAGCTAAG		10746-10727				
plp6F	AAAGATATCAACACATTTCAG	Z73964	18796-18815	56	B	239	--
plp6R	TTGCCTTTCAGAATAGCTGT		19034-19015				
ch6p24.2F	CTCACCATGGGAAGCATTTTT	AL022098	245-264	58	B	263	-
ch6p24.2R	TCACCCATCAAGACCTCACA		507-488				
sryF	GCTGGGATACCAGTGGAAAA	AC006040	63117-63098	58	B	271	-
sryR	TAAGTGGCCTAGCTGGTGCT		62847-62866				
ch4q26F	TGCGACAGTTCTGACCACTC	AC027613	4489-4508	56	H	276	-
ch4q26R	CCCTGGACTTGACCTGTGTT		4764-4745				
Xq12F	TGGAAACAGTCCCTTCCTTG	Z82212	6286-6267	58	B	281	-
Xq12R	CTTCCTGCTGCCTACACTCC		6006-6025				
198p4F	AGGGGATCTATGTTGGGACC	AL008708	22496-22515	56	B	293	-
198p4R	TGTTGGTCATGCTGTGGTTT		22788-22769				
43h13F	CACCAGTATCCCTTGCCACT	AL035444	22647-22666	58	B	300	-
43h13R	CTTGGTTCTTGAGCCTGGAG		22946-22927				

Primer name	Primer sequence	Accession	Position within clone	Annealing temperature	Enzyme	Size of product (bp)	Other reaction conditions
plp2F	TTCCCTGGTCTCGTTTGTCT	Z73964	15993-16012	58	B	309	-
plp2R	TGAGGATGATCACCTTGTCG		16301-16282				
plp7F	TCACCCACAGAAAGAGAGCA	Z73964	20741-20760	58	B	314	-
plp7R	GAGGGCCATCTCAGGTTACA		21054-21035				
ch17q21.32F	AAGGTCCTGGGCTAGAAGGA	AC103702	105825-105844	58	B	331	-
ch17q21.32R	ACCCACACAAGGAGGTTTCAG		106155-106136				

APPENDIX B

Genomic clones used or referred to within this thesis. The name of the clone, its accession number, position in Mb on the X chromosome (NCBI release 34), orientation (+, centromere to telomere; -, telomere to centromere) and the band to which it maps, are all shown.

Clone	Accession	Mb	Orientation	band
dJ570L12	AL049589	76.065-76.206	+	Xq21.1
dJ795G23	AL031000	77.976-78.108	-	Xq21.1
bA102P23	AL357115	79.176-79.341	+	Xq21.1
bA346E8	AL359885	79.340-79.417	+	Xq21.1
bB52K8	AL445213	79.417-79.608	+	Xq21.1
dJ2A2	AL645817	79.955-80.084	+	Xq21.1
dJ717L12	AL021706	81.141-81.237	+	Xq21.1
dJ326L13	Z82170	81.448-81.575	+	Xq21.1
dA43C13	AL009175	83.940-84.047	+	Xq21.2
dJ769N13	AL035427	100.649-100.810	+	Xq22.1
bA522L3	AL590407	100.810-100.890	+	Xq22.1
cU157D4	Z95624	100.888-100.926	-	Xq22.1
dJ1054G24	AL669904	100.925-100.931	+	Xq22.1
cU237H1	Z95624	100.931-100.967	-	Xq22.1
cU237H3	Z93943	100.968-100.983	-	Xq22.1
cU61F10	Z75895	100.983-101.028	-	Xq22.1
cU73E8	Z73361	101.025-101.066	-	Xq22.1
dJ198P4	AL008708	101.066-101.101	+	Xq22.1
cU221F2	Z57546	101.101-101.139	-	Xq22.1
cU101D3	Z85997	101.138-101.179	-	Xq22.1
dJ635G19	AL035494	101.179-101.248	+	Xq22.1
cU65A4	Z81014	101.248-101.265	-	Xq22.1
cU177E8	Z68694	101.265-101.306	-	Xq22.1
dJ79P11	AL133348	101.306-101.346	+	Xq22.1
cU105G4	Z92846	101.346-101.390	+	Xq22.1
bB349O20	AL606763	101.388-101.450	+	Xq22.2
dJ823F3	AL079333	101.448-101.455	+	Xq22.2
dJ421I20	AL117327	101.455-101.508	+	Xq22.2
cU250H12	Z69733	101.508-101.549	-	Xq22.2
dJ43H13	AL035444	101.549-101.576	+	Xq22.2
cU246D9	AL021308	101.576-101.598	+	Xq22.2
cV857G6	Z73965	101.598-101.640	-	Xq22.2

Clone	Accession	Mb	Orientation	band
cU25D11	Z68327	101.640-101.645	-	Xq22.2
dJ1055C14	AL049610	101.645-101.745	+	Xq22.2
cU35G3	Z93848	101.745-101.784	+	Xq22.2
dJ764D10	AL034409	101.785-101.797	+	Xq22.2
cV698D2	Z73964	101.797-101.839	+	Xq22.2
dJ540A13*	AL139228	101.839-101.871	+	Xq22.2
cV461C10	Z75896	101.871-101.909	+	Xq22.2
dJ540A13*	AL139229	101.908-101.925	+	Xq22.2
bA370B6	AL390022	101.925-101.934	+	Xq22.2
cU116E7	Z70273	101.934-101.960	+	Xq22.2
cV362H12	Z70227	101.960-102.002	+	Xq22.2
dJ839M11	AL034485	102.002-102.050	+	Xq22.2
cU240C2	Z73497	102.050-102.090	-	Xq22.2
cV467E10	Z74620	102.090-102.101	+	Xq22.2
cU46H11	Z82254	102.100-102.139	+	Xq22.2
dJ513M9	AL049631	102.206-102.324	+	Xq22.2
bA541I12	AL121868	102.324-102.442	+	Xq22.2
cU144A10	Z70224	102.793-102.828	+	Xq22.3
dJ75H8	AL158821	104.940-105.048	+	Xq22.3
dJ378P9	AC005002	115.201-115.321	+	Xq23
dJ404F18	AC004000	117.309-117.437	+	Xq24
dJ136O17	Z72001	126.611-126.696	-	Xq25
dJ297J13	Z77723	129.009-129.036	-	Xq26.1
dJ154J13	AL049734	129.263-129.390	+	Xq26.2
bA324A3	AL355366	129.390-129.416	+	Xq26.2
bA453F18*	AL513242	129.416-129.431	+	Xq26.2
dJ765F13	AL109853	129.431-129.538	+	Xq26.2
bA453F18*	AL513243	129.538-129.597	+	Xq26.2
bA512H23	AL121572	129.596-129.638	+	Xq26.2
dJ197O17	AL008629	129.638-129.799	-	Xq26.2
dA213H19	AL109749	129.799-129.938	+	Xq26.2
dJ305B16	AL049792	129.939-130.048	+	Xq26.2
dJ37M17	Z78022	130.048-130.149	+	Xq26.2
dJ965E19	AL353676	130.149-130.173	+	Xq26.2
dJ842K24	AL050310	130.173-130.282	+	Xq26.2
dJ119E23	Z99570	131.749-131.815	+	Xq26.2
dJ260J9	Z82193	136.447-136.545	+	Xq26.3
dJ656F14	AL034416	136.787-136.818	+	Xq27.1
bA35F15	AL590077	137.697-137.768	+	Xq27.1
bA364B14	AL589987	137.767-137.911	+	Xq27.1
bA189F12*	AL449183	137.912-137.927	+	Xq27.1
bA189F12*	AL449184	138.038-130.038	+	Xq27.1
bA51C14	AL121875	138.186-138.334	+	Xq27.1

Clone	Accession	Mb	Orientation	band
dJ595A18	AL137016	138.334-138.450	+	Xq27.1
dJ177G6	AL078639	138.516-138.645	+	Xq27.1
dJ507I15	Z98950	138.838-138.962	-	Xq27.1
dJ406C18	AL023773	139.856-139.972	+	Xq27.2
dJ73H14	AL080272	140.343-140.438	+	Xq27.2
dJ357K22	AL022720	140.847-140.993	+	Xq27.3
bG256O22	AL080239	141.291-141.508	+	Xq27.3
dJ145B12	AL008706	142.395-142.510	-	Xq27.3
bA550B3	AL589671	143.396-143.425	+	Xq27.3
bA159A24	AL356503	143.863-144.010	+	Xq27.3
bG278N14	AL109654	144.021-144.203	+	Xq27.3
bA183K14	AL109913	144.203-144.346	+	Xq27.3
bA79A21	AL513491	144.346-144.360	+	Xq27.3
dJ203P18	Z97180	144.360-144.493	-	Xq27.3

APPENDIX C

In each table the occurrence of various sequence motifs close to the different sequenced breakpoints from this study are shown. The actual sequences corresponding to each motif are given in Table 2.4. The number of times each motif was found in each 5Kb region is shown in the “Number found” column, and where this number is less than 10, the actual positions of each motif within the 5Kb region are shown, with (-) in front of a position indicating that the motif was on the reverse strand. The number of times each motif was expected to occur on both strands of the 5Kb sequence given the nucleotide frequencies and assuming random assortment of nucleotides is shown in the “Expected” column. The relative incidence of each sequence is shown in the final column, and is calculated as the difference between the observed and expected incidences, divided by the expected incidence, plus one. Where a motif was found at five times or greater than the expected incidence within the 5Kb region, the relative incidence is underlined. Motifs that were found within 50bp of a breakpoint have their positions highlighted in bold, where those positions within the 5Kb were already shown. In those cases where too many copies of the motif had been found in the 5Kb region to show all of them, the occurrences within the region around the breakpoint were shown bracketed in bold. Where a motif was found more than 10 times close to the breakpoint, just the number of occurrences was shown. For each motif occurrence table, the actual position of the 5Kb region within the relevant genomic clone is given, and the smaller region around the breakpoint is shown in brackets. For cU177E8 in family 2, there were two breakpoints within the 5Kb region, the region just around the short inserted sequence is shown in italics, and motif occurrences within this region are also italicised.

Family 1 distal duplication breakpoint:

dJ839M11 8310-13309 (10759-10858) (100bp)

Motif	Number found	Positions	Expected	Relative incidence
<i>S. pombe</i> ARS	1	3427	0.03	31.63
Scaffold motif 1	1	-3438	0.03	31.69
Scaffold motif 2	1	-4432	0.13	7.42
Scaffold motif 3	7	3453, 4592, -3438, -3460, -3485, -3499, -3709	1.41	4.98
Scaffold motif 4	4	-3270, -4432, -4483	1.49	2.01
Topoisomerase I 1	199	(2456, 2463, 2476, 2544, -2466, -2519, -2528)	163.26	1.22
Topoisomerase I 2	363	(2460, 2485, -2474, -2495, -2502, -2506, -2541)	312.34	1.16
Topoisomerase I 3	243	(2450, 2468, 2488, -2514, -2523, -2535)	312.19	0.78
Topoisomerase I 4	321	(2516, 2525, -2459, -2479, -2547)	344.13	0.93
Vaccinia topoisomerase I	32		19.58	1.63
Heptamer	5	1902, -2146, -2340, -3143, -3815	0.57	8.73
Ig class switch 1	11		9.19	1.20
Ig class switch 2	18	(-2498)	8.32	2.16
Ig class switch 3	8	214, 605, 697, 4881, -1263, -3190, -4090, -4689	8.31	0.96
Ig class switch 4	21		8.31	2.53
Ig class switch 5	12		9.19	1.31
Translin consensus 3	5	113, 597, 1268, -339, -4031	2.41	2.07
Translin consensus 4	5	2045, -752, -2285, -2300, -2364	1.97	2.54
Chi-like sequence	3	1684, 2276, -4029	0.49	6.10
Deletion hotspot consensus	69	(-2464)	38.82	1.78
Mouse deletion hotspot	145	(2481, 2498, 2514)	74.07	1.96
Parvovirus deletion hotspot	17		11.50	1.48
Mouse MHC recombination hotspot	3	1571, 2398, 3286	0.55	5.42
Polymerase α pause site 1	190	(2493, 2500, 2504, 2541, -2462)	147.91	1.28
Polymerase α pause site 2	42	(2534)	147.86	0.28
Polymerase α pause site 3	186	(2474, 2495, 2519, 2530, -2485, -2510, -2531)	266.91	0.70
Polymerase arrest site	51	(-2464)	18.44	2.77
Polymerase α frameshift site 1	3	935, 1918, -1845	1.97	1.52
Polymerase β frameshift site 1	14		9.23	1.52
Polymerase β frameshift site 2	43		48.68	0.88
Polymerases α/β frameshift site 1	10	213, 604, 2339, 2774, 3315, 4880, -1412, -3002, -3561, -4091	9.24	1.08
Polymerases α/β frameshift site 2	4	4086, -218, -609, -4885	2.19	1.83
DNA bending motif	2	3829, 4944	3.26	0.61

Family 1 proximal duplication breakpoint:

dJ635G19: 66710-69648 (69160-69259), cU65A4: 16814-14654 (100bp)

Motif	Number found	Positions	Expected	Relative incidence
<i>S. cerevisiae</i> ARS	1	-1140	0.08	12.01
Scaffold motif 2	4	-1302, -1312, -3129, -4975	0.51	7.82
Scaffold motif 3	2	1130, 1549	2.52	0.79
Scaffold motif 4	6	1472, -1138, -1302, -1312, -3129, -4975	3.13	1.92
Topoisomerase I 1	164	(2481, -2534)	158.21	1.04
Topoisomerase I 2	350	(2468, 2537, -2454, -2494, -2509)	320.08	1.09
Topoisomerase I 3	271	(2485, 2488, 2496, -2477)	321.85	0.84
Topoisomerase I 4	315	(2531, -2453)	350.96	0.90
Vaccinia topoisomerase I	24	(2466)	21.72	1.11
Ig class switch 1	13		8.04	1.62
Ig class switch 2	8	351, 575, 2975, 3971, -57, -1488, -2051, -2158	7.11	1.12
Ig class switch 3	9	131, 1591, 1807, 1930, -1010, -1120, -1328, -2345, -2462	7.14	1.26
Ig class switch 4	20	(-2475, -2516)	7.14	2.80
Ig class switch 5	7	51, 1235, 2117, 3022, -1030, -1918, -2616	8.04	0.87
Translin consensus 3	4	1611, 2177, -4238, -4636	2.20	1.82
Translin consensus 4	2	2091, -445	1.58	1.26
Chi-like sequence	1	1863	0.40	2.53
Deletion hotspot consensus	46	(2490, -2475, -2516)	35.78	1.29
Mouse deletion hotspot	127	(2501, 2505, 2540)	72.34	1.76
Parvovirus deletion hotspot	22		16.85	1.31
Mouse MHC recombination hotspot	3	1064, 1698, 2501	0.60	4.98
Polymerase α pause site 1	169	(2452, 2492, 2507, -2539)	139.00	1.22
Polymerase α pause site 2	27	(-2521, -2527)	139.50	0.19
Polymerase α pause site 3	180	(2463, 2511, 2524, 2535, 2545, -2464, -2501, -2505, -2536)	229.02	0.79
Polymerase arrest site	40	(2490, -2541)	17.18	2.33
Polymerase α frameshift site 1	4	2011, 2286, -950, -1735	1.58	2.51
Polymerase α frameshift site 2	3	2142, 2445, -2407	1.60	1.87
Polymerase β frameshift site 1	9	1395, 2764, 2789, 3362, 3949, 4362, -21, -4647	9.03	1.00
Polymerase β frameshift site 2	79		80.72	0.98
Polymerases α/β frameshift site 1	8	66, 2446, -1011, -1121, -1329, -1771, -1943, -2634	9.33	0.86
Polymerases α/β frameshift site 2	3	1006, 1116, 1324	2.11	1.42
DNA bending motif	1	839	3.46	0.29

Family 2 cU177E8: 39838-34839 (37388-37289) 100bp (36104-35956) (149bp)

Motif	Number found	Positions	Expected	Relative incidence
<i>S. pombe</i> ARS	2	4878, 4890	0.11	<u>17.81</u>
Scaffold motif 1	3	-4885, -4897, -4909	0.10	<u>30.94</u>
Scaffold motif 2	5	672, 995, 1681, -2394, -3864	0.48	<u>10.36</u>
Scaffold motif 3	14	(-2487)	4.06	3.44
Scaffold motif 4	11	(2487) (-3864)	4.90	2.25
Topoisomerase I 1	187	(2457, 2466, -2465) (3776, 3804, -3738)	179.09	1.04
Topoisomerase I 2	364	(14 occurrences) (14 occurrences)	310.09	1.17
Topoisomerase I 3	218	(-2511) (3750, 3755, 3790, 3829, -3769, -3820, -3836, -3869)	294.41	0.74
Topoisomerase I 4	367	(2462, 2476, 2488, 2498, -2460, -2469, -2479, -2483, -2491, -2501) (3735, 3858, 3873, -3779, -3807)	445.06	0.82
Vaccinia topoisomerase I	38	(-3845)	20.12	1.89
Heptamer	1	4302	0.41	2.43
Ig class switch 1	7	1377, -1394, -1762, -2714, -3328, -3746, -4451	7.05	0.99
Ig class switch 2	5	652, 4986, -265, -2383, -3604	4.80	1.04
Ig class switch 3	11		5.54	1.98
Ig class switch 4	15		5.54	2.71
Ig class switch 5	5	2335, 3196, -1395, -1752, -3762	7.04	0.71
Translin consensus 3	5	2233, 3507, -296, -4357, -4754	2.12	2.35
Translin consensus 4	1	1104	0.95	1.06
Chi-like sequence	1	-612	0.27	3.71
Deletion hotspot consensus	51	(3871, -3748, -3777)	38.41	1.31
Mouse deletion hotspot	101	(-2508) (3765, 3816, 3836, -3772, 3782, -3812, -3829, -3872)	59.10	1.71
Parvovirus deletion hotspot	29	(2452, 2513) (-3843, -3867)	17.17	1.69
Mouse MHC recombination hotspot	1	3523	0.36	2.80
Polymerase α pause site 1	179	(-2473, -2503, -2523, -2526, -2529, -2535, -2538, -2541, -2547, -2550) (3846, 3878, -3746, -3761, -3775, -3798, -3804, -3810, -3833)	126.60	1.41
Polymerase α pause site 2	19	(3762)	117.54	0.16
Polymerase α pause site 3	114	(3814, 3825, -3765, -3815)	158.30	0.72
Polymerase arrest site	41	(-2525, -2537, -2549) (-3748, -3777)	16.18	2.53
Polymerase α frameshift site 1	3	3471, 3797, -2886	1.32	2.28
Polymerase β frameshift site 1	9	46, 1717, 4066, -566, -876, -1058, -1225, -2645, -3532	7.36	1.22
Polymerase β frameshift site 2	93	(2453, 2490, 2514, -2498) (-3842, -3858, -3864)	81.15	1.15
Polymerases α/β frameshift site 1	11		7.51	1.47
Polymerases α/β frameshift site 2	4	4196, 4253, -1220, -1366	1.61	2.49
DNA bending motif	7	75, 186, 745, 1559, 2135, 2491, 4363	7.05	0.99

Family 2 dJ1055C15: 57309-62308 (59759-59892) (100bp)

Motif	Number found	Positions	Expected	Relative incidence
Chi	1	-496	0.05	<u>19.10</u>
<i>S. cerevisiae</i> ARS	1	-2661	0.15	<u>6.71</u>
<i>S. pombe</i> ARS	1	-4815	0.13	<u>7.52</u>
Scaffold motif 1	1	4804	0.13	<u>7.62</u>
Scaffold motif 2	8	1803, -360, -2725, -3966, -3976, -4138, -4477, -4487	0.69	<u>11.60</u>
Scaffold motif 3	10	118, 389, 1663, 2723, 3751, 3950, 4804, -548, -1813, -3760	4.81	2.08
Scaffold motif 4	11		6.02	1.83
Topoisomerase I 1	180	(2457, 2582, -2453, -2460, -2524, -2537, -2569)	178.28	1.01
Topoisomerase I 2	337	(2526, 2576, 2579, -2485, -2488, -2499, -2508, -2512, -2515)	304.61	1.11
Topoisomerase I 3	239	(2469, 2479, 2515, 2547, -2462, -2526)	298.69	0.80
Topoisomerase I 4	394	(13 occurrences)	452.95	0.87
Vaccinia topoisomerase I	23		19.05	1.21
Nonamer	1	239	0.07	<u>15.04</u>
Ig class switch 1	7	599, 751, 1472, 4070, -2672, -2736, -3458	6.89	1.02
Ig class switch 2	6	327, 773, 1344, -661, -876, -4368	4.50	1.33
Ig class switch 3	4	1314, 1434, -1459, -2703	4.50	0.89
Ig class switch 4	4	1018, 1313, -1460, -2704	4.50	0.89
Ig class switch 5	10	277, 750, 994, 2973, 4069, 4736, 4941, -770, -2673, -4529	6.89	1.45
Translin consensus 3	4	830, -568, -1550, -2323	2.10	1.90
Translin consensus 4	3	4365, -2809, -3986	0.85	3.51
Chi-like sequence	1	-496	0.22	4.63
Deletion hotspot consensus	39		36.78	1.06
Mouse deletion hotspot	128	(2530)	59.23	2.16
Parvovirus deletion hotspot	28	(-2499)	18.27	1.53
Mouse MHC recombination hotspot	3	3387, 3883, -1212	0.37	<u>8.19</u>
Polymerase α pause site 1	146	(2506, 2510, 2513, -2528, -2578, -2581)	119.01	1.23
Polymerase α pause site 2	26		117.17	0.22
Polymerase α pause site 3	123		152.11	0.81
Polymerase arrest site	30	(2511, -2580)	14.34	2.09
Polymerase β frameshift site 1	13		6.92	1.88
Polymerase β frameshift site 2	100	(-2494, -2498)	91.92	1.09
Polymerases α/β frameshift site 1	9	1313, 1975, 3072, 3609, -1460, -2466, -2704, -4123, -4765	6.92	1.30
Polymerases α/β frameshift site 2	3	1455, 2699, -1318	1.36	2.21
DNA bending motif	4	623, 678, 3313, 4214	7.43	0.54

Family 3 bA346E8 deletion breakpoint 2487-7486 (4937-5036) (100bp)

Motif	Number found	Positions	Expected	Relative incidence
<i>S. cerevisiae</i> ARS	1	948	0.44	2.29
Scaffold motif 2	4	154, 771, 950, 1697	3.26	1.23
Scaffold motif 3	6	-284, -958, -1701, -3166, -3341, -4502	9.34	0.64
Scaffold motif 4	8	154, 771, 944, 1693, 1989, 2083, 4321, -595	14.70	0.54
Topoisomerase I 1	176	(2457, 2469, 2487, 2498, -2460)	161.95	1.09
Topoisomerase I 2	324	((2474, 2502, 2510, 2519, -2525)	298.18	1.09
Topoisomerase I 3	277	(2452, 2481, 2507, 2517, 2527, 2545)	320.80	0.86
Topoisomerase I 4	436	(2461, -2464, -2468, -2472, -2490, -2537)	441.76	0.99
Vaccinia topoisomerase I	35		18.99	1.84
Nonamer	1	-4786	0.31	3.20
Heptamer	1	-14	0.23	4.37
Ig class switch 1	8	668, -1219, -2303, -2718, -2777, -3458, -4397, -4531	5.41	1.48
Ig class switch 2	5	1008, 2226, 2350, -3627, -4306	3.75	1.33
Ig class switch 3	7	548, 630, 1413, 3230, 4018, 4569, 4895	4.05	1.73
Ig class switch 4	7	629, 3229, 4017, 4568, -989, -3248, 3966	4.05	1.73
Ig class switch 5	9	313, 1061, 3513, -814, -853, -879, -1014, -2232, -4990	5.41	1.66
Translin consensus 3	7	1208, 3696, -1437, -1775, -1956, -3323, -4157	1.67	4.19
Translin consensus 4	1	-1011	0.61	1.64
Deletion hotspot consensus	38	(-2470)	30.77	1.23
Mouse deletion hotspot	101	(-2523, -2541)	56.43	1.79
Parvovirus deletion hotspot	25		27.05	0.92
Mouse MHC recombination hotspot	1	-3218	0.40	2.50
Polymerase α pause site 1	114	(2523, -2521)	101.64	1.12
Polymerase α pause site 2	12		105.85	0.11
Polymerase α pause site 3	69		122.53	0.56
Polymerase arrest site	23		11.62	1.98
Polymerase α frameshift site 1	2	983, 3242	0.61	3.29
Polymerase β frameshift site 1	6	-755, -1629, -2414, -3181, -3987, -4727	7.41	0.81
Polymerase β frameshift site 2	106	(2482, 2489, 2528)	173.92	0.61
Polymerases α/β frameshift site 1	12		7.95	1.51
Polymerases α/β frameshift site 2	4	-634, -3234, -4022, -4573	1.54	2.61
DNA bending motif	4	146, 2918, 3420, 4923	6.90	0.58

Family 3 dJ203P18 deletion breakpoint 101270-96271 (98818-98719) (100bp)

Motif	Number found	Positions	Expected	Relative incidence
<i>S. cerevisiae</i> ARS	1	-4056	0.12	8.17
Scaffold motif 2	2	1885, -4054	0.49	4.08
Scaffold motif 3	29		4.31	6.73
Scaffold motif 4	7	1624, 1885, 2182, 3309, -2838, -3008, -4054	5.20	1.35
Topoisomerase I 1	212	(2490, 2502, 2529, -2476, -2485, -2497, -2514)	180.11	1.18
Topoisomerase I 2	323	(-2470, -2509, -2525, -2545)	299.13	1.08
Topoisomerase I 3	245	(2486, 2514, -2463, -2501, -2529)	299.50	0.82
Topoisomerase I 4	384	(2464, 2473, 2476, 2511, 2537, -2481, -2493)	452.44	0.85
Vaccinia topoisomerase I	21	(-2527)	17.91	1.17
Topoisomerase II consensus	1	-1399	0.37	2.70
Heptamer	1	673	0.43	2.35
Ig class switch 1	9	403, 918, 991, 4151, 4339, -109, -142, -515, -4612	7.12	1.26
Ig class switch 2	6	889, 1814, -170, -315, -369, -2950	4.73	1.27
Ig class switch 3	6	304, 912, 1412, 2071, -438, -1110	4.78	1.26
Ig class switch 4	14		4.78	2.93
Ig class switch 5	10	2163, 2251, 3067, 3218, 4125, 4181, -143, -1956, -2626, -2866	7.12	1.40
Translin consensus 3	4	211, 899, 970, 2319	2.14	1.87
Translin consensus 4	2	51, 954	0.90	2.22
Deletion hotspot consensus	67	(2535)	37.41	1.79
Mouse deletion hotspot	102	(2454, 2459)	59.47	1.72
Parvovirus deletion hotspot	10	588, 3308, 3367, 4513, -849, -1282, -2284, -2525, -2553, -3799	16.29	0.61
Mouse MHC recombination hotspot	2	-2129, -2582	0.35	5.65
Polymerase α pause site 1	159		119.16	1.33
Polymerase α pause site 2	20		118.98	0.17
Polymerase α pause site 3	138	(2457, -2458)	157.37	0.88
Polymerase arrest site	27		14.30	1.89
Polymerase α frameshift site 1	1	-3189	0.95	1.05
Polymerase α frameshift site 2	1	4023	0.94	1.07
Polymerase β frameshift site 1	17		7.18	2.37
Polymerase β frameshift site 2	71	(-2474, -2523, -2550)	82.17	0.86
Polymerases α/β frameshift site 1	11		7.20	1.53
Polymerases α/β frameshift site 2	2	434, -2075	1.44	1.39
DNA bending motif	4	1266, 1630, 3262, 3909	7.41	0.54

Family 3 dJ1055C14 65684-70683 (68134-68233) (100bp)

Motif	Number found	Positions	Expected	Relative incidence
Ade-M26 heptamer	1	3184	0.65	1.55
Human replication origin consensus	1	1656	0.003	<u>357.36</u>
<i>S. cerevisiae</i> ARS	1	4953	0.09	<u>10.75</u>
<i>S. pombe</i> ARS	1	-658	0.09	<u>10.84</u>
Scaffold motif 2	7	510, 783, 3204, 3316, 3335, -4478, -4577	0.38	<u>18.40</u>
Scaffold motif 3	8	809, 1713, 1967, 3844, 4468, -791, -1703, -2073	3.43	2.33
Scaffold motif 4	16		4.03	3.97
Topoisomerase I 1	132	(-2506)	177.41	0.74
Topoisomerase I 2	306	(2477, -2497, -2500, -2514)	307.09	1.00
Topoisomerase I 3	221	(2500)	300.96	0.73
Topoisomerase I 4	420	(17 occurrences)	428.12	0.98
Vaccinia topoisomerase I	34	(-2516)	19.21	1.77
Topoisomerase II consensus	1	4106	0.41	2.46
Ig class switch 1	7	1310, 3443, -633, -1268, -1868, -2323, -2772	7.54	0.93
Ig class switch 2	3	-1219, -3219, -3986	5.36	0.56
Ig class switch 3	11		5.66	1.94
Ig class switch 4	14		5.66	2.47
Ig class switch 5	5	3442, 3653, -22, -1869, -4543	7.534	0.66
Translin consensus 3	1	-1118	2.21	0.45
Translin consensus 4	7	212, 3220, 3659, 3983, -208, -2743, -2991	1.09	<u>6.45</u>
Deletion hotspot consensus	47		38.23	1.23
Mouse deletion hotspot	101	(2508)	62.33	1.62
Parvovirus deletion hotspot	33	(2468)	15.65	2.11
Polymerase α pause site 1	136		127.93	1.06
Polymerase α pause site 2	69		124.39	0.55
Polymerase α pause site 3	172	(2528, -2508, -2517)	176.05	0.98
Polymerase arrest site	26		15.90	1.64
Polymerase α frameshift site 1	5	1840, 2021, 2356, 3383, 4398	1.26	3.98
Polymerase α frameshift site 2	1	3704	1.09	0.92
Polymerase β frameshift site 1	11		7.65	1.44
Polymerase β frameshift site 2	101	(2471, -2457)	74.06	1.36
Polymerases α/β frameshift site 1	8	2262, 3640, 3705, 4881, -271, -1599, -1831, -4034	7.71	1.04
Polymerases α/β frameshift site 2	2	1826, -3645	1.64	1.22
DNA bending motif	4	425, 1071, 1906, 3926	6.28	0.64

Family 3 dJ305B16 51553-46552 (43978-44029) (152bp)

Motif	Number found	Positions	Expected	Relative incidence
Pur binding motif	1	-1344	0.0002	<u>5800.51</u>
<i>S. cerevisiae</i> ARS	1	2248	0.20	4.96
Scaffold motif 1	1	3577	0.16	<u>6.14</u>
Scaffold motif 2	5	1696, 1706, 2390, -4052, -4217	0.84	5.94
Scaffold motif 3	20		6.47	3.09
Scaffold motif 4	16		8.24	1.94
Topoisomerase I 1	209	(2435, 2443, 2464, 2513, 2558, -2434, -2540)	182.98	1.14
Topoisomerase I 2	315	(13 occurrences)	293.24	1.07
Topoisomerase I 3	245	(2439, -2525, -2544, -2566)	290.71	0.84
Topoisomerase I 4	388	(11 occurrences)	495.88	0.78
Vaccinia topoisomerase I	31	(2466)	17.36	1.79
Heptamer	3	52, 4020, -3664	0.36	<u>8.30</u>
Ig class switch 1	7	2929, 3382, 3836, 4151, -2784, -2934, -3387	6.22	1.13
Ig class switch 2	4	-229, -877, -2585, -3994	3.64	1.10
Ig class switch 3	10	648, 704, 3753, -1010, -1184, -1311, -3885, -4033, -4176, -4250	3.65	2.74
Ig class switch 4	13		3.65	3.56
Ig class switch 5	9	74, 2460, 3381, 3520, 4125, 4150, -2935, -3388, -3941	6.22	1.45
Translin consensus 3	2	3929, -1270	1.97	1.02
Translin consensus 4	1	3991	0.63	1.59
Chi-like sequence	1	-41	0.16	6.32
Deletion hotspot consensus	54	(2516)	36.29	1.49
Mouse deletion hotspot	69		53.79	1.28
Parvovirus deletion hotspot	31	(2468, 2544, -2555)	18.82	1.65
Mouse MHC recombination hotspot	1	4964	0.29	3.41
Polymerase α pause site 1	125	(2457, 2461, 2517, -2479, -2500)	108.06	1.16
Polymerase α pause site 2	11		107.24	0.10
Polymerase α pause site 3	102	(-2454, -2498, -2531, -2576)	125.42	0.81
Polymerase arrest site	26	(-2502)	12.65	2.06
Polymerase α frameshift site 1	3	832, 1339, 1814	0.69	4.33
Polymerase β frameshift site 1	12		6.22	1.93
Polymerase β frameshift site 2	104	(2469, 2484, 2532, 2545, -2524m - 2553)	101.26	1.03
Polymerases α/β frameshift site 1	11		6.21	1.77
Polymerases α/β frameshift site 2	7	1180, 4029, 4172, 4246, -652, -708, -3757	1.14	<u>6.11</u>
DNA bending motif	8	274, 1091, 2163, 2546, 2559, 3202, 3359, 3570	9.75	0.82

Family 3 dJ197O17 71905-66906 (69483-69329) (155bp)

Motif	Number found	Positions	Expected	Relative incidence
<i>S. cerevisiae</i> ARS	1	-1542	0.32	3.17
Scaffold motif 1	1	771	0.27	3.72
Scaffold motif 2	1	-1024	1.51	0.66
Scaffold motif 3	8	163, 771, 1011, 1539, 3150, 4604, -1750, -3159	8.64	0.93
Scaffold motif 4	18		11.82	1.52
Topoisomerase I 1	200	(2455, 2499, 2519, -2502)	180.57	1.11
Topoisomerase I 2	293	(13 occurrences)	283.26	1.03
Topoisomerase I 3	237	(2430, 2469, 2476, 2556, 2572, -2453, -2496, -2553, -2562)	292.60	0.81
Topoisomerase I 4	494	(2443, 2462, 2487, 2545, -2446, -2458, -2465, -2490, -2550)	512.89	0.96
Vaccinia topoisomerase I	25		16.14	1.55
Ig class switch 1	8	1663, 1874, 2435, 3326, 3583, 4501, -3911, -4405	5.56	1.44
Ig class switch 2	6	2946, 3992, -3704, -3827, -3922, -3974	3.18	1.89
Ig class switch 3	6	2671, 4482, -3801, -3865, -3896, -4446	3.35	1.79
Ig class switch 4	12		3.35	3.58
Ig class switch 5	7	2080, 3325, 3407, 4046, -2571, -2831, -4406	5.56	1.26
Translin consensus 3	6	492, 3715, 4149, 4280, -2078, -2207	1.78	3.37
Translin consensus 4	2	-3439, -4157	0.50	4.01
Deletion hotspot consensus	52	(-2451, -2529, -2565)	34.53	1.51
Mouse deletion hotspot	80	(2558, 2564, -2467)	50.39	1.59
Parvovirus deletion hotspot	32		20.50	1.56
Polymerase α pause site 1	118	(2435, -2427, -2450, -2482, -2485, -2528, -2537, -2564, -2570)	98.66	1.20
Polymerase α pause site 2	11		99.77	0.11
Polymerase α pause site 3	74	(2511, -2568)	108.42	0.68
Polymerase arrest site	16	(-2484)	11.25	1.42
Polymerase α frameshift site 1	1	2362	0.55	1.83
Polymerase β frameshift site 1	13	(-2470)	6.16	2.11
Polymerase β frameshift site 2	133	(2439, 2538)	124.68	1.07
Polymerases α/β frameshift site 1	11		6.33	1.74
Polymerases α/β frameshift site 2	6	3797, 3861, 3892, 4442, -2675, -4486	1.15	<u>5.20</u>
DNA bending motif	9	90, 158, 414, 535, 1215, 1674, 2113, 3449, 4657	10.80	0.83

Family 3 bA453F18 22455-27454 (24905-25004) (100bp)

Motif	Number found	Positions	Expected	Relative incidence
<i>S. pombe</i> ARS	3	1423, 1435, 1447	0.10	28.81
Scaffold motif 1	4	-1430, -1442, -1454, -1466	0.09	45.33
Scaffold motif 2	4	-2818, -2828, -3560, -3960	0.42	9.57
Scaffold motif 3	11		3.78	2.91
Scaffold motif 4	11		4.48	2.45
Topoisomerase I 1	145	(2459, 2465, -2456, -2519)	178.76	0.81
Topoisomerase I 2	348	(2463, 2478, 2484, 2542, -2475, -2494, -2503, -2506, -2535)	301.69	1.15
Topoisomerase I 3	243	(2469, 2475, 2482, -2459, -2529)	301.89	0.80
Topoisomerase I 4	415	(2453, 2487, 2498, 2516, -2462, -2490, -2510, -2523)	438.57	0.99
Vaccinia topoisomerase I	29		18.22	1.59
Ig class switch 1	2	4452, -445	7.41	0.27
Ig class switch 2	9	194, 1751, 1757, 4989, -88, -1134, -2563, -3715, -3826	5.12	1.76
Ig class switch 3	8	1164, 1293, 1634, 2986, -3778, -4813, -4887, -4906	5.18	1.55
Ig class switch 4	11		5.18	2.12
Ig class switch 5	12	(-2545)	7.41	1.62
Translin consensus 3	1	3450	2.19	0.46
Translin consensus 4	4	85, 1131, -780, -3418	1.01	3.98
Chi-like sequence	1	-3045	0.25	3.94
Deletion hotspot consensus	48	(2531)	37.76	1.27
Mouse deletion hotspot	120	(-2532)	61.40	1.95
Parvovirus deletion hotspot	20		15.67	1.28
Polymerase α pause site 1	161	(2533, -2465, -2544)	123.08	1.31
Polymerase α pause site 2	34	(2545)	122.81	0.28
Polymerase α pause site 3	151	(2547, -2542, -2548)	169.00	0.89
Polymerase arrest site	33	(2531)	14.90	2.21
Polymerase β frameshift site 1	12		7.47	1.61
Polymerase β frameshift site 2	112	(-2499)	77.08	1.45
Polymerases α/β frameshift site 1	11		7.50	1.47
Polymerases α/β frameshift site 2	3	-1168, -1297, -1638	1.54	1.95
DNA bending motif	9	299, 914, 1721, 2419, 2853, 3533, 4439, 4685, 4913	6.75	1.33