

TECHNOLOGICAL AND BIOLOGICAL STUDIES OF HUMAN STRUCTURAL VARIATION

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Katherine Emily Reekie BSc (Sheffield)

Department of Genetics

University of Leicester

September 2010

© Katherine E Reekie 2010

This thesis is copyright material and no quotation from it may be published without proper
acknowledgement

Technological and Biological Studies of Human Structural Variation

Katherine Emily Reekie

Abstract

Extensive regions of copy number variation (CNV) located throughout the genome play a significant role in common disease. This thesis describes the study of a structural variation on chromosome 12p13.31, and its involvement in susceptibility to complex disease, specifically the autoimmune disorder rheumatoid arthritis (RA).

Methods of studying CNV include oligo-array Comparative Genomic Hybridisation (oaCGH), for which we developed a relatively optimised protocol on an in-house customisable microarray platform. In parallel, studies of the 12p13.31 locus using PCR-based methods revealed a large novel tandem duplication. Using quantitative assays we detected copy number variation within this duplication which occurs at a frequency of ~4% in European populations. At least two distinct points of recombination have been identified, supporting our theory that CNV in this region initiated from NAHR between the two units of the tandem duplication.

We assessed copy number of sequences within the tandem duplication in a Swedish RA cohort, and revealed that a low copy number within this region occurs at a significantly higher rate in control samples compared to cases ($p=0.001$, OR=2.3 (95% CI 1.4-3.9)), suggesting a protective role for this variant. This was replicated in a UK cohort ($p=0.036$, OR=1.90 (95% CI 0.93-3.82)). We believe that the size of this effect is as large as any previously reported impact of CNV on common disease. An association was also detected in a Swedish psoriasis cohort ($p=0.013$, OR=2.16 (95% CI 1.2-4.1)).

Future investigations into the effect of CNV at 12p13.31 on gene and protein expression may provide an insight into mechanisms of RA susceptibility and development. Given the tendency for autoimmune disease loci to share susceptibility regions, as well as the biological importance of genes located with the tandem duplication, we consider it likely that this region may also play a role in other complex disorders.

Acknowledgements

The last four years would not have been possible without the help and support of a great number of people. Firstly, thanks must go to my supervisor, Anthony Brookes, for taking me on and for his supervision and guidance. His never-ceasing enthusiasm and optimism were a constant source of inspiration. I would also like to thank the MRC for funding my PhD.

A big thank you must go to Colin Veal for his advice, patience with my constant questions and for reading the thesis multiple times! Also, thanks to all other members of the Brookeslab past and present, I have been very lucky to work with such a brilliant group of people. Special mention must go to Reshma, without whom the lab would not function, Pete for advice and company in the lab, and Owen for help with most of the bioinformatics-related aspects of this thesis. Thanks also to my 3rd year project students Kelly and Jaya for tying up a lot of the loose ends (good times!).

I also acknowledge everyone from the University of Leicester and collaborators elsewhere, who have provided me with sets of DNA samples for this work, and Bettina, Ioannis and others at Febit for their help with the Geniom experiments.

My PhD experience would not have been the same without all the colleagues and friends I met along the way; there are far too many to name, but you know who you are! Special thanks must go to Christine, RobM (and his amazing chocolate desserts) and the ‘Monkees’, especially Adam, RobH, Sirisha (for good advice and always being able to make me laugh) and RobF (a good distraction).

Finally I must thank my family, especially my parents...not only for the genes (I am rather attached to them!) but also for their constant encouragement and belief in me. You support me in everything I do and I couldn’t have done this without you!

Table of Contents

Abstract.....	i
Acknowledgements.....	ii
List of Abbreviations.....	vii
List of Figures.....	x
List of Tables.....	xii

Chapter 1: Introduction

1.1 Genetic Variation	1
1.1.1 Genetic Variation and Disease	4
1.1.1.1 <i>Linkage Studies and Single Gene Disorders.....</i>	<i>4</i>
1.1.1.2 <i>Association Studies and Complex Disorders.....</i>	<i>5</i>
1.1.1.3 <i>Determining the Statistical Significance of Results.....</i>	<i>6</i>
1.1.1.4 <i>Adjusting for Multiple Testing.....</i>	<i>7</i>
1.1.1.5 <i>Calculating the Strength of an Association</i>	<i>8</i>
1.1.1.6 <i>Genome-Wide Association Studies.....</i>	<i>10</i>
1.2 Copy Number Variation.....	12
1.2.1 Early Genome-Wide Studies of CNV	12
1.2.2 CNV Genomics.....	15
1.2.3 Differences in Copy Number Variation between Populations	16
1.2.4 Categorising and naming CNVs	16
1.2.5 Mechanisms of CNV Formation	17
1.2.5.1 <i>Recurrent Rearrangements.....</i>	<i>18</i>
1.2.5.2 <i>Non-Recurrent Rearrangements</i>	<i>19</i>
1.2.5.3 <i>Somatic Recombination</i>	<i>20</i>
1.3 Phenotypic Effects of Copy Number Variation.....	22
1.3.1 The Role of CNV in Adaptation and Evolution	22
1.3.2 CNV and Disease	24
1.3.3 'Missing Heritability'	26
1.3.4 Genome-wide CNV Studies	27
1.4 Methods of Studying Copy Number Variation.....	29
1.4.1 Methods Employed for Early Studies of CNV.....	29
1.4.2 PCR-based Methods.....	31
1.4.3 Comparative Genomic Hybridisation	34
1.4.4 Sequence-based Approaches.....	36
1.4.5 SNP-based Approaches	38
1.5 Rheumatoid Arthritis.....	40
1.5.1 Environmental Factors Contributing to RA Susceptibility.....	41
1.5.2 Auto-antibodies in RA	42
1.5.3 RA Pathogenesis.....	43
1.5.3.1 <i>Chondrocytes.....</i>	<i>44</i>

1.6	Rheumatoid Arthritis Genetics.....	47
1.6.1	Identification of HLA Genes as the Major RA Susceptibility Locus	47
1.6.2	RA Linkage Studies	48
1.6.3	<i>PTPN22</i> : The Second RA Susceptibility Locus.....	50
1.6.4	SNPs Associated with RA Identified Using GWAS	51
1.6.5	CNV and autoimmune disease	52
1.6.6	Involvement of Pathways.....	53
1.6.7	Interaction Between Genetic and Environmental Factors.....	53
1.6.8	Population Differences.....	54
1.6.9	The Current State of RA Genetics	55
1.7	A Putative RA Susceptibility Locus on Chromosome 12.....	57
1.7.1	Identification of an RA Susceptibility locus on Rat chromosome 4q42	57
1.7.2	Human 12p13.31.....	58
1.8	Summary and Project Aims	62

Chapter 2: Materials and Methods

2.1	Materials and Equipment.....	63
2.1.1	Chemicals, Enzymes and Oligonucleotides	63
2.1.2	DNA Samples	63
2.1.3	Equipment and Computer Software	65
2.2	Oligo-arrayCGH on the Geniom Platform	67
2.2.1	Array Design	67
2.2.2	Array Synthesis.....	69
2.2.3	Preparation of DNA	69
2.2.3.1	<i>Whole Genome Amplification</i>	69
2.2.3.2	<i>Restriction Digests</i>	70
2.2.3.3	<i>Purification of Digest Products</i>	70
2.2.3.4	<i>Precipitation of DNA</i>	71
2.2.3.5	<i>Biotin Labelling</i>	71
2.2.4	Pre-Hybridisation	72
2.2.5	Hybridisation	72
2.2.6	Wash Steps.....	73
2.2.6.1	<i>Machine Washes</i>	73
2.2.6.2	<i>Manual Washes</i>	73
2.2.7	Detection.....	74
2.2.8	Signal Amplification.....	74
2.2.9	Analysis.....	74
2.3	Polymerase Chain Reaction (PCR).....	76
2.3.1	Primer design	76
2.3.2	Reaction Conditions	76
2.3.2.1	<i>Standard PCR Conditions</i>	76
2.3.2.2	<i>Betaine Supplemented PCR Conditions</i>	76
2.3.2.3	<i>PCR using 11.1 x buffer</i>	77
2.3.2.4	<i>Touchdown PCR</i>	77
2.3.2.5	<i>FastStart High Fidelity PCR</i>	77

2.3.3	Agarose Gel Electrophoresis	78
2.3.4	Alkaline Gel Electrophoresis.....	79
2.3.5	DNA Sequencing.....	79
2.4	Parologue Ratio Test (PRT)-based Assays.....	81
2.4.1	Series A Assays	81
2.4.1.1	<i>Data Analysis</i>	82
2.4.1.2	<i>Identifying Classes of Copy Number Variation</i>	82
2.4.2	Series B Assays	83
2.5	<i>In Silico</i> Sequence Studies	84
2.5.1	<i>De Novo</i> Sequence Assembly	84
2.5.2	Dot Plots.....	84
2.5.3	BLAT Alignments	85

Chapter 3: Development of Oligo-Array CGH on the Geniom One Microarray Platform

3.1	Introduction.....	87
3.1.1	The Geniom One Microarray System	89
3.2	Optimisation of oaCGH on the Geniom Platform.....	92
3.2.1	Array Design	92
3.2.2	DNA Quantity and Concentration	93
3.2.3	Optimisation of Labelling Conditions	95
3.2.3.1	<i>Reaction Time</i>	96
3.2.3.2	<i>Concentration of Nucleotides</i>	96
3.2.3.3	<i>Addition of extra enzyme</i>	99
3.2.3.4	<i>Optimal Labelling Protocol</i>	99
3.2.4	Hybridisation Time	101
3.2.5	Adjusting the Stringency of Hybridisation	103
3.2.6	Mixing During Hybridisation	104
3.2.7	Smearing within Microchannels.....	106
3.2.8	Digestion of DNA with Restriction Enzymes	107
3.2.9	Summary of Development of oaCGH on the Geniom Platform.....	109
3.2.10	Probe Optimisation	109
3.2.10.1	<i>GC Content</i>	110
3.2.10.2	<i>Number of BLAT hits</i>	112
3.3	Data Analysis and Identification of Structural Variation using the Geniom Platform	114
3.3.1	Removal of Background	115
3.3.2	Normalisation.....	116
3.3.3	Displaying oaCGH Data Using a Logarithmic Scale.....	117
3.4	Summary & Discussion.....	120
3.4.1	DNA Concentration and Labelling	120
3.4.2	Identification of Structural Variation	121
3.4.3	Network Formation.....	121
3.4.4	Summary	123

Chapter 4: Investigating Structural Variation within a Novel Tandem Duplication on Chromosome 12p13.31

4.1	Introduction	124
4.2	Sequence Analysis	129
4.3	Structural Investigations	132
4.3.1	Analysis of Variant Boundaries	132
4.3.2	De Novo Sequence Assembly	134
4.3.3	Confirmation of Region Structure Using PCR.....	136
4.3.4	Identification of a Polymorphic Alu Element	137
4.4	Development of PRT-based Assays	142
4.4.1	Correction of Variation Between Different Batches of DNA Samples	143
4.5	Studying Variation in HapMap Samples	148
4.6	Investigating CNV Inheritance	152
4.7	Summary and Discussion	156

Chapter 5: Association Analysis of a CNV Located at 12p13.31 with Rheumatoid Arthritis and Other Complex Disorders

5.1	Introduction	159
5.1.1	Samples	161
5.2	Swedish Rheumatoid Arthritis Cohort	162
5.2.1	Genotypes Distinguishable using the A9 Assay.....	162
5.2.2	Determining Boundaries for Each Category of Variation.....	165
5.2.3	Results of a Swedish RA Association Study	167
5.3	UK RA Replication Cohort	171
5.3.1	Determining Category Boundaries for UK Samples	171
5.3.2	Replicating the A9(B) Deletion Association with RA	176
5.3.3	Investigating Other Forms of Variation.....	178
5.3.4	The Effect of Adjusting Category Boundaries	179
5.4	Swedish Psoriasis Cohort	181
5.5	Cardiovascular Disease.....	183
5.6	Summary and Discussion	185

Chapter 6: Investigating Historical Recombination Events at 12p13.31

6.1	Introduction.....	189
6.2	Characterisation of Sequence Identity	192
6.2.1	Dot Plots.....	192
6.2.2	BLAT Alignments	194
6.3	Development of a New Series of Assays for CNV.....	198
6.3.1	Assay Design.....	199
6.3.2	Optimisation.....	201
6.3.3	Detection of Putative Recombination Intervals.....	203
6.4	Identifying Points of Recombination.....	213
6.4.1	Primer Design.....	213
6.4.2	Primer Optimisation.....	215
6.4.3	Investigating Recombination within the B9/B10 Interval.....	216
6.4.3.1	<i>A9(B) Duplication</i>	<i>216</i>
6.4.3.2	<i>A9(B) Deletion</i>	<i>220</i>
6.4.4	Investigating Recombination within the B5/B6 Interval.....	222
6.4.5	Validation of Genotyping Results.....	223
6.5	Summary and Discussion	226

Chapter 7: Discussion

7.1	Development of oaCGH on the Geniom Platform.....	232
7.2	Characterisation of CNV at 12p13.31	236
7.2.1	Assay Development.....	236
7.2.2	Studies of Inheritance	238
7.2.3	Population Differences.....	238
7.3	Association of CNV at 12p13.31 with Complex Disease	240
7.3.1	RA Association.....	241
7.3.2	Physiological Effects of a Deletion within 12p13.31	242
7.3.3	Chondrocytes in RA Pathogenesis.....	243
7.3.4	Future Perspectives.....	244

Appendix A.....	246
------------------------	------------

Appendix B.....	253
------------------------	------------

Appendix C.....	258
------------------------	------------

References.....	261
------------------------	------------

List of Abbreviations

BAC	Bacterial Artificial Chromosome
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-like Local Alignment Tool
bp	base pair
cDNA	Complementary DNA
CCP	Cyclic Citrullinated Peptide
CEPH	Centre d'Etude du Polymorphisme Humain
CGH	Comparative Genomic Hybridisation
CI	Confidence Interval
CNV	Copy Number Variation
CVD	Cardiovascular Disease
DASH	Dynamic Allele-Specific Hybridisation
DNA	Deoxyribonucleic acid
ECM	Extra-cellular Matrix
EDTA	Ethyl-enediaminetetraacetic acid
FISH	Flourescent <i>in situ</i> Hybridisation
GWAS	Genome-wide Association Study
HGNC	HUGO Gene Nomenclature Committee
HIV	Human Immunodeficiency Virus
HLA	Human Leukocyte Antigen
HUGO	Human Genome Organisation
JRA	Juvenile Rheumatoid Arthritis
kb	kilobase
LD	Linkage Disequilibrium
LTR	Long Terminal Repeat
MAF	Minor Allele Frequency
MAPH	Multiplex Amplifiable Probe Hybridisation
MCC	Molecular Copy-number Counting
MEPS	Minimum Efficient Processing Segment
MES	2-(N-morpholino)ethanesulphonic acid
MHC	Major Histocompatibility Complex

MLPA	Multiplex Ligation-dependent Probe Amplification
mRNA	Messenger RNA
MSV	Multi-site Variant
NAHR	Non-allelic Homologous Recombination
NHEJ	Non-homologous End Joining
nt	Nucleotides
oaCGH	Oligo-array Comparative Genomic Hybridisation
OR	Odds Ratio
PCR	Polymerase Chain Reaction
PFGE	Pulsed-field Gel Electrophoresis
PLIS	Pooled Local Index of Significance
PRT	Paralogue Ratio Test
qPCR	Quantitative Polymerase Chain Reaction
QTL	Quantitative Trait Locus
RA	Rheumatoid Arthritis
RF	Rheumatoid Factor
RNA	Ribonucleic acid
RNAi	Ribonucleic acid Interference
SAPE	Streptavidin Phycoerythrin
SINE	Short Interspersed Nuclear Element
SLE	Systemic Lupus
SNP	Single Nucleotide Polymorphism
SSPE	Saline-Sodium Phosphate-EDTA buffer
Sw	Swedish
TBE	Tris-borate buffer
TE	Tris-EDTA buffer
Tm	Melting Temperature
UCSC	University of California Santa Cruz
UV	Ultraviolet
VNTR	Variable Number Tandem Repeat
WGA	Whole Genome Amplification
WTCCC	Wellcome Trust Case Control Consortium

List of Figures

Figure 1.1: Global Variation in Copy Number	13
Figure 1.2: Non-Allelic Homologous Recombination.....	18
Figure 1.3: BAC Array CGH.....	36
Figure 1.4: Effects of Rheumatoid Arthritis	40
Figure 1.5: Association of SNPs on Chromosome 12p13.31 with RA.....	59
Figure 1.6: Structural Variation at 12p13.31.....	61
Figure 2.1: Selection of X Chromosome Clones	69
Figure 3.1: Structure of a Geniom Biochip.....	90
Figure 3.2: Relationship Between Target DNA concentration and Spread of Signal	94
Figure 3.3: Relationship Between Length of Reaction & Yield of Labelled Product	97
Figure 3.4: Effect of Nucleotide Concentration on Labelling Efficiency.....	98
Figure 3.5: Addition of Extra Enzyme During Labelling Reaction.....	100
Figure 3.6: Results of Labelling Optimisation.....	101
Figure 3.7: Comparison of a Range of Different Hybridisation Times	102
Figure 3.8: Inclusion of Formamide in Hybridisation Solution.....	104
Figure 3.9: Mixing Hybridisation Solution During Hybridisation	106
Figure 3.10: Network Formation May Lead to ‘Smearing’ within Microchannels.....	108
Figure 3.11: Fragmenting DNA with a Double Rather than Single Digest.....	109
Figure 3.12: Relationship Between Probe GC Content and Signal Intensity.....	111
Figure 3.13: Detection Image of Geniom Biochip Used for Analysis.....	115
Figure 3.14: Log ₂ Plot Comparing Relative Hybridisations from Two DNA Samples	118
Figure 3.15: Identifying Structural Variation Using the Geniom Platform	119
Figure 4.1: Structural Variation at 12p13.31	125
Figure 4.2: Trace Data Reveals a Tandem Duplication at 12p13.31	126
Figure 4.3: Tandem Duplication at 12p13.31	127
Figure 4.4: Investigating the Evolutionary Age of the Tandem Duplication	129
Figure 4.5: Primers Designed Around Theoretical Boundaries.....	133
Figure 4.6: Results of a <i>de novo</i> Sequence Assembly	135
Figure 4.7: Gaps between Contigs Identified using <i>de novo</i> Sequence Assembly.....	136
Figure 4.8: Identification of a variant sample	137
Figure 4.9: Separation and Purification of the Products from a DNA Variant.....	138
Figure 4.10: Identification of a Polymorphic Alu Element	139
Figure 4.11: Investigating Variant Products Using an Alkaline Gel	140
Figure 4.12: The Location of PRT-based Assays Within the Tandem Duplication	144
Figure 4.13: A9 Assay	144
Figure 4.14: Elimination of Batch Effect.....	145
Figure 4.15: Identification of Copy Number Variation in Swedish Control Samples..	147
Figure 4.16: Inheritance in HapMap Trios	151

Figure 4.17: Patterns of CNV Inheritance in CEPH Families	153
Figure 5.1: Four Types of Structural Variation	163
Figure 5.2: Density Plot of Swedish RA Data	166
Figure 5.3: Density Plots comparing of Swedish RA Case and Control Data.....	169
Figure 5.4: Frequency of variants in Swedish RA cases compared to controls.....	170
Figure 5.5: Frequency of Variants in UK & Swedish Control Populations	172
Figure 5.6: Comparing UK and Swedish Control Sample Data to Determine Category Boundaries	174
Figure 5.7: Density Plots of UK RA Case and Control Data	175
Figure 5.8: Swedish Psoriasis Cohort Data	182
Figure 5.9: Frequency of Variation in UK CVD Case-Control Cohort.....	184
Figure 6.1: NAHR within the Tandem Duplication at 12p13.31.....	190
Figure 6.2: Dot Plots to show Regions of Sequence Identity	193
Figure 6.3: The Location of the Longest Regions of Identical Sequence.....	194
Figure 6.4: Graphical View of Sequence Similarity Between the Two Units of the Tandem Duplication	197
Figure 6.5: Investigating the Location of Recombination Events	198
Figure 6.6: Design of B Assays	200
Figure 6.7: Location of B Assays	200
Figure 6.8: Optimisation of an Unequal B Assay	202
Figure 6.9: Investigating of Sites of Recombination Using B Assays.....	204
Figure 6.10: Locating Site of Recombination using Assay B9.....	206
Figure 6.11: Identification of Two Distinct Points of Recombination	207
Figure 6.12: B Assays Reveal Two Distinct Sites of Recombination	210
Figure 6.13: Putative Sites of Historical Recombination	211
Figure 6.14: Design of Primers to Reveal Recombination Events	214
Figure 6.15: Recombination Within the B9/B10 Interval Resulting in a Duplication .	217
Figure 6.16: Identification of Recombination Within a 1.1 kb Region (1)	218
Figure 6.17: Identification of Recombination Within a 1.1 kb Region (2)	219
Figure 6.18: Direct PCR to Identify a B Duplication Event.....	221
Figure 6.19: Recombination at B9/B10 resulting in a Deletion	222
Figure 6.20: Structure of Sequence Similarity in the B5/B6 Interval.....	223
Figure 6.21: NAHR in the B Unit of the Tandem Duplication.....	229

List of Tables

Table 1.1: RA Susceptibility Loci for which Association has been Replicated.....	52
Table 2.1: DNA Samples	64
Table 2.2: Equipment.....	65
Table 2.3: Computer Software.....	66
Table 2.4: Probes included on Geniom Biochip	68
Table 3.1: Conditions which Enhance Network Formation	107
Table 3.2: Summary of Conditions used for oaCGH on the Geniom Platform.....	110
Table 4.1: Frequency of Variation in Four Populations	148
Table 5.1: Classes of Genotype Detectable With the A9 Assay.....	164
Table 5.2: Product Ratios for Detectable Genotypes.....	164
Table 5.3: Categories of Variation.....	167
Table 5.4: Swedish Rheumatoid Arthritis Data	168
Table 5.5: Frequency of Variants in UK and Swedish Control Populations	171
Table 5.6: UK RA Data	177
Table 5.7: Summary of RA results from Swedish and UK studies	178
Table 5.8: UK RA Frequency Data for all Variants	179
Table 5.9: Effect of Changing Category Boundaries.....	180
Table 5.10: Swedish Psoriasis Data	181
Table 5.11: UK Cardiovascular Disease Data	183
Table 6.1: Location of Historical Recombination Events.....	212
Table 6.2: Control Primers for B5/B6 Interval	215
Table 6.3: Control Primers for B9/B10 Interval	215
Table 6.4: Confirmation of Variants in UK Samples using Recombination PCRs	224
Table 6.5: Population Differences in the Location of Recombination Events	225
Table 6.6: Effect of Recombination Events within the B5/B6 and B9/B10 Intervals on Genes within the Tandem Duplication.....	229

Chapter 1

Introduction

1.1 Genetic Variation

Genetic variation is the term used to describe differences in the genetic code seen between individuals. This can take many forms, from changes involving a single base of the nucleotide sequence, up to the gain or loss of one or more of the 46 chromosomes found in every human cell. The different alleles of genetic variants can be formed in many ways, for example as a result of UV damage or replication errors. New combinations of chromosomal alleles are continuously being generated through recombination, a process which occurs during meiotic cell division in which homologous chromosomes align and exchanges of DNA sequence occur between them. An average of 30 recombination events are thought to occur during each meiotic division, although the frequency is higher in female compared to male meiosis (Cheung *et al*, 2007). Recombination is an important source of genome variation and plays a role in environmental adaption and evolution.

Large-scale variations, such as changes in chromosome number and translocations involving extensive regions of DNA, are amongst the easiest forms of variation to detect as they can often be visualised using a microscope. X or Y (sex chromosome) copy number changes are usually tolerated, although they may have phenotypic effects; for example, the most common sex chromosome disorder is Klinefelter's syndrome, which is characterised by an XXY karyotype. In contrast, variations from two copies of each of the 22 autosomal (non-sex) chromosomes are not usually viable. However, there

are exceptions, for example a trisomy of chromosome 21 results in the Down's Syndrome phenotype.

Single base changes to the DNA sequence are one of the most common forms of human genetic variation. The term single nucleotide polymorphism (SNP) is traditionally used to refer to single base changes for which the most common allele occurs in the population at a frequency of less than 99%. SNPs may take the form of substitutions, deletions or insertions and most are either diallelic or, less commonly, triallelic. The minor allele frequency (MAF) is used to describe the rate of occurrence of the least common allele of a SNP. This value tends to vary between populations (International HapMap Consortium, 2005).

The functional impact of a SNP is dependent upon its location. Due to the degeneracy of the genetic code, base substitutions within a coding region of the genome may not always alter the amino acid sequence of the protein. SNPs that do not alter the protein sequence are referred to as 'silent'. Other substitutions may change the amino acid sequence ('mis-sense'), or code for a stop codon which may prematurely terminate translation ('non-sense'). An insertion or deletion can shift the reading frame and alter the amino acid sequence of the protein. This may lead to a change in protein structure, which can result in a loss, decrease or change in protein function. SNPs located in non-coding regions may have little effect, although those in regulatory regions can have phenotypic consequences.

SNPs tend to be inherited as haplotype blocks, which are made up of combinations of alleles between which there is little evidence for historical recombination events (Gabriel *et al*, 2002). The transmission of alleles together at a higher frequency than would be expected by chance is known as Linkage Disequilibrium (LD). For markers in

LD, the genotype at one locus can be used to predict the genotype at another (usually nearby) locus. SNPs within a haplotype block can therefore be used as proxies for each other, as well as for hidden disease-associated variants. This provides a useful tool for the study of disease, as it reduces the number of markers needed for a study, and may also overcome difficulties associated with the inclusion of problematic regions of the genome, for example those with a high density of repeat elements. Haplotype blocks are also useful for the study and location of recombination ‘hotspots’, a term used to describe regions of the genome in which recombination events frequently take place (Myers *et al*, 2005).

Although SNPs are probably the most well studied form of genetic variation, there are also other types of variation with unique characteristics which can be exploited for research purposes. Variable number tandem repeats (VNTRs) are short nucleotide sequences which are highly polymorphic in terms of the number of times they are repeated in tandem. Types of VNTRs include minisatellites, for which the repeated unit is typically 10-60 base pairs long, and the smaller microsatellites which are 1-6 bp in length. VNTRs are found at many locations in the genome. The length of the tandem repeat at each position is used to build an individual’s unique VNTR profile. Since the number of tandem repeats at each position acts as an inherited allele, an individual’s VNTR profile is a combination of those of their parents (Wyman & White, 1980). The fact that each individual has a distinguishable inherited genetic profile of these repeat elements forms the basis of the DNA fingerprinting technique (Jeffreys *et al*, 1985) which has multiple uses, for example in forensics and paternity determination.

More recently there has been increasing interest in a form of variation known as copy number variation (CNV). Extensive regions of the genome have been shown to vary in

copy number between individuals, resulting in a range of phenotypic effects. This form of variation is discussed in detail in section 1.2.

1.1.1 Genetic Variation and Disease

Genetic variation contributes towards many of the phenotypic differences between individuals, including susceptibility to disease. An improved understanding of the genetic basis of disease has great potential to lead to improvements and developments in diagnosis, treatment and prognosis; therefore this field is currently a major focus of research.

Diseases which involve a genetic component are typically categorised as either single gene disorders (used to refer to those caused by mutations affecting just one gene) or complex disorders (which develop as a result of multiple genetic and environmental factors). Early studies focussed on single gene disorders, since the causal variants responsible for these phenotypes were the simplest to identify. As the sensitivity and throughput of techniques improved, it became possible to study genetic variation on a larger scale, which led to advances in the study of complex disorders, susceptibility to which typically involves multiple variations with relatively small effects. The differing approaches used to study single and complex gene disorders are discussed below.

1.1.1.1 *Linkage Studies and Single Gene Disorders*

Two or more markers which are located close together on a chromosome are more likely to be transmitted together than those further apart or on different chromosomes, due to the low chance of recombination occurring between them. This is known as 'linkage'. Linkage maps are created to show the relative location of known genes or

markers on chromosomes, based on the frequency of recombination events between them. These maps can be used to determine the location of the genetic variants responsible for other phenotypic traits, including those involved in disease. By studying how alleles segregate between phenotypically different individuals, the relative locations of these alleles can be determined.

The first linkage studies were carried out using fruit flies (*Drosophila*) at the start of the 20th Century. Although a useful tool for studying model organism genetics, this process is difficult to apply to humans due to the inability to create desired crosses and the tendency towards small families. For most of the 20th century, this was a significant limitation to human genetic research. However, the discovery in the 1980s that naturally occurring DNA variations could be used as markers to track inheritance and recombination (Botstein *et al*, 1980) went some way to solving this problem. Due to recombination, any marker which is seen to co-segregate with the trait of interest must lie nearby to the causal variant. Many of the causal genes and mutations responsible for single gene disorders have been identified using linkage studies, for example cystic fibrosis (Kerem *et al*, 1989) and Huntington's disease (The Huntington's Disease Collaborative Research Group, 1993).

1.1.1.2 *Association Studies and Complex Disorders*

Complex disorders such as rheumatoid arthritis, diabetes and cancer occur as a result of multiple environmental and genetic factors. The genetic component is thought to involve variations at a number of distinct loci which each confer a relatively modest effect to disease risk. Since such disorders rarely follow a simple pattern of inheritance, it is difficult to identify susceptibility loci using linkage studies.

Association studies have become the method of choice for the study of complex disease. This approach aims to determine whether a particular genetic variant is associated with a certain phenotype by comparing its frequency in two groups; one of individuals with the required phenotype and the other a control cohort without the phenotype of interest. Genome-wide association studies became possible as a result of the publication of the human genome sequence (Lander *et al*, 2001; Venter *et al*, 2001), and subsequent mapping of common genetic variations in four populations (European, African and Asian (Chinese and Japanese)) as part of the HapMap Project (International HapMap Consortium, 2003).

The association study approach was suggested as an effective method to investigate the common disease-common variant hypothesis, which proposed that common variants are responsible for a significant proportion of susceptibility to common disorders (Risch & Merikangas, 1996; Lander, 1996). Whereas the aim of linkage studies is to localise alleles by segregation and recombination in pedigrees, genetic association studies determine whether an allele, and sequence in LD with that allele, is implicated in a phenotype of interest. If a statistically significant association is detected, depending on the degree of significance, replication may be required to confirm the association.

1.1.1.3 *Determining the Statistical Significance of Results*

The p-value, or probability value, is used to determine whether an observed association is statistically significant. This describes the likelihood of obtaining the same result purely by chance. P-values take the form of a number between 1 and 0; a value close to 1 indicates that it is likely the result in question has occurred due to chance and the effect seen is not significant. The smaller the p-value, the more statistically significant the results are considered to be. The cut off point used to determine significance is

dependent on the application and the number of factors being tested. 0.05, or the likelihood of obtaining a false result 1 time in 20, is often considered to be the universal threshold for statistical significance for a single factor. The reason for this is largely convention; Ronald Fisher suggested this value in his 1925 book *Statistical Methods for Research Workers*, as a matter of convenience and also due to the fact that, in a normal distribution curve, the area for which $p=0.05$ is located approximately 2 standard deviations from the mean (1.96) (Fisher, 1925). Since then, a p-value of ≤ 0.05 has become synonymous with a statistically significant result. However, it should be emphasised that this is an arbitrary method of determining significance and that it is important to consider the results themselves in each case; for example, there is not a great deal of difference between a p-value of 0.06 and 0.05, yet one is considered significant while the other is not. Factors such as sample size may affect the p-value, since a study containing a small number of samples may not have enough power to detect an association with a relatively low effect.

1.1.1.4 *Adjusting for Multiple Testing*

As the number of factors under investigation increases, the more likely it is that the case and control groups will appear to differ significantly with regard to at least one of them by chance. Therefore as the number of factors being tested increases, confidence in the results decreases. A common method employed to overcome the problem of multiple testing is to adjust significance levels. For example, the Bonferroni correction tests each individual hypothesis at a significance level of $1/n$ times what it would be for a single hypothesis, where n is the number of hypotheses being tested (Rice *et al*, 2008). Although this approach works well for smaller studies, the Bonferroni correction tends to overcorrect, especially in the case of large investigations such as Genome-Wide

Association Studies (GWAS), where many hundreds of thousands of SNPs may be tested simultaneously. It also makes the assumption that each test is independent, which is not the case for SNPs due to correlations between markers, for example those in LD.

Many GWAS attempt to correct for multiple testing by raising the significance level threshold. For example, a recent Wellcome Trust Case Control Consortium (WTCCC) study which carried out a genome-wide search for loci associated with seven common diseases used a significance level of 5×10^{-7} . This threshold was calculated by taking into account the power of the study as well as an estimation of the odds of a true association (Wellcome Trust Case Control Consortium, 2007). Other methods of correction include permutation testing, which empirically determines how often a given p-value would occur by chance if the study was repeated in the absence of any true associations, using p-values calculated from a number of repeated permutations of the data (discussed in Rice *et al*, 2008). Also proposed is a method termed Pooled Local Index of Significance (PLIS), which takes into account the dependency between adjacent SNPs (Wei *et al*, 2009).

1.1.1.5 *Calculating the Strength of an Association*

The odds ratio or risk ratio can be used to determine the strength of an association. The odds ratio (OR) measures the likelihood (odds) of an event occurring in one group against the likelihood of the same event occurring in a second group. A similar measurement is calculated by the risk ratio, or relative risk, however this determines the probability, rather than the odds, of the event occurring in each group. Although these are slightly different variables, they are often used interchangeably, especially in disease association studies. In such investigations, the relative risk (probability of developing

the disease) is in some ways the more useful measurement. However, since the fundamental design of case-control studies means there is a bias for selecting patients who are at risk (those in the case group), as well as the fact that these studies observe the situation at a single time point rather than tracking individuals over a period of time, it is not appropriate to calculate the relative risk. In this situation, as long as the disease outcome is rare (has an occurrence of less than 10%), the odds ratio is a good approximation of the relative risk (discussed in Cummings, 2009).

The odds ratio is calculated using the formula:

$$\frac{p_1/(1 - p_1)}{p_2/(1 - p_2)} = \frac{p_1/q_1}{p_2/q_2} = \frac{p_1q_2}{p_2q_1}$$

In the above equation, p represents the probability of the event occurring in two groups, 1 and 2. The resulting value describes the size of the effect; a ratio of 1 indicates that the event under investigation is equally likely to occur in both groups, whereas an odds ratio of >1 indicates that the event is more likely to occur in one group rather than another. For example, in a disease association study the two groups in question would be the case and control populations, and an odds ratio of 3 would suggest that an individual with the allele of interest is approximately 3 times more likely to develop the disease than an individual with a different allele at the same position. Odds ratios are usually given with 95% confidence intervals (CI). For example, OR=3 (95% CI 2.4-3.6) indicates that one can be 95% sure that the odds ratio lies between the confidence limits of 2.4 and 3.6.

1.1.1.6 *Genome-Wide Association Studies*

There has been a recent explosion of genome-wide association studies (GWAS), which have identified putative susceptibility loci for a wide range of diseases from neurological disorders such as schizophrenia (International Schizophrenia Consortium, 2008) to autoimmune diseases such as diabetes and rheumatoid arthritis (WTCCC, 2007), as well as many cancers (for example, Diskin *et al*, 2009; Turnbull *et al*, 2010). Associations identified through such studies are often independently validated using follow-up investigations, which may also investigate the potential pathways involved. One approach is to attempt to locate the precise causal variant using methods such as in-depth resequencing (McCarthy *et al*, 2008). However, it is often not possible to identify a causal variant within the region of association. A proposal has been put forward which suggests that the clustering of rare susceptibility variants, which may be more strongly associated with one allele of the marker site rather than the other, can create so called ‘synthetic associations’ (Dickson *et al*, 2010). These signals are derived from rare variants acting over large distances, which are falsely credited to the haplotype block containing the marker. It is thought that the more rare causal variants there are located within the surrounding region, the greater the chance of a synthetic association. Currently it is not known what proportion of GWAS signals can be attributed to such associations, however this is an important factor to consider as a large number of rare mutations may collectively contribute towards a significant portion of the missing risk for complex disorders. Identifying such variants may provide an important insight into the cause and development of disease.

The identification of susceptibility genes often provides a valuable insight into disease pathogenesis. This is an important field of research, as an understanding of disease pathology can aid the development of treatments as well as improve diagnosis and

preventative care. For example, an individual with an increased genetic susceptibility to heart disease can be continuously and carefully monitored for signs of the illness. There is also increasing interest in the field of pharmacogenomics, which studies the way that variations in an individual's genome may affect the way they respond to different drugs (Reviewed in Katz & Bhathena, 2009). It is likely that in the future there will be greater emphasis on an individual's genetics when considering susceptibility, diagnosis and treatment of disease.

1.2 Copy Number Variation

Over the last decade there has been increasing interest in a form of genetic variation known as copy number variation (CNV). This describes regions of sequence for which the number of copies differs between individual genomes, and is an important class of variation present in humans as well as other species (Reviewed by Feuk *et al*, 2006). The Database of Genomic Variants, which aims to provide a summary of known structural variations present in the human genome (Iafrate *et al*, 2004), currently (as of June 2010) lists 57829 CNVs. This number has rapidly increased over the last few years as new technologies have been applied.

1.2.1 Early Genome-Wide Studies of CNV

Variations in the copy number of large regions of sequence have been recognised for many years, but these were thought to be rare and usually identified as a result of their association with disease phenotypes. For example a range of structural variations, which result in rearrangements and changes to the genomic architecture rather than alteration of the underlying DNA sequence, were found to occur in patients with mental retardation (Jacobs *et al*, 1978). However, the extent to which CNV occurs in the normal population was not realised until 2004 when technological advances made it possible to carry out genome-wide studies of variation. Two parallel studies used microarrays to reveal the presence of large scale copy number variants in the genomes of phenotypically normal individuals (Iafrate *et al*, 2004; Sebat *et al*, 2004). These studies detected variants which were typically several 100s of kb in size, many of which were found to be present in multiple individuals. Both studies noted that a significant proportion of the variants detected seemed to overlap with coding regions of the

genome. However, in both of these examples the arrays used contained probes covering less than 13% of the genome, leaving many regions unstudied.

A couple of years later, another genome-wide study was carried out using a combination of two methods; SNP genotyping and arrayCGH (Comparative Genomic Hybridisation, see Section 1.4.3 for more detail) (Redon *et al*, 2006). This investigation covered a much greater proportion of the human genome (up to 93.7% of euchromatic regions) and used a larger number of samples (270 compared to 20 (Sebat *et al*, 2004) and 39 (Iafrate *et al*, 2004) in the two studies previously described). The results of this study suggested that as much as 12% of the genome may be copy number variable (Figure 1.1).

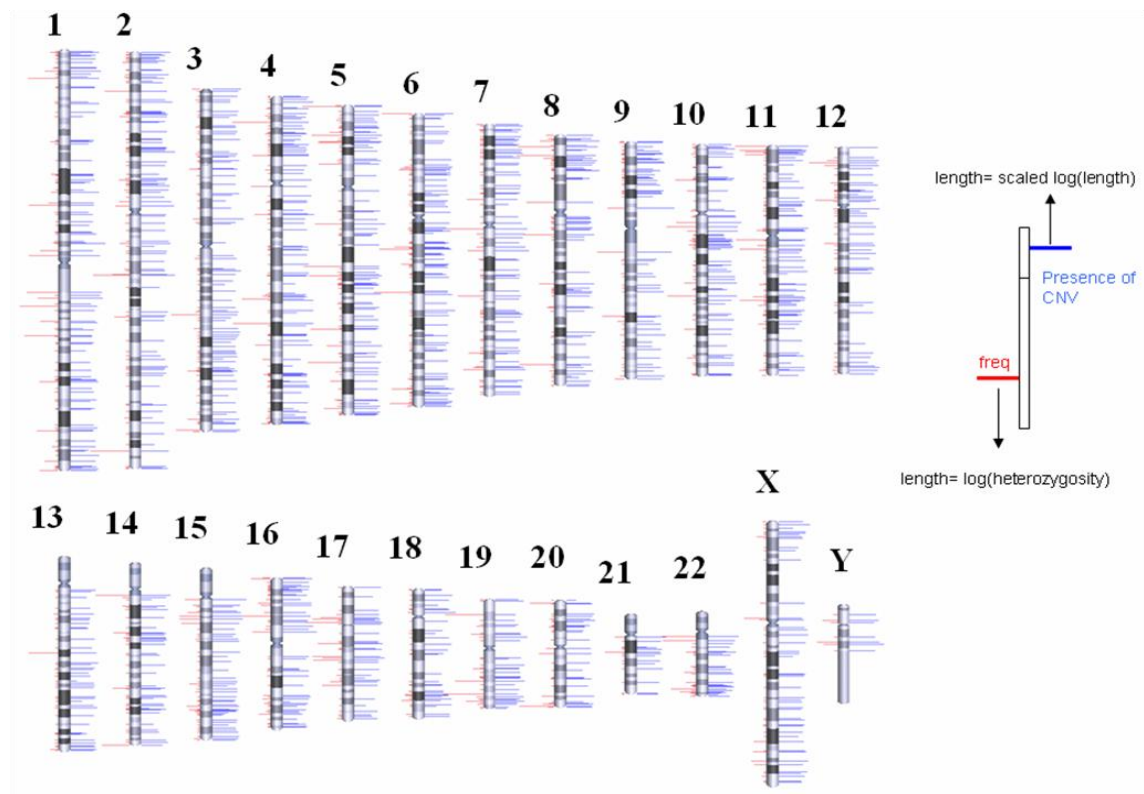


Figure 1.1: Global Variation in Copy Number

Redon *et al* (2006) performed a genome-wide study of copy number variation in 270 individuals which revealed the presence of widespread copy number variation in the genome of apparently healthy individuals. The location of CNVs is shown on each of the 24 chromosomes. Blue lines indicate the presence CNV, whereas red lines show the frequency of each variant.

At the time, the figure of 12% was considered to be a conservative estimate, since resolution limitations of the technologies employed for this study meant that only CNVs of greater than 1 kb in size were included. However, recent investigations have concluded that the figure of 12% is likely to be an overestimate; it has been suggested that as many as 88% of CNVs are actually smaller than recorded in the Database of Genomic Variants, highlighting the need for fine scale mapping of CNVs and identification of breakpoints (Perry *et al*, 2008). Nevertheless, it is estimated that CNVs are likely to encompass more nucleotide content in the human genome than SNPs (Redon *et al*, 2006), which were previously considered to be the main source of genome variability.

Investigations of CNV typically employ their own definitions relating to the size of variants. For example, many studies define copy number variation as involving segments of DNA greater than 1 kb in size (Redon *et al*, 2006). However, this arbitrary cut-off may in part be due to the resolution restrictions of common techniques used to reveal variation. Improvements in resolution and the use of new methods for CNV detection, such as in-depth resequencing, are likely to lead to the identification of smaller regions of CNV. Applying a size restriction means that studies may miss potentially important pathogenic CNVs which involve regions of less than 1 kb. It has already been shown that many known variants involve much smaller regions of sequence than initially thought due to difficulties in precisely identifying variant boundaries (Perry *et al*, 2008).

1.2.2 CNV Genomics

A number of similarities have been identified between regions of CNV. Many lie within complex regions of the genome, such as those containing segmental duplications, which are defined as blocks of sequence larger than 1 kb in size which have a sequence similarity of greater than 90% (reviewed by Eichler, 2001). Unlike CNVs, the number of segmental duplications was thought not to vary between individuals; however, there is considerable overlap between the two classes and many segmental duplications have been shown to be copy number variable. Segmental duplications are often present in regions of the genome which are susceptible to recombination and structural changes, which has lead to the suggestion that they may play a role in defining hotspots of chromosomal rearrangement (Sharp *et al*, 2005). The presence of CNV in such regions supports the idea that certain areas of the genome are more unstable (and therefore more likely to undergo recombination and rearrangements) than others.

Wong *et al* (2007) identified 3654 autosomal CNVs using whole genome BAC arrayCGH, 800 of which were found to occur at a frequency of at least 3%. Of these, 68% were found to overlap with genes. Certain functional classes of genes seem to be enriched in CNVs, including genes involved with the senses (including olfactory and taste receptors), immune system and also many genes associated with cancer for example tumour suppressor genes (Wong *et al*, 2007). CNV in regions of the genome which contain genes and regulatory elements may have a significant effect on gene expression and phenotypic variation. This is discussed in detail in section 1.3. Within genes, duplications occur more frequently than deletions (Redon *et al*, 2006). This suggests that sequence duplications are less detrimental, and therefore are more likely to be retained.

1.2.3 Differences in Copy Number Variation between Populations

Population differences have been detected in both the number of copy number variable regions identified, as well as the diversity of individual CNVs (Redon *et al*, 2006). Results from the HapMap project revealed a higher number of SNPs in samples of African origin compared to the other populations studied (International HapMap Consortium, 2005), which is in agreement with the African origin of the human species. A study of copy number variation in the same set of samples saw a similar effect with CNVs; the number of variations in individuals of Yoruban descent from Nigeria, Africa was double that seen in the other individual populations (Armengol *et al*, 2009).

CNVs may be easier to detect in some populations than others, due to the differing levels of heterozygosity. This emphasises the importance of studying a diverse range of populations to allow a comprehensive map of CNVs to be produced. Population variation in allele frequencies also highlights the importance of obtaining closely matched case and control groups for CNV association studies, to prevent such differences leading to false positive results.

1.2.4 Categorising and naming CNVs

CNVs are typically classed according to the number and complexity of rearrangements. Distinctions are made between diallelic events (involving a single duplication or deletion) and complex CNVs, which may be a result of numerous rearrangements and the exchange of genetic material between different chromosomes (Redon *et al*, 2006). Other possible categories include separating events based on the features of the surrounding genomic region, for example differentiating between CNVs which are

associated with segmental duplications or tandem repeats. Due to the relative youth of this field, there is currently no universal nomenclature for CNVs.

1.2.5 Mechanisms of CNV Formation

Copy number variations occur as a result of structural rearrangements which create duplications, deletions, insertions, translocations and complicated combinations of the above. The exact mechanisms by which CNVs are generated, and the frequency at which this occurs, remains unclear. It has been suggested that new CNVs are constantly being created (Egan *et al*, 2007); however, other studies have shown that more than 99% of CNVs are inherited rather than newly formed (McCarroll *et al*, 2008). When *de novo* events do take place, it has been shown that deletions are generated at a higher frequency than duplications in the male germ line (Turner *et al*, 2008).

Most CNVs have been shown to follow classical patterns of Mendelian inheritance (Locke *et al*, 2006). However, the process of investigating inheritance and accurately identifying *de novo* events is complicated by the fact that most CNV detection techniques are only able to detect total diploid copy number rather than identify copy number on each chromosome separately. This means that it is not possible to distinguish between *cis* (copies on same chromosome) and *trans* (copies on different chromosomes) rearrangements. This contributes to difficulties in determining inheritance patterns, as mixtures of copy number may lead to apparent “missing” or undetected CNV in one or more generations.

There are a number of mechanisms by which CNVs are thought to be created. These fall into two categories, recurrent rearrangements and non-recurrent rearrangements.

1.2.5.1 *Recurrent Rearrangements*

Recurrent rearrangements occur independently in unrelated individuals, and share common sites of rearrangement. These are typically the result of ectopic recombination between stretches of homologous sequences, termed Non-Allelic Homologous Recombination (NAHR). Normally, recombination occurs between equivalent sequences on homologous chromosomes when they align during meiosis. However, in the case of NAHR, low copy repeats misalign leading to the loss or gain of genetic material (Figure 1.2). For NAHR to occur, stretches of sequence above a minimal length which share extremely high levels of sequence similarity are required. These are termed Minimal Efficient Processing Segments (MEPS).

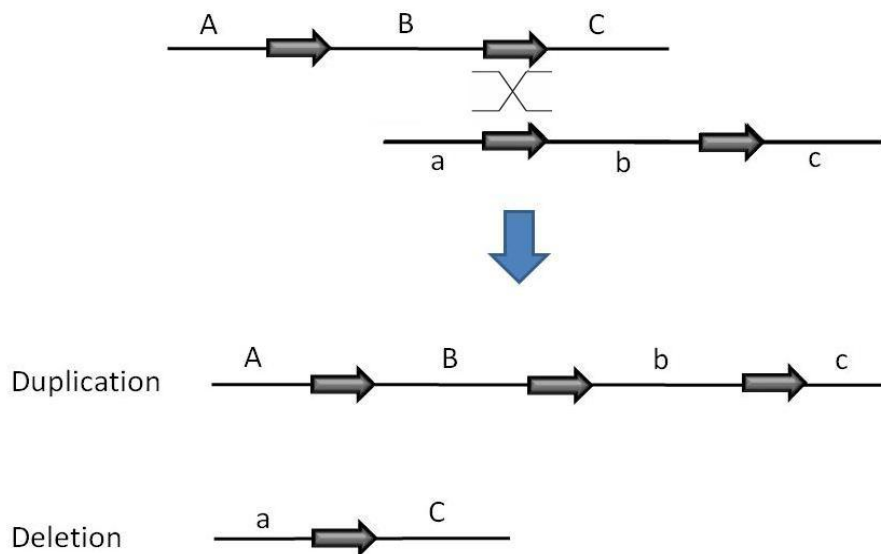


Figure 1.2. Non-Allelic Homologous Recombination

This figure illustrates the outcome of NAHR between low copy repeats on different chromosomes. Low copy repeats are indicated by arrows, and letters represent three genes for which the two sequences are heterozygous (represented by upper and lower case letters). NAHR recombination occurs between two low copy repeats, represented by a crossing over. This results in two products, one containing a duplication of sequence and the other a deletion.

In humans there is evidence to suggest that MEPs are required to be a minimum of 300-500 bp in length (Reiter *et al*, 1998), although examples of recombination between much shorter homologous sequences have been described. For example, Lam & Jeffreys describe NAHR between human α -globin genes which is mediated by homologous sequence fragments of less than 50 bp in size (Lam & Jeffreys, 2006).

NAHR is thought to be the most common mechanism by which copy number variants originate. A recent study which mapped structural variation in eight human genomes estimated that 47% of variations for which a mechanism could be assigned occurred as a result of ectopic recombination between regions of homologous sequence (Kidd *et al*, 2008). It has been suggested that NAHR may also occur between mobile sequence elements which have been inserted at different locations in the genome (Xing *et al*, 2009).

1.2.5.2 *Non-Recurrent Rearrangements*

Unlike recurrent rearrangements, non-recurrent rearrangements vary in size and have different breakpoints. Non-Homologous End Joining (NHEJ) is a mechanism involved in the repair of double-stranded breaks. As the name suggests, this process does not require the presence of regions of homology. It is thought that this mechanism may be responsible for many simple non-recurrent rearrangements (Reviewed in Gu *et al*, 2008).

Additional mechanisms have been proposed in an attempt to explain complex rearrangements which cannot be explained by a single recombination event. For example, whilst studying the gene PLP1 in patients suffering from the central nervous system disorder Pelizaeus-Merzbacher Disease, Lee *et al* observed non-recurrent

arrangements in the region surrounding the gene which could not be explained by NAHR or NHEJ. Duplications were seen to be interrupted by other rearrangements including deletions and even stretches of normal (non-variable) genomic sequence (Lee *et al*, 2007a). To explain this, they proposed a mechanism called Fork Stalling and Template Switching (FoSTeS). This describes a situation where the process of DNA replication stalls and switches template to another region of the genome which shows microhomology to the original sequence. This leads to either duplication or deletion of DNA sequence, depending on whether the second locus is located upstream or downstream of the initial replication site (Lee *et al*, 2007a). This process may occur several times, leading to a complex series of rearrangements. Similar events which involve a collapsed replication fork rather than fork stalling have been termed Microhomology-mediated Break-induced Replication (MMBIR) (Hastings *et al*, 2009).

1.2.5.3 *Somatic Recombination*

Whilst the vast majority of new CNVs are thought to be created in gametes as a result of recombination events which occur during meiosis, there is evidence to suggest that structural variations can also arise in somatic tissues during mitosis (Piotrowski *et al*, 2008). For example, it is possible for monozygotic twins, which arise from an early division of the same embryo and therefore should be genetically identical, to show regions of the genome which differ in copy number (Bruder *et al*, 2008). Somatic occurring changes in copy number such as these have historically been associated with disease phenotypes, for example cancerous cells often contain multiple large genomic rearrangements (Weir *et al*, 2007; Mullighan *et al*, 2007).

Since somatic cells do not undergo meiosis, structural variants must instead arise during the process of mitosis. However, compared to meiosis, in which recombination plays an integral role, mitotic recombination events are considerably rarer and less well understood. How widespread the generation of variants in somatic cells is, and the frequency at which this occurs, remains to be determined. It is thought that variations may occur as a result of errors in DNA damage repair, and therefore tend to be spontaneous rather than recurrent (discussed by LaFave & Sekelsky, 2009).

1.3 Phenotypic Effects of Copy Number Variation

CNV is abundant in the genomes of normal healthy individuals (Redon *et al*, 2006) and is therefore likely to contribute towards phenotypic variation seen between them. The effect conferred by individual CNVs very much depends on their location, for example whether they are found within a coding region. In the case of genes, the class of gene also appears to be important, as some are more tolerant of variation in copy number than others. For example, the β -defensin cluster on chromosome 8p23.1 is highly variable; normal individuals carry between 2 and 7 copies, with a median of 4 (Hollox *et al*, 2003).

1.3.1 The Role of CNV in Adaptation and Evolution

The relative abundance of CNV seen to occur within genes associated with the senses and immune system (Wong *et al*, 2007) has lead to the suggestion that such events may have been maintained due to the fact that they conferred a selective advantage, for example in times when heightened senses were beneficial (discussed by Nguyen *et al*, 2006).

Population differences in the allele frequency of copy number variants could be taken to support the idea that selection pressures act on this form of variation and may play a role in enabling populations to adapt to a changing environment. An example of this is seen in the case of the gene *AMY1*, which codes for the salivary enzyme amylase. This enzyme acts on starch and breaks it down into simple sugars (Lebenthal, 1987). The *AMY1* locus is amongst the most copy number variable in the genome (Iafrate *et al*, 2004), with diploid individuals ranging from 2 to over 10 copies. Populations which tend to have a high starch content in their diet also have, on average, a higher copy

number of this gene than those with low starch diets (Perry *et al*, 2007). This suggests a role for natural selection in favour of those individuals best able to digest their main food source.

Further evidence to support the theory that some CNVs may confer a selective advantage comes from a recent study which identified a number of loci which showed significant copy number differences between three populations (Armengol *et al*, 2009). Interestingly, within these loci there appeared to be an enrichment of genes with functions such as immune response, lipid metabolism and other factors relating to environmental adaptation. The population differences seen here were reflected in gene expression levels, which were shown to be altered in over half of the copy number variable loci studied. Genes at these loci provide interesting candidates for further study into their involvement of copy number variation in adaptation and evolution.

Despite a wealth of evidence to suggest that CNV within certain classes of genes may have, at some point, conferred a selective advantage, it is unlikely that all copy number variants would have been affected by selective pressures in this way. It is probable that many of the differences in copy number variation seen between populations may be a result of genetic drift, which describes the change in frequency of a variant which occurs purely by chance. Such changes are not affected by environmental pressures, and are likely to have a larger effect on allele frequency within small or isolated populations. It has been theorised that the majority of changes in variant frequency may occur as a result of drift rather than selection (Kimura, 1968), and therefore the likely effect of this process on CNVs should not be disregarded.

1.3.2 CNV and Disease

Despite the fact that CNV occurs at a significant frequency in the healthy population, many CNVs have been identified as causal variants for an increasing number of diseases. Initially structural variation was thought to contribute mainly to rare, sporadic disorders, but a role for this form of variation has also been identified in many common complex disorders including schizophrenia (International Schizophrenia Consortium, 2008), autism (Sebat *et al*, 2007) and HIV (Gonzalez *et al*, 2005).

A common mechanism by which CNV contributes towards disease susceptibility is by varying the gene count in dosage sensitive regions of the genome, leading either to an excess or deficiency of gene product, which may have detrimental effects. An example of this is seen in variation in the gene *PMP22* (Peripheral Myelin Protein 22) which lies on chromosome 17p11.2. The PMP22 protein is involved in the production of myelin, which is essential for proper functioning of the neurons. CNV at this locus arises as a result of recombination between two repeat units, the interval between which includes the gene *PMP22* (Pentao *et al*, 1992). As a result of this ectopic recombination, two products are formed, one containing a deletion and the other a duplication. These each give rise to distinct phenotypes; a duplication of *PMP22* results in the neuronal disorder Charcot-Marie-Tooth Disease Type 1A (Lupski *et al*, 1991), whereas a deletion in this region results in the peripheral nerve disorder Hereditary Neuropathy with liability to Pressure Palsies (HNPP) (Chance *et al*, 1993).

If the insertion of a new copy number variant disrupts a gene or regulatory region, this may cause disease through a loss of function mutation. An example of this is seen in red-green colour blindness. The red-green pigment genes involved in colour vision are present in a tandem array. Recombination between these genes which results in

disruption or deletion events is a cause of red-green colour blindness (Nathans *et al*, 1986).

While extensive non-pathogenic CNV is seen within the normal population, extremes are often detrimental. For example, the β -defensin cluster located on chromosome 8p23.1 has a diploid copy number of between 2 and 7 in normal individuals (carriers of a euchromatic variant typically have 9-12 copies). The median copy number of this locus is 4 (Hollox *et al*, 2003). Genes in this region code for small cationic proteins which play an important role in the immune response to infection (Ganz, 1999). Extremes of copy number at this locus are known to be pathogenic; for example a low copy number is associated with Crohn's disease (Fellermann *et al*, 2006), whereas an increased copy number is associated with susceptibility to psoriasis (Hollox *et al*, 2008). Copy number variation of genes in this region, such as *DEFB4*, has been shown to correlate with gene expression levels, suggesting that efficiency of the immune response may vary according to copy number of the defensin genes (Hollox *et al*, 2003).

CNVs with mildly pathogenic phenotypes may also confer positive effects, resulting in them being maintained in the population. For example, variation in the α -globin genes is associated with α -thalassaemia (Kan *et al*, 1976). However, decreased α -globin copy number has a protective effect against malaria, and therefore low copy number at this locus is more common in parts of Africa where malaria is rife.

Compared to SNPs, which are usually associated with large haplotype blocks, CNVs often provide more direct clues to their functional impact as they often contain genes which may infer a possible phenotypic effect of the variant. However this can also be a disadvantage; there is a risk that a potential candidate gene selected purely because of a putative biological link with the disease of interest may be a false lead.

It is likely that there is a role for CNV in many forms of cancer due to the fact that there is a considerable level of CNV in putative oncogenes and tumour suppressor genes (Wong *et al*, 2007). Despite this, few CNVs have as yet been shown to directly affect susceptibility to cancer. Those that have been described include variation at chromosome 1q21.1, which is associated with susceptibility to the paediatric cancer neuroblastoma (Diskin *et al*, 2009). It is likely that detection of CNVs which predispose to cancer, and the way in which CNV in tumours effects disease progression, will be a major area of research in the coming years.

The role of a CNV in disease susceptibility can vary between populations. For example, a region of CNV on chromosome 8q24 which is associated with prostate cancer contributes to a greater proportion of population risk in African Americans than Caucasians, due to the fact that this variant occurs at a higher frequency in this population (Haiman *et al*, 2007).

1.3.3 ‘Missing Heritability’

After the initial flurry of identification of disease-associated marker SNPs, it became increasingly obvious that loci identified by GWAS would not be able to explain the genetic basis of all common disease. The causes of many complex diseases, for example, still remain elusive. In cases where variants which modify susceptibility to complex disorders have been identified, these tend to contribute a relatively small effect towards overall heritability (Reviewed by McCarthy, 2009). Therefore a considerable percentage of heritability for many diseases remains unaccounted for. Several possible explanations for this missing heritability have been proposed (Maher, 2008), one of which is the involvement of copy number variation. A role for CNV has been implicated in susceptibility for a number of complex diseases, however it is not yet

known what proportion of heritability for such disorders will be attributed to this form of variation. It is likely that a significant amount of the genetic risk will still remain unaccounted for. This remaining heritability may be explained by factors such as epigenetics, a term which describes inherited changes in gene expression or phenotype which occur as a result of mechanisms other than changes in the underlying DNA sequence, for example sequence modifications such as methylation.

1.3.4 Genome-wide CNV Studies

Recently there has been a move towards using large-scale GWAS to investigate the involvement of CNVs in complex disease. One such investigation was carried out by the WTCCC, who used an Agilent CGH array platform to detect associations between common CNVs and eight common complex disorders, including breast cancer, rheumatoid arthritis and cardiovascular disease (WTCCC, 2010). The frequency of variants in approximately 2000 case DNA samples for each of the eight diseases was compared to a common cohort of 3000 control DNA samples. The results of this study were somewhat disappointing, as only three loci were confirmed, each of which had previously been detected using SNP studies. This led the Consortium to conclude that common CNVs are not responsible for a significant proportion of risk for complex disease, and therefore do not explain much of the missing heritability as had previously been suggested. However, the relatively small sample sizes used in this study mean that there would have been low power to detect associations with rare CNVs. Additionally, CNVs which were difficult to genotype on the array platform employed, such as those within complex genomic regions, were excluded from this investigation. A number of regions of CNV which have previously been associated with disease are known to occur within complex regions of the genome. For example, variation within the complex β -

defensin region is known to play a role in susceptibility to Crohn's disease (Fellerman *et al*, 2000), however this association was not detected in the WTCCC study (WTCCC, 2010). Therefore it is possible that the WTCCC investigation may have failed to detect many associations. Findings from other genome-wide studies suggest that rare variants play an important role in complex disorders, for example a recent genome-wide investigation into autism spectrum disorders revealed an increased frequency of rare CNVs in affected individuals (Pinto *et al*, 2010). These findings are similar to those reported previously for schizophrenia, which also suggested that the effect of multiple rare structural variants is an important factor in disease susceptibility (The International Schizophrenia Consortium, 2008).

It is likely that the number of variants associated with common disease will increase rapidly over the next few years. Along with other studies, for example the 1000 genomes project, which aims to map SNPs and CNVs in 1000 phenotypically normal individuals from a number of different ethnic backgrounds (www.1000genomes.org), this will significantly increase the amount of CNV information available in the public domain. The involvement of individuals from a variety of ancestral backgrounds will aid comparison of genetic variation both within and across populations.

To date the majority of CNV studies have focused on the effect of variation on gene expression and the mechanisms by which this may influence susceptibility to disease. It is likely that in the future there will be an increasing emphasis on the global effects of changes in copy number, for example how they may affect biological pathways. Such studies will provide a greater understanding of the phenotypic and physiological effects of copy number variation, and aid advances in diagnosis and treatment of complex disease.

1.4 Methods of Studying Copy Number Variation

Since CNVs often lie in complex regions of the genome, studying this form of variation poses great challenges. For example many CNVs are surrounded by segmental duplications, or located in regions rich in repetitive elements (reviewed by Eichler, 2001). This can make designing suitable assays difficult, and consequently common SNP chips and arrayCGH platforms are often not well suited to study complex regions of CNV. Also it is very difficult, especially when using large-scale methods, to precisely determine where the boundaries of a CNV lie. It has been suggested that many studies have overestimated the size of CNVs, which may also lead to difficulties in distinguishing two variants which lie close together (Perry *et al*, 2008). For these reasons there is some uncertainty surrounding current estimates as to what proportion of the genome is copy number variable. Accuracy of CNV detection is currently considerably less than for SNPs, since compared to qualitative SNP genotyping, the quantitative assaying of copy number is far more difficult. Whilst it may be possible to distinguish between 1 and 2 copies (a doubling), it becomes increasingly more difficult to distinguish copy number as numbers increase.

Due to the many difficulties associated with the study of CNV, there is great demand for robust and reliable techniques for studying this form of variation. Some of the most common are described below.

1.4.1 Methods Employed for Early Studies of CNV

The earliest studies of CNV were carried out using microscopy, which allowed visualisation of variations involving large stretches of sequence. As technologies improved, methods such as Fluorescent *in situ* hybridisation (FISH) enabled detection

of sub-microscopic variants. FISH uses fluorescently labelled DNA probes which bind specifically to target DNA sequences, for example on a metaphase chromosome spread. Using a fluorescent microscope it is possible to detect where the fluorescent probes have hybridised (and therefore where the sequence of interest is localised). The number of copies of the region can be determined by the number of locations to which the probe has bound. Whilst this method is useful for visualising large copy number variations (Qiao *et al*, 2007), it is of limited use when looking at smaller regions and inversions. A development of this technique is FIBRE FISH (Heng *et al*, 1992) in which DNA probes are hybridised to chromosomes or DNA which has been mechanically stretched on a glass slide. This provides a significantly higher resolution, often down to a few kilobase pairs, which enables much more specific localisation of the probe binding point and accurate breakpoint detection (Florijn *et al*, 1995). Other forms of FISH have also been described including Q-FISH, which is often employed to assess telomere length (Slijepcevic, 2001), and CO-FISH (Chromosome Orientation-FISH) which is able to uniquely label sister chromatids (Meyne & Goodwin, 1994).

Southern Blotting (Southern, 1975) is used to detect the presence of a specific DNA sequence within a sample, and can also be used to determine copy number. To carry out this technique, genomic DNA is digested with endonucleases to cut the molecules into smaller fragments, which are then separated according to their size using agarose gel electrophoresis. These fragments are transferred from the agarose gel to a nitrocellulose or nylon membrane, using a process of capillary action, and fixed onto the membrane using heat or UV light. The presence of a sequence of interest is detected using fluorescently or radioactively labelled DNA probes which hybridise to their specific targets on the membrane. This method can be used to identify copy number variation by observing the number of fragments to which the labelled probes bind. In the case of a

single copy sequence the probes would only be expected to bind to one fragment, whereas if the sequence has been duplicated the probes will bind to multiple fragments (assuming the presence of internal cut sites).

Size differences resulting from copy number variation in high molecular weight restriction fragments can be visualised using Pulsed-Field Gel Electrophoresis (PFGE) (Schwartz & Cantor, 1984). Unlike conventional agarose gel analysis, PFGE enables the effective separation of large DNA molecules, up to around 2000 kb. In this technique an electric field is pulsed through the gel, and the direction of the current is altered at regular intervals. The larger the size of the DNA fragments, the slower they will be to adjust to this change in direction. Over a period of time, with the current constantly alternating, this will lead to increased separation between even very larger molecules. Although this is an effective method of separating molecules, its use is limited by the fact that each lane of the gel requires an extremely high concentration of DNA (10-20 µg).

1.4.2 PCR-based Methods

PCR-based methods are often used to determine copy number of a specific variant, or to characterise previously identified structural rearrangements in more detail. One common technique is Multiplex Amplifiable Probe Hybridisation (MAPH), in which short probes for regions of interest are hybridised to genomic DNA immobilised on a nylon filter (Armour *et al*, 2000). Probes are prepared by cloning sequence into a vector, so that each probe will share the same two flanking regions. Rather than detecting the hybridisations directly, as in the case of FISH or Southern blotting, the hybridised probes are recovered and amplified using a single pair of primers specific to the flanking regions. Since every probe is designed to be a different length, they can be

separated on an agarose gel. The amount of each probe is then quantified to determine relative copy number. Since an excess of probe is used, the amount of probe recovered will reflect the copy number of the sequence in the target DNA sample. This technique has been shown to be high-throughput and reproducible, with a resolution of 100-300 bp (Reid *et al*, 2003), and has been useful in determining copy number at α -defensin loci such as *DEFA1* and *DEFA3* (Aldred *et al*, 2005). However, this method does have significant limitations in that it can be labour intensive and requires large quantities of genomic DNA.

Another technique used to reveal copy number variation is Multiplex Ligation-dependent Probe Amplification (MLPA) (Schouten *et al*, 2002). Like MAPH, this method uses a single primer pair to amplify multiple targets. However, in the case of MLPA, each probe is made up of two oligonucleotides which hybridise to adjacent sequences on the target DNA. Only in a situation where both parts of the probe have hybridised to the correct sequence are they able to ligate and form the complete probe. Subsequent amplification of the target sequence occurs, and the products are separated on an agarose gel and quantified. This technique is highly specific due to the fact that the two parts of the probe will only join, permitting amplification, if they are hybridised to sequences close together. However, the preparation of the probes is an expensive and time-consuming process.

Both MAPH and MLPA are labour intensive and require considerable amounts of input DNA. However, there are a number of PCR-based methods available for the study of CNVs which are rapid and require relatively little DNA. Quantitative PCR (qPCR, also known as real time PCR) determines the relative amount of specific DNA sequences by monitoring the levels of fluorescent product produced during the course of a PCR reaction. This technique is used to amplify and simultaneously quantify a region of

DNA, and provides a rapid alternative to methods which are time consuming or require large quantities of input DNA (De Preter *et al*, 2002). qPCR is used to detect copy number variation by comparing the amplification of a single copy reference sequence to the region of interest. Differences in the rate of amplification indicate that the test region is copy number variable. qPCR is often used to confirm the presence of variations detected using other methods such as arrayCGH and FISH. For example, this technique was used to confirm the discovery of several large CNVs (Qiao *et al*, 2007). However, a limitation of this technique is that it assumes the PCR reactions being compared (test and reference) are equivalent in terms of variables such as primer efficiency, reaction conditions and rate of amplification, which may not always be the case.

The Parologue Ratio Test (PRT) goes some way to overcoming the problems associated with differing reaction variables by using a single pair of primers to compare the amplification of two diverged paralogous sequences. In a conventional PRT, one amplicon is a region of known copy number, which acts as a reference to verify the copy number of the second locus. The two loci are amplified simultaneously in the same reaction, and the amount of product produced from each locus is quantified. The ratio of product from each locus is compared and used to determine copy number (Armour *et al*, 2007). This is an extremely useful technique and is suitable for high-throughput, unlike many techniques available for the study of CNVs. Another advantage is that it requires a relatively low amount of input DNA compared to other techniques, typically around 10-20 ng. However, the success of this technique depends on being able to design primers which amplify both the sequence of interest as well as a paralogous reference region.

Another locus-specific method used to study CNV is Molecular Copy-number Counting (MCC). This technique uses DNA which has been highly diluted so that most wells on a PCR plate will contain just one copy of the genome. Specific sequences are then amplified, and the number of wells which are positive for the sequence of interest reflects the copy number of that sequence in the genome. This technique can be useful in identifying rearrangement breakpoints, for example Daser *et al* used MCC to locate the junctions of a recurrent translocation seen in renal cell carcinoma (Daser *et al*, 2006). The major disadvantage of this method is that it is very sensitive to contamination.

1.4.3 Comparative Genomic Hybridisation

In recent years there has been a move towards studying CNV on a genome-wide scale, which has been enabled by the development of techniques to support this approach. One of the most powerful techniques available for this purpose is Comparative Genomic Hybridisation (CGH). Initially described in the early 90s (Kallioniemi *et al*, 1992) this procedure uses genomic DNA from two samples (one test and one reference) which are labelled with different fluorochromes and co-hybridised to a metaphase chromosome spread. This is then visualised using fluorescent microscopy and the intensities of each fluorochrome determined. Any change in copy number between the two samples is revealed as a change in relative intensity between the fluorochromes. The initial protocol for CGH had the disadvantage of a low resolution and was not suitable for detecting small CNVs, balanced translocations or inversions.

A few years after CGH was first described, a further development followed, termed arrayCGH, which overcame some of the limitations of the initial method. Rather than metaphase chromosomes, total genomic DNA is hybridised to a series of probes spotted

uniformly onto an array. These probes are typically long regions of human genomic sequence which have been isolated from BACs (bacterial artificial chromosomes), although other sources may also be used, for example cosmids or cDNA clones. To this array, two differentially fluorescently labelled target DNA samples (test and reference) are co-hybridised. For each probe, hybridisation from each target DNA sample is calculated by measuring the relative fluorescence, allowing the detection of regions of CNV (Pinkel *et al*, 1998) (Figure 1.3).

Recently, a further advance of CGH has been described, termed oligo-array CGH (oaCGH), in which the cloned fragments are replaced by long oligonucleotides. These are either spotted onto a glass slide or synthesised *in situ* (Carvalho *et al*, 2004). Oligonucleotide arrays have the advantage of allowing a higher resolution than BAC arrays. Commercial platforms provided by companies such as Affymetrix and Nimblegen are capable of producing many hundreds of thousands of features per array, which enables an extremely high throughput system generating a huge amount of data. However, there is currently a paucity of cheap, reliable and effective in-house platforms for arrayCGH, which is a limiting factor in the study of CNV and is holding back progress in this field.

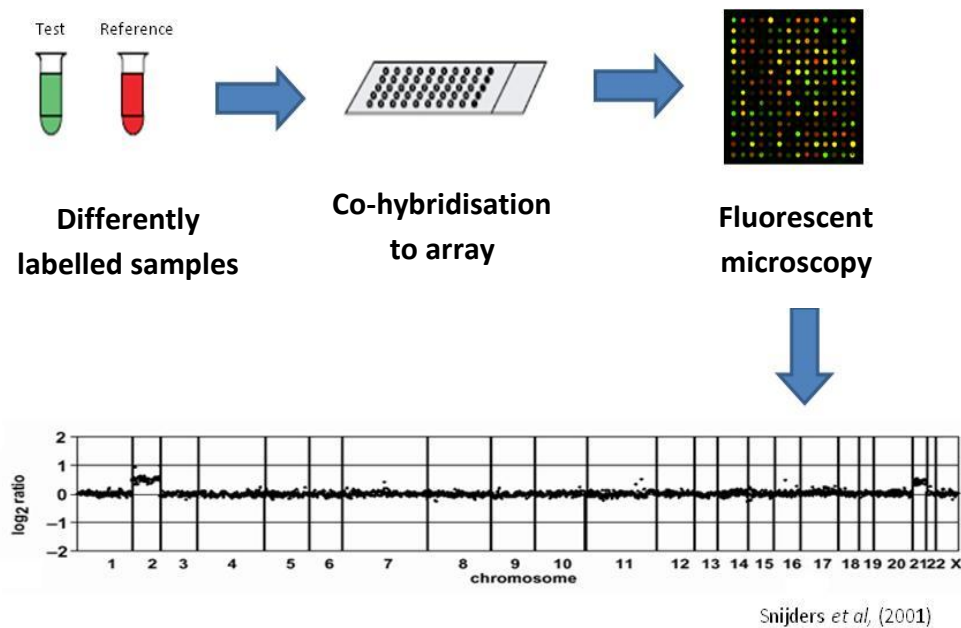


Figure 1.3: BAC Array CGH

Array Comparative Genomic Hybridisation is a powerful method used to detect the presence of CNV. Two samples, one test and one reference, are fluorescently labelled with different fluorochromes and co-hybridised to an array of BAC clones. The array is visualised using fluorescent microscopy and a change in copy number between regions of the test and reference genomes is revealed as a change in the ratio of signal intensities between the two fluorochromes. Array CGH data is typically visualised on a \log_2 scatter plot, the graph shown here is taken from Snijders *et al*, 2001 and shows a trisomy of chromosomes 2 and 21.

1.4.4 Sequence-based Approaches

Determining the nucleotide sequence of the whole or part of a genome is a useful method for establishing its structure and identifying variants. Over recent years it has become possible to sequence entire genomes at high throughput and at an increasingly lower cost. This has led to a rapid increase in the number of sequenced genomes which are available, from a variety of ethnic backgrounds including Caucasian, African and Korean individuals (Levy *et al*, 2007; Wheeler *et al*, 2008; Wang *et al*, 2008; Bentley *et al*, 2008). This increase in the number of available reference genomes, as well as advancements in sequencing technologies, means that sequencing-based methods are becoming more popular as a means to study copy number variation. However, these

genomes have been aligned to a reference sequence rather than assembled *de novo*, which means that they will reflect any errors present in the reference, and are not useful for identifying new copy number variants. Also, cost limitations mean that whole-genome sequencing is not a practical method for studying copy number in a large number of individuals (Bentley, 2006). A more common approach is to resequence selected regions, for example candidate genes, data from which is aligned to a reference sequence to allow direct comparison of the results from a number of samples.

Paired-end mapping is another sequencing-based approach which is useful for revealing the presence of structural variation. Genomic DNA is fragmented into segments of around 3 kb in size, the ends of which are sequenced and mapped to a reference genome (Korbel *et al*, 2007). If the relative orientation of the two end reads has altered or if the distance between them on the reference sequence after mapping is not approximately equal to the size of the fragment, then this can indicate the presence of structural rearrangements, for example deletions, duplications and insertions. This technique is especially useful for detecting inversions, which may be missed by other methods since they do not usually result in a dosage change. Paired-end mapping was recently used to identify structural variations present in eight HapMap individuals (Kidd *et al*, 2008). Advantages of this technique are that it does not require full sequencing of genomes, and the development of next-generation sequencing technologies, such as that from 454 Life Sciences (Margulies *et al*, 2005), makes it feasible to use this method to study variation in a large number of individuals at relatively low cost.

1.4.5 SNP-based Approaches

It is possible to reveal the presence of CNVs through the genotyping of markers such as SNPs in the region of interest. A normal diploid individual will carry two copies of a specific region of sequence, and so will either be homozygous (two copies of the same allele) or heterozygous (one copy of each allele) for each SNP. Therefore, the allele ratio can either be 2:0 or 1:1. However, in regions which are copy number variable, a much wider variety of ratios are possible. For example, ratios such as 2:1, 1:2 and 3:1 are possible in regions of CNV, depending on the degree of variability in the region of interest. Techniques such as Dynamic Allele-Specific Hybridisation (DASH) can be used to score alleles of SNPs (Howell *et al*, 1999). This method works by genotyping variants through the study of melting curves, on the principle that sequences which match exactly will denature at a higher temperature than those which contain mismatched bases. This method can be used to detect SNPs which deviate from the expected 2:0 or 1:1 allele ratios, termed Multisite Variants (MSVs), suggesting the presence of structural variation within the locus under study (Fredman *et al*, 2004).

Arrays have been developed which make it possible to simultaneously genotype regions of CNV as well as SNPs. For example, the Affymetrix SNP 6.0 array includes probes for the detection of CNV as well as the traditional probes to assay SNPs (McCarroll *et al*, 2008). The CNV probes are specifically targeted to known regions of variation. This approach is a combination of SNP genotyping and arrayCGH in a single hybrid array, although due to limits in the number of features that can be included on each array, the coverage is compromised to some extent.

Bearing in mind the difficulties associated with identifying and studying regions of CNV, one of the aims of the research for this thesis was to develop oaCGH on an in-

house customisable microarray facility which had not previously been used for oaCGH. This was to enable the study of a putative region of copy number variation on chromosome 12, for which involvement in RA susceptibility has been suggested. The rest of this chapter will describe RA, current knowledge of the genetic basis of this disorder, and evidence for a CNV on chromosome 12 putatively associated with the disease.

1.5 Rheumatoid Arthritis

Rheumatoid Arthritis (RA) is a chronic inflammatory autoimmune disease which affects approximately 0.8% of the UK adult population, and is around three times more common in women compared to men (Symmons *et al*, 2002). It has been estimated that the UK economic burden relating to RA is around £8 billion per year (National Rheumatoid Arthritis Society, 2010).

RA is characterised by the generation of an autoimmune response against cells of the synovium, a thin membrane found in synovial (freely moving) joints which lines the joint capsule and secretes synovial fluid. This leads to an infiltration of inflammatory cells into the synovium, causing swelling and progressive destruction of the bones and cartilage within the affected joint (Figure 1.4).

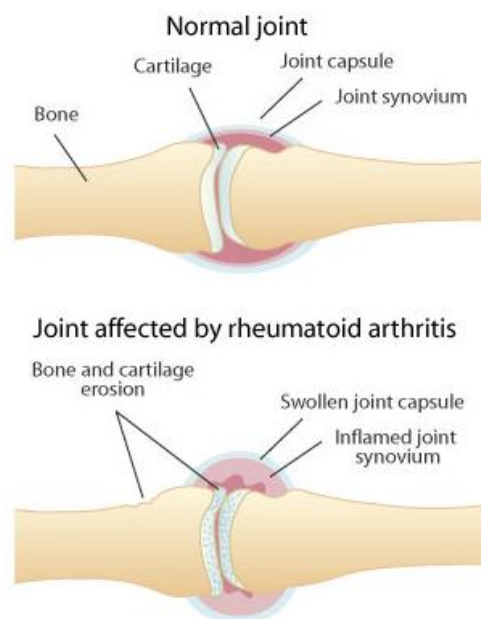


Figure 1.4: Effects of Rheumatoid Arthritis

A normal joint is shown, compared to one from a rheumatoid arthritis patient. Characteristic features of this autoimmune disease which can be seen in this figure include inflammation of the joint synovium and cartilage erosion (Medical Look, n.d).

Characteristic symptoms of RA include pain, stiffness in joints and loss of mobility (Arnett *et al*, 1988). Advanced cases of RA may show systemic inflammation, and in 15-25% of cases extra-articular manifestations of RA involving tissues such as blood vessels, heart and lungs are apparent (Turesson *et al*, 2003). RA patients show an increased risk of heart disease, which is a major cause of their mortality (Turesson *et al*, 1999; Gabriel, 2010). There is also evidence to suggest that cases of lymphoma are increased in patients with severe RA (Zintzaras *et al*, 2005). There is currently no cure for this debilitating disease, although anti-inflammatory drugs such as steroids can be used to suppress disease symptoms. Biological agents which target molecules thought to be involved in RA pathogenesis may also be used for this purpose, for example TNF- α blockers have been shown to slow disease progression and joint destruction (Listing *et al*, 2008).

RA is the most prevalent autoimmune disorder, others of which include the inflammatory skin disease psoriasis, type 1 diabetes, the central nervous system disorder multiple sclerosis, and systemic lupus erythematosus which affects connective tissue.

1.5.1 Environmental Factors Contributing to RA Susceptibility

The exact causes of RA are poorly defined, although it is known that both environmental and genetic factors play a role in disease development. Studies of disease concordance in monozygotic and dizygotic twins have placed the heritability of this disorder at around 60% (MacGregor *et al*, 2000). The remaining 40% of disease risk must therefore be due to environmental factors. Many environmental agents have been suggested as contributing factors to RA susceptibility, the most well studied of which is cigarette smoking. A study which compared the smoking history of twins with RA and

their unaffected co-twins detected a strong association between smoking and development of RA in both monozygotic (OR=12.0 95% CI 1.78-513) and dizygotic twins (OR=2.5 95% 0.92-7.87) (Silman *et al*, 1996). This effect has been confirmed by a number of independent studies (Morgan *et al*, 2009; Symmons *et al*, 1997), which agree on a more modest effect size of around 1.7 (95% CI 1.3-2.2 and 0.95-3.06 respectively). One of these studies, which examined a range of potential risk factors for RA, also identified strong associations with obesity (OR= 3.74; 95% CI 1.14-12.27) and a history of blood transfusion (OR=4.83; 95% CI 1.29-18.07) (Symmons *et al*, 1997). A role for hormones, particularly oestrogen and progesterone, has been suggested due to the observation that women are three times more likely to develop RA, particularly after periods of hormonal changes such as post-partum (Silman *et al*, 1992) and during the menopause (Goemaere *et al*, 1990).

1.5.2 Auto-antibodies in RA

Rheumatoid arthritis is a heterogeneous disease, in that it does not manifest itself in the same way in every patient. For example, one of the common features of autoimmune diseases is the presence of auto-antibodies in blood serum. There is evidence to suggest that there may be several distinct subtypes of RA, characterised by the different auto-antibodies produced. One autoantibody commonly associated with RA is Rheumatoid Factor (RF), which was first described in 1940 (Waalder, 1940) and later found to be specific to the antibody Immunoglobulin G (IgG) (Reviewed in Dorner *et al*, 2004). The presence of this autoantibody is often used as a diagnostic marker for RA (Arnett *et al*, 1988), with cases being categorised as being either RF positive (RF+) or RF negative (RF-). However, it has been shown that the presence of this antibody is not specific to RA but may also be a feature of other autoimmune diseases including SLE and Sjögren

syndrome, as well as non-autoimmune conditions including chronic infections (Renaudineau *et al*, 2005).

As well as RF, autoantibodies which are highly specific to RA have also been described, which are collectively termed anti-cyclic citrullinated peptide (anti-CCP) antibodies (Schellekens *et al*, 1998). These are estimated to be present in around 80% of RA patients (reviewed in Zendman *et al*, 2006). It has been shown that these antibodies can be present in the blood many years before onset of disease symptoms (earlier than RF) and may be correlated with disease severity (Rantapaa-Dahlqvist *et al*, 2003; Nielen *et al*, 2004; Kastbom *et al*, 2004).

There is currently no conclusive evidence as to whether environmental factors such as smoking are associated with the development of particular subtypes of RA. One study reported that anti-CCP-positive RA is associated with exposure to cigarette smoke (Klareskog *et al*, 2006a) whereas another shows evidence to suggest that smoking is associated with the generation of both CCP and RF specific antibodies ($p=0.0001$, OR 1.7; 95% CI 1.3-2.2) but failed to find association with the two types of auto-antibody individually (Morgan *et al*, 2009). There is evidence to suggest that smoking may contribute towards the citrullination of proteins in target tissues such as the lungs, therefore acting as a trigger to anti-CCP autoimmunity (Klareskog *et al*, 2006a).

1.5.3 RA Pathogenesis

The exact pathogenesis of RA is currently unclear. Autoantibodies such as RF and anti-CCP have been detected in the blood of individuals several years before the onset of disease symptoms (Rantapaa-Dahlqvist *et al*, 2003), which suggests that either the

development of RA is a delayed effect, or that the production of these molecules alone is not sufficient to trigger disease development.

Early symptoms of RA include joint swelling and angiogenesis, followed by rapid proliferation of cells within the synovial lining as well as infiltration of immune system cells including T-cells, B-cells and macrophages into the synovium (reviewed in Weyand *et al*, 2000). Later stages of the disease involve remodelling of the synovium and formation of a mass of tissue called ‘pannus’. This serves as the origin for joint destruction and produces enzymes such as matrix metalloproteinases, which are involved in the destruction of the extracellular matrix (ECM) (discussed by Goronzy & Weyand, 2009).

Research has suggested a role for various cells and molecules of the immune system in the initiation of the immune response and production of autoantibodies. There is evidence to suggest the involvement of immune system cells such as B and T lymphocytes in RA pathogenesis (Reviewed in De Keyser *et al*, 1995; Weyand *et al*, 2000; Martinez-Gamboa *et al*, 2006). Also implicated are cytokines, molecules which act as chemical messengers between cells and are involved in many biological processes, including the immune response. It has been suggested that RA progression may be linked to excess production of pro-inflammatory cytokines, in particular tumour necrosis factor α (TNF- α) (Feldmann, 1996).

1.5.3.1 *Chondrocytes*

Joint cartilage is an inflexible connective tissue which cushions the joints, preventing the bones from rubbing against each other and causing joint pain. Cartilage is populated by a small number of cells known as chondrocytes, which are important in the

maintenance of homeostasis in the joints. Under normal conditions they maintain a stable equilibrium between the synthesis and degradation of components of the extracellular matrix (ECM), including several types of collagen (Reviewed by Goldring & Marcu, 2009).

Chondrocytes vary in size, shape and metabolic activity depending on their exact location within the cartilage. Since there is a limited blood supply to the cartilage, nutrients are supplied by diffusion. Chondrocytes have therefore adapted to survive in an environment which is deficient in the provision of oxygen and other nutrients. This is reflected in features of chondrocyte metabolism, which operates at an unusually low oxygen tension, as low as <1% (Silver, 1975), and has a particularly high glucose requirement. The cells are also well equipped to respond to environmental changes in order to maintain homeostasis in the cartilage. For example, they are able to modulate the intracellular expression of survival factors such as HIF-1 α which can stimulate the expression of glucose transporters (Mobasheri *et al*, 2008).

Chondrocyte metabolism is affected by factors such as stress, which can lead to the breakdown of cartilage. Since the pathogenesis of RA is known to involve destruction of the ECM, a process in which chondrocytes play a crucial role, it has been hypothesised that these cells may be involved in disease progression. This is supported by evidence of chondrocyte apoptosis in joint destruction (Yatsugi *et al*, 2000). It has been proposed that chondrocytes themselves may also be a source of pro-inflammatory cytokines, which aid joint destruction by increasing the breakdown of tissue and suppressing repair mechanisms. As a result, cartilage is degraded faster than it can be repaired, leading to destruction of the joint (reviewed by Otero & Goldring, 2007). Improved understanding of chondrocyte physiology and the role that these cells play in

joint disorders may provide important insights into the pathology of rheumatoid arthritis.

1.6 Rheumatoid Arthritis Genetics

Rheumatoid arthritis is a multifactorial disease which develops as a result of both environmental and genetic causes. The genetic component has been identified through twin and family studies, and it is estimated as much as 60% of the risk for RA may be due to genetic factors (Macgregor *et al*, 2000). However, the role of genetics in rheumatoid arthritis is far from simple. For many years the only genes which were known to confer susceptibility to RA were HLA alleles located within the Major Histocompatibility Complex (MHC) on chromosome 6 (Stastny, 1976). Recent GWAS experiments have led to putative associations with several non-HLA loci, however only a handful of these have been replicated, and in the majority of cases, the actual causal variant remains to be identified. Many RA susceptibility loci have also been implicated in other autoimmune diseases including SLE, multiple sclerosis and inflammatory bowel disease (Jawaheer *et al*, 2001). There is no doubt that many more putative risk loci will be discovered over the next few years.

1.6.1 Identification of HLA Genes as the Major RA Susceptibility Locus

In the 1970s it was shown that HLA genes within the MHC play a role in RA susceptibility (Stastny, 1976). This discovery was later built upon by Gregersen *et al*, who identified causal genes within the MHC HLA loci, specifically *HLA-DRB1*. Alleles of *HLA-DRB1* which contribute to RA susceptibility share a highly conserved amino acid sequence at a certain position (residues 67-74), which has been termed the “shared epitope” (Gregersen *et al*, 1987). It has been estimated that alleles at this locus may contribute to around 40% of the total genetic risk for RA (Deighton *et al*, 1989). Despite the discovery of several new RA association loci as a result of the recent explosion of GWAS, this still remains by far the most significant association detected. *HLA-DRB1*

association appears to be restricted to the anti-CCP positive subcategory of the disease (Huizinga *et al*, 2005; Ding *et al*, 2009). Studies have suggested that proteins encoded by the HLA shared epitope alleles may bind to citrullinated peptides with a higher affinity, which leads to an enhanced T-helper cell response (Hill *et al*, 2003).

1.6.2 RA Linkage Studies

For many years, the HLA genes were the only loci confirmed to show association with RA susceptibility. In order to identify non-HLA regions which contributed to RA risk, several approaches were taken, including linkage studies of RA families and investigations involving rat models of the disease (reviewed by Worthington, 2005).

A number of genome-wide linkage screens for RA were carried out using multi-case families with affected sibling pairs. Early studies of this kind used relatively small numbers of families. For example, a European study investigated 97 families (Cornelis *et al*, 1998) whereas an early Japanese study used 41 (Shiozawa *et al*, 1998). These studies confirmed the HLA region as having the strongest linkage with RA, and identified a number of other putative susceptibility regions; however the small number of samples used meant that they had low power to detect loci with a smaller contribution to RA susceptibility. Later studies of Caucasian populations which used much larger sample sizes of several hundreds of families (377 and 512 affected families, respectively) (MacKay *et al*, 2002; Jawaheer *et al*, 2003) repeated these observations. Data from these linkage studies supported the hypothesis that although genes within the HLA region show the strongest linkage to RA, there are also a number of other susceptibility loci which each contribute smaller effects to risk of developing the disease. The four studies described above detected evidence for linkage in wide

range of markers, although not all of these were replicated in multiple investigations. However, a number of regions were identified as putative susceptibility loci by more than one study, including those on chromosomes 1q and 14q. It is likely that failure to replicate many of the findings is due to factors such as false positive results in the initial studies, and limitations of sample size. A number of meta-analyses (in which the results of several studies are combined prior to analysis) have been carried out using data from linkage studies in an attempt to overcome this problem. Such studies have identified strong evidence for linkage at a number of loci, including chromosomal regions 6q and 16p (Etzel *et al*, 2006; John *et al*, 2006).

Rat models have also been employed to investigate RA susceptibility loci. These provide a useful model for investigating RA pathogenesis since a number of rat strains are sensitive to the induction of an inflammatory arthritis using agents including collagen and oil (reviewed in Holmdahl *et al*, 2001). The resulting condition shares many features with the human disorder, for example MHC association, involvement of T-cells and chronic erosion of bones and joints. Studies of rat models and the identification of putative susceptibility loci which have orthologous regions in the human genome provides a good starting point for further investigation and candidate gene studies. For example, genes within the quantitative trait locus *Oia2*, located on rat chromosome 4, are associated with susceptibility to oil-induced arthritis in rats (Jansson *et al*, 1999; Lorentzen *et al*, 1998). This discovery has led to the identification of putative candidate genes in the corresponding region on human chromosome 12 (discussed in detail in section 1.7).

Linkage studies were able to identify a number of putative RA susceptibility loci, however consistent replication was a problem due to the low power of the studies, and the rarity of many of the variants in question. More recently, the ability to genotype

increasingly large numbers of markers simultaneously using microarrays has led to a rapid increase in genome-wide association studies. These have now, to a large extent, taken over from linkage analysis studies as a means to identify disease susceptibility loci.

1.6.3 *PTPN22*: The Second RA Susceptibility Locus

PTPN22 was first identified as a susceptibility locus for Type I Diabetes using association studies (Bottini *et al*, 2004). The same variant was then shown to be associated with RA in a Caucasian population, becoming the second gene for which association with RA susceptibility was confirmed (Begovich *et al*, 2004). A SNP within *PTPN22* (allele 1858T) leads to an amino acid substitution (arginine to tryptophan), which is associated with an increased susceptibility to RA. This effect has been replicated in a number of case-control studies and is therefore well established (for example Hinks *et al*, 2005; Michou *et al*, 2007). This polymorphism also provides evidence of population differences in susceptibility to RA, as the 1858T allele is not present in Asian populations (Kawasaki *et al*, 2006; Ikari *et al*, 2006).

PTPN22 is located on chromosome 1p13 and encodes a lymphoid-specific tyrosine phosphatase which has a potential role in regulating the thresholds for B and T cell activation (Rieck *et al*, 2007; Bottini *et al*, 2006). This makes it a good candidate gene for involvement in autoimmune disorders. This hypothesis is supported by the fact that, as well as RA, there are also proven associations between *PTPN22* and other autoimmune diseases including Type I diabetes and SLE in Caucasian populations (Reviewed in Gregersen *et al*, 2006). Whilst this is a well known association, it is estimated to contribute a much smaller effect (OR = 1.8 (Michou *et al*, 2007)) to RA susceptibility than the only other previously described association locus, *HLA-DRB1*.

Estimates have suggested that the 1858T polymorphism within *PTPN22* may contribute to around 8% of RA susceptibility (Lee *et al*, 2007b), compared to HLA which is thought to confer around 40% of the genetic risk (Deighton *et al*, 1989).

1.6.4 SNPs Associated with RA Identified Using GWAS

A landmark GWAS study published in 2007 investigated disease associations using a common panel of 3000 control samples and a total of 14000 case samples from 7 common complex diseases including RA. Affymetrix 500K GeneChips were used to genotype loci across the entire human genome and identify putative regions of association (Wellcome Trust Case Control Consortium, 2007). Excluding association in the HLA region, 10 SNPs were identified which showed putative association with RA. One of these polymorphisms maps to *PTPN22*, which has previously been established as an RA susceptibility locus (Begovich *et al*, 2004). Since the publication of the WTCCC study, association of a number of the other SNPs with RA has been replicated (Table 1.1).

Despite the success in identifying RA susceptibility loci using GWAS, this may not be a suitable approach for detecting all regions of association. For example, the number of susceptibility loci detected may vary depending on the number of markers; a study containing a relatively small number of markers may miss some variants. Also, complex regions of the genome may be poorly represented due to a lack of suitable probes. This is particularly likely to affect CNVs, which are often rich in repeat elements and located within or nearby regions containing segmental duplications, both of which may result in probes from these loci being excluded from a study.

Table 1.1: Established RA SNP Associations (correct as of January 2010)

SNP	Chromosome	Candidate Gene	Date	Reference
rs2476601	1p13.2	<i>PTPN22</i>	2004	Begovich <i>et al</i>
rs7574865	2q32.3	<i>STAT4</i>	2007	Remmers <i>et al</i>
rs3761847	9q33.1	<i>C5/TRAFF1</i>	2007	Plenge <i>et al</i>
rs6920220	6q23.3	<i>TNFAIP3</i>	2007	Thomson <i>et al</i>
rs3087243	2q33.2	<i>CTLA4</i>	2005	Lei <i>et al</i> ;
rs13031237	2p16	<i>REL</i>	2009	Gregersen <i>et al</i>
rs4810485	20q13	<i>CD40</i>	2009	Orozco <i>et al</i>
rs2240340	1p36.13	<i>PADI4</i>	2003	Suzuki <i>et al</i>
rs4750316	10p15.1	<i>PRKCQ</i>	2008	Barton <i>et al</i>
rs1678542	12q13.3	<i>KIF5A</i>	2008	Barton <i>et al</i>
rs2812378	9p13.3	<i>CCL21</i>	1998	Cornelis <i>et al</i>
rs6822844	4q27	<i>IL2/IL21</i>	2007	Zhernakova <i>et al</i>

1.6.5 CNV and autoimmune disease

Despite an increase in the number of association studies in recent years, leading to an improved understanding of causes of many diseases, the pathogenesis of autoimmune disorders remains poorly understood. However, results from twin and family studies, as well as human linkage and association studies, have suggested that many of the same common genetic factors underpin a range of different autoimmune diseases (Reviewed by Gregersen & Olsson, 2009).

CNV has been implicated in susceptibility to several autoimmune diseases including systemic lupus erythematosus (Yang *et al*, 2007) and psoriasis (Hollox *et al*, 2008).

However, currently there is not much known about the contribution of CNV to RA. There was a report of an association of RA with CNV of *CCL3L1*, a chemokine ligand which binds to several pro-inflammatory cytokine receptors. It has been suggested that increased copy number of this gene may have proinflammatory effects and therefore increase the risk of autoimmune diseases such as rheumatoid arthritis ($p = 0.009$, OR 1.34) and type 1 diabetes (McKinney *et al*, 2008). However, this effect was only seen in a disease cohort from New Zealand ($n=834$) and not in a smaller UK cohort ($n=302$). It had previously been shown that a decreased copy number of *CCL3L1* is associated with increased risk to HIV (Gonzalez *et al*, 2005).

1.6.6 Involvement of Pathways

A number of the SNPs which have been associated with RA lie within the same biological pathways. Many of these are involved in the immune response and other processes important in RA pathogenesis. For example, genes such as *PTPN22*, *IL21* and *CTLA4* are all involved in T-cell inactivation or signalling pathways, and a number of these as well as others including *CD40* and *IL2* are associated with apoptosis and cell death (Hill *et al*, 2003; Ettinger *et al*, 2008; Slavik *et al*, 1999; Xu *et al*, 2004; Burchill *et al*, 2007). These observations provide clues as to pathways important in disease development and pathogenesis, and such discoveries may ultimately inform the development of treatments.

1.6.7 Interaction Between Genetic and Environmental Factors

It has been suggested that rheumatoid arthritis may develop as a result of environmental triggers acting on individuals with a genetic predisposition (Reviewed by Klareskog *et al*, 2006b). For example, individuals with a *PTPN22* polymorphism which confers

susceptibility to RA may be unable to destroy auto-reactive T-cells, thereby predisposing them to autoimmune diseases (Vang *et al*, 2005). This susceptibility, combined with environmental factors such as smoking or changes in hormones, may therefore lead to the development of RA.

1.6.8 Population Differences

There appears to be considerable genetic heterogeneity in RA susceptibility amongst different ancestral groups. Begovich *et al* showed that the *PTPN22* 1858T polymorphism occurs at a frequency of ~17% in Caucasians (n=1961), but the same study failed to detect the variant allele in smaller populations of Africans (n=21) or Han Chinese (n=100) (Begovich *et al*, 2004). The failure of this investigation to detect the causal allele in individuals of Asian ancestry could have been attributed to the small size of the cohorts studied; however, a subsequent investigations using larger Asian cohorts also failed to detect the 1858T polymorphism in this population (Kawasaki *et al*, 2006; Ikari *et al*, 2006). In Asian individuals, the second most important RA susceptibility locus (after the HLA genes) has been shown to be *PADI1* (Takata *et al*, 2008).

Additionally, a study of multiple RA risk loci in Caucasian and Korean populations showed that none of the loci identified in Caucasian populations showed significant association with the disease in Koreans (Lee *et al*, 2009). Therefore, although it is known that a number of loci universally confer RA susceptibility in all populations tested to date, including the well known association with alleles of the *HLA-DRB1* locus and also *STAT4* (Remmers *et al*, 2007; Lee *et al*, 2007c), it appears that many of the risk loci which show smaller effects may vary considerably between populations.

1.6.9 The Current State of RA Genetics

Although in recent years there has been a considerable increase in the number of new RA susceptibility loci identified, most of these contribute small effect sizes to overall heritability of the disease. A significant proportion of RA susceptibility therefore remains to be explained, some of which may be due to the involvement of CNV, rare variants, or epigenetic factors which have not yet been identified.

As well as identifying loci which contribute to a significant proportion of RA risk, continual identification of risk factors which contribute small effect sizes remains important. Although such variants may not contribute greatly to RA susceptibility, their location may highlight additional pathways which are important in disease pathogenesis and therefore advance knowledge of RA pathogenesis, as well as providing novel targets for therapies. This approach could make it possible to design a single drug to target a pathway containing a number of susceptibility loci, rather than attempting to design multiple drugs to target individual variants. It is likely that in the coming years there will also be further investigation into the contribution of environmental factors to RA development, and the ways in which these act in combination with genetic susceptibility loci to initiate and maintain a disease state, as well as closer examination of the pathways involved.

Populations with different ancestry are known to show variation in susceptibility loci for complex disorders. For example, a number of RA risk loci which are important in Caucasians did not show association with the disease in a Korean cohort (Lee *et al*, 2009). So far, the vast majority of association studies have been carried out using samples from individuals with Caucasian ancestry, which mostly limits current knowledge to risk loci which are important in this population. In order to obtain a

broader picture of disease susceptibility, it is therefore necessary to study individuals from a wider range of ancestral backgrounds. Such investigations will provide further insight into population differences in RA susceptibility, and may highlight new genes and pathways which are involved in disease pathogenesis.

1.7 A Putative RA Susceptibility Locus on Chromosome 12

A number of studies, including preliminary investigations within our research group, have suggested a putative association between chromosome 12p13.31 and RA. There is also evidence to suggest that this region may contain structural variation. Work which led to the identification of this locus as a putative susceptibility region is described in this section.

1.7.1 Identification of an RA Susceptibility locus on Rat chromosome 4q42

A rat model of oil induced arthritis (OIA) is commonly used in linkage studies to detect putative susceptibility regions which may also be involved in human RA. *Oia2*, the second quantitative trait loci for RA to be described in rats (aside from *Oia1* which contains MHC genes), is located on rat chromosome 4 (Jansson *et al*, 1999; Lorentzen *et al*, 1998). The region involved in RA susceptibility was later narrowed down to a 1.2 Mb interval at chromosome 4q42 using high resolution recombinant mapping (Ribbhammar *et al*, 2003). A gene complex termed *APLEC* (Antigen Presenting cell Lectin-like gene Complex) is present at this locus, which contains a number of lectin-like receptors. These are members of the *CLEC* super family and code for group II Calcium-dependent Lectins, which are glycoproteins that bind carbohydrates (Flornes *et al*, 2004). The *CLEC* proteins are expressed on the surface of many cells of the immune system, including neutrophils and antigen presenting cells, a fact which makes them good candidate genes for involvement in RA. Using rats which showed resistance to OIA, linkage of this trait was shown to be conferred by certain alleles of *APLEC* (Lorentzen *et al*, 2007).

As well as *Oia2*, there are a number of other quantitative trait loci (QTLs) which map to rat chromosome 4q42. These are associated with traits including blood pressure, serum cholesterol, insulin and glucose levels, suggesting a role for genes within this region in other complex disorders, such as diabetes and heart disease.

1.7.2 Human 12p13.31

The rat *APLEC* gene complex on chromosome 4q42 is orthologous with a gene cluster on human chromosome 12p13.31, which contains a number of the same genes including *CLEC4A* (also known as *DCIR*) and *CLEC4E* (also known as *MINCLE*) (Flornes *et al*, 2004). There are also a number of other genes present within both the rat QTL *Oia2* and the human 12p13.31 locus which are not part of the C-type lectin superfamily, including *SLC2A3*, a glucose transporter, and *NANOG*, which has a role in cell development.

Interestingly, linkage of the human 12p13 region with RA has previously been described. In one of the first major genome-wide screens of RA families, which aimed to identify susceptibility loci outside of the HLA region, 257 multi-case families (including 301 affected sibling pairs), were screened for shared alleles which were transmitted along with the disease (Jawaheer *et al*, 2001). This investigation identified a number of non-HLA susceptibility loci which showed linkage with RA, including one located across the 12p13.31 region ($p = 0.0051$).

As well as RA, there is evidence to suggest that the 12p13.31 locus may be involved in susceptibility to other complex diseases. In particular, the region surrounding the gene *CLEC2D* has been implicated in Type I Diabetes (The Wellcome Trust Case Control Consortium, 2007). There is also a QTL within this region which is associated with blood pressure.

Given this evidence, a study involving our group as well as a number of collaborators investigated the association of SNPs within the *APLEC* complex with RA (Lorentzen *et al*, 2007). Around 35 SNPs were genotyped using DASH, and the allele frequencies in RA cases and control samples compared. Analysis of data revealed a peak of association with anti-CCP negative RA, which localised around *CLEC4A* (*DCIR*) (OR=1.27; 95% CI 1.06-1.52) (Figure 1.5). The exact nature of the variation responsible for this association was not identified. It is possible that this result may reflect an association with variation found in genes of the *APLEC* cluster which are not members of the *CLEC* family, however due to their involvement with the immune system, *CLEC* genes make interesting candidate genes for involvement in autoimmune disease.

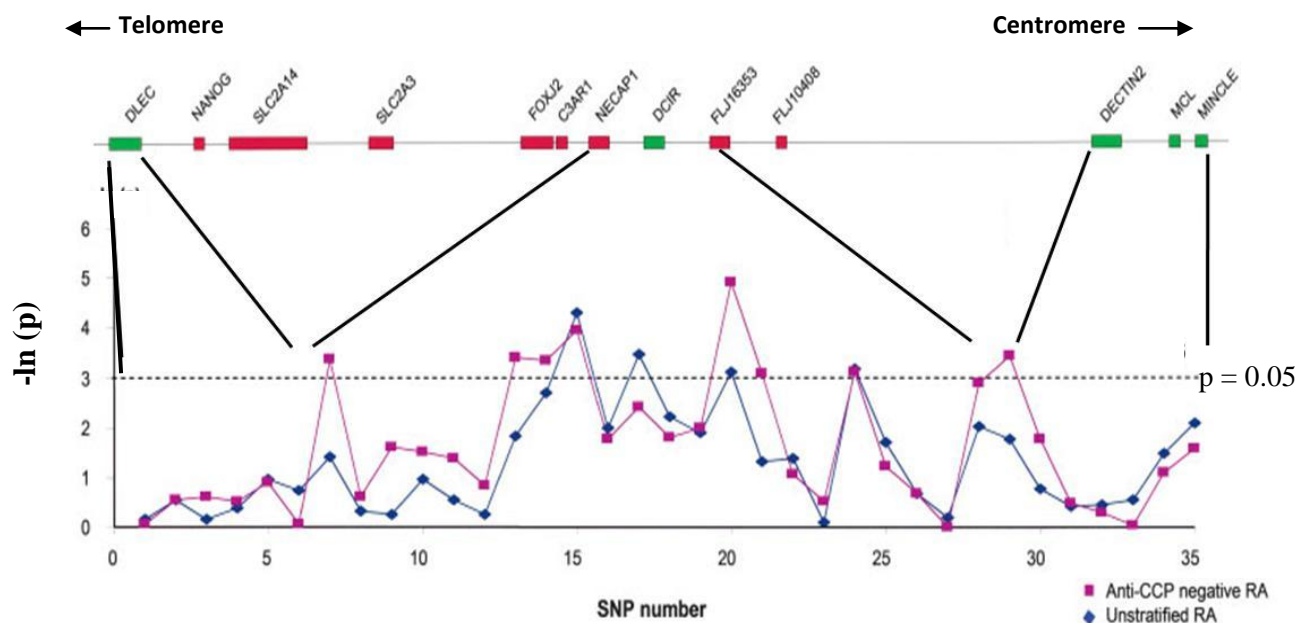


Figure 1.5: Association of SNPs on Chromosome 12p13.31 with Rheumatoid Arthritis

Lorentzen *et al* (2007) genotyped SNPs within a putative RA susceptibility locus on chromosome 12p13.31, and revealed a statistically significant association of SNPs surrounding the gene *CLEC4a* (*DCIR*) with the disease. Genes are represented as coloured bars; green bars indicate genes which are members of the C-type lectin receptor family, whereas red bars show other genes within this region. Black lines show intervals within which the genotyped SNPs are located. P-values for both anti-CCP negative RA and unstratified samples are shown as different coloured lines on the graph, as shown in the key.

Whilst SNP genotyping for this study was being carried out within our research group, it was observed that the DASH data for many of the SNPs in this region suggested the presence of structural variation. Rather than show allele ratios of either 2:0 or 1:1, as expected, a variety of different ratios were observed. Similar complex genotyping patterns, termed multisite variants (MSVs), have previously been described (Fredman *et al*, 2004). These occur due to the presence of structural variation, for example a segmental duplication, where the sequences of individual copies may vary resulting in different allele ratios being detected for the same SNP. Further studies using 200 markers and covering a larger area of 12p13.31 detected the presence of many MSVs located all along the region of interest. These results strongly suggest the presence of structural variation at this locus.

In order to further investigate this putative structural variation, arrayCGH experiments were carried out using the Nimblegen microarray platform. Several pairs of DNA samples were compared, including an RA patient verses control sample. Results highlighted a region of CNV located distal to *CLEC4A*, a gene which had previously been identified as a putative candidate gene for RA susceptibility (Lorentzen *et al*, 2007) (Figure 1.6).

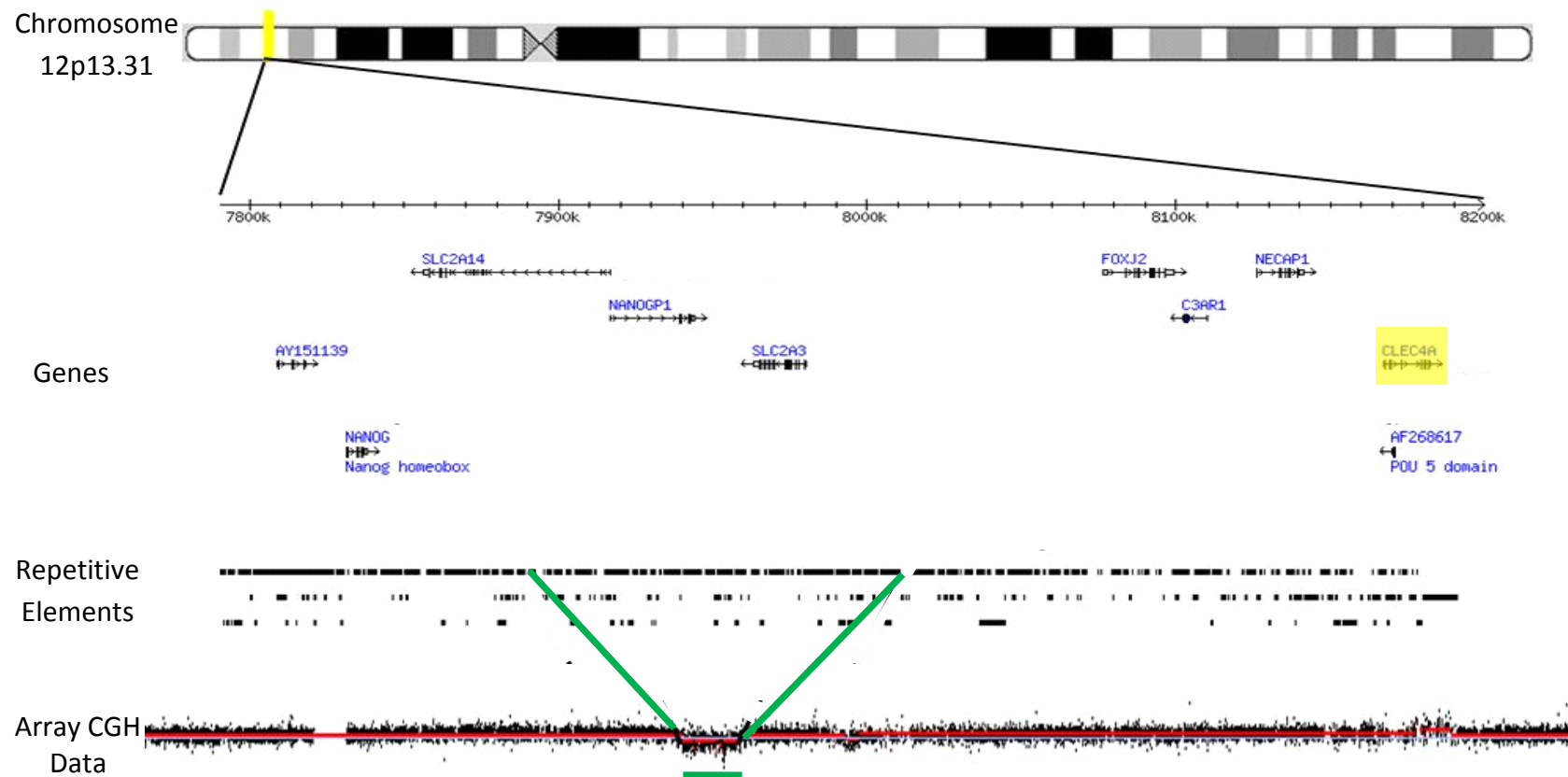


Figure 1.6: Structural Variation at 12p13.31

The location of 12p13.31 is shown, along with a close-up of the genes and repetitive elements within this region (image taken from UCSC genome browser). Also shown is a graph of sample arrayCGH data from within this region, produced using Nimblegen oaCGH. The graph displays a comparison of the relative hybridisation signal to each probe from two DNA samples, shown as the log₂ of each ratio plotted against the genomic position of the probes. It can be seen that a region of structural variation was identified within this locus, the position of which is indicated on the region diagram by green lines. *CLEC4a*, our original candidate gene, is highlighted.

1.8 Summary and Project Aims

Since the extent to which the genome varies in copy number was realised in the mid-2000s, there has been an increasing interest in this form of genetic variation and its role in disease susceptibility. A number of methods are available to study CNV, including arrayCGH, a powerful comparative method which can be used to study copy number variation on a genome-wide scale. However, this technique is mainly available commercially which is a significant limitation to the uptake of this method, and therefore also research in this field. There is currently a paucity of in-house systems for arrayCGH; therefore one of the aims of research for this thesis was to develop this technique on an in-house microarray platform. Once optimised, we planned to use this method to investigate structural variation, with particular interest in a putative region of CNV located on chromosome 12p13.31, for which there was evidence for an association with RA.

Despite the detection of a number of susceptibility regions, a considerable proportion of the genetic basis of complex disorders such as diabetes and RA remains to be explained. Some of this may be explained by the involvement of CNV; however, so far only one copy number variable locus, *CCL3L1*, has been implicated in RA susceptibility (McKinney *et al*, 2008). Investigations with our research group identified a putative region of structural variation on chromosome 12p13.31, located adjacent to an interval previously shown to be associated with RA. The following thesis is concerned with characterising this putative region of structural variation in greater detail, using a number of methods to firstly confirm its presence, and secondly to investigate the role which CNV within this region plays in susceptibility to RA.

Chapter 2

Materials & Methods

2.1 Materials and Equipment

2.1.1 Chemicals, Enzymes and Oligonucleotides

All chemicals were supplied by Sigma-Aldrich (Gillingham, UK), unless otherwise stated. Buffers were made up according to Molecular Cloning (Sambrook & Russell, 3rd Ed.), except those used in oaCGH which were either supplied by Febit (Heidelberg, Germany) or made up according to their protocols. All H₂O was purified using a Millipore Milli-Q water purification system.

Restriction enzymes and standard DNA ladders were provided by New England Biolabs (Hitchin, UK). DNA *Taq* polymerase was obtained from Kapa Biosystems (Boston, USA) and oligonucleotides were provided by Biomers (Ulm, Germany).

2.1.2 DNA Samples

Unless otherwise stated, all DNA samples used for method development and optimisation were Northern European control samples purified from blood lymphocytes. DNA samples used for oaCGH were products of whole genome amplifications (see Section 2.2.3.1). Sets of DNA samples for association and population studies were obtained from collaborators, as detailed in Table 2.1.

Table 2.1: DNA Samples

Sample Set	Provided By
HapMap Phase I Samples*	Prof. Mark Jobling (University of Leicester, UK)
Centre d'Etude du Polymorphisme Humain (CEPH) Families*	Prof. Alec Jeffreys (University of Leicester, UK)
Swedish RA Case-Control Cohort	Prof. Lars Klareskog & Dr Leonid Padyukov (Karolinska Institutet, Stockholm, Sweden)
Swedish Psoriasis Cohort	Dr Fabio Sanchez (Karolinska Institutet, Stockholm, Sweden)
UK RA Cohort	Prof. Jane Worthington (University of Manchester, UK)
UK Cardiovascular Case-Control Cohort	Prof. Nilesh Samani (University of Leicester, UK)
1958 British Birth Cohort	Centre for Longitudinal Studies (London, UK)

*HapMap and CEPH samples from the initial studies were provided by collaborators at the University of Leicester. Individual samples of interest used in further studies were purchased from the Coriell Institute for Medical Research.

2.1.3 Equipment and Computer Software

Table 2.2: Equipment

Equipment	Model	Supplier	Manufacturer
Microarray Facility	Geniom One	Febit (Heidelberg, Germany)	Febit
Hybridisation oven	INE 200-800	Fisher Scientific (Loughborough, UK)	Memmert
Gel electrophoresis power units	PowerPac Basic	BioRad Laboratories	BioRad Laboratories
Gel documentation system	G-Box	Syngene (Cambridge, UK)	Syngene
Electrophoresis tanks	1200 ml & 800 ml	-	University of Leicester Workshop
Electrophoresis cassettes	100 ml & 300 ml	-	University of Leicester Workshop
Microtube centrifuge	5415-D & 5415-R	Eppendorf (Cambridge, UK)	Eppendorf
Plate centrifuge	Jouan B4i	Thermo Scientific (Basingstoke, UK)	Thermo
Thermal cyclers	MBS 0.2G	Thermo Scientific	Hybaid
Single channel pipettes	Finnpipette	Thermo Scientific	Thermo Scientific
Multichannel pipettes	Finnpipette Novus	Thermo Scientific	Thermo Scientific
Rocking platform	OS-500	VWR (Lutterworth, UK)	VWR
Water bath	SUB 6	Grant Instruments (Shepreth, UK)	Grant
Heating block	Driblock DB 2A	Bibby-Scientific Ltd (Stone, UK)	Techne
Dark Reader Transilluminator	DR195M	GRI (Braintree, UK)	Clare Chemical Research
Benchtop Minicentrifuge	Galaxy Ministar	VWR	VWR

Table 2.3: Computer Software

Software	Reference/Company	Use
Geniom Software	Febit (Heidelberg, Germany)	oaCGH experiments, extracting signal intensities
Microsoft Office Excel 2007	Microsoft	Data manipulation e.g. calculating averages, normalisation
Newbler	Roche	<i>De novo</i> sequence assembly & analysis
GeneSnap	Syngene	Imaging agarose gels
GeneScan	Syngene	Extracting signal intensities from agarose gels
Gepard 1.30	Krumsiek <i>et al</i> , 2007	Dotplots
UCSC genome browser	Kent <i>et al</i> , 2002	BLAT alignments, primer design, sequence visualisation & comparison
TraceSpace	Owen Lancaster (Personal communication)	Viewing & analysis of sequencing trace data
GraphPad	www.graphpad.com	Statistics
R 2.10.1	www.r-project.org	Creating density plots
Standalone BLAT software	Kent, 2002	High volume BLAT alignments
Image J 1.43	Abramoff <i>et al</i> , 2004	Quantification of Labelling blots

2.2 Oligo-arrayCGH on the Geniom Platform

Febit provide a detailed instruction manual and workflows which take the user through each process carried out using the Geniom One Microarray System. This section provides a basic outline of the processes and programmes used, for more detail refer to the Geniom workflows which can be found in Appendix A.

2.2.1 Array Design

The oligonucleotide probes used in the array design are shown in Table 2.4. A number of probes previously used in arrayCGH carried out on the Nimblegen (Waldkraiburg, Germany) array platform corresponding to the chromosome 12p13.31 region were included. However, since probes up to a maximum length of 80 nt can be synthesised on the Nimblegen arrays, compared to 60 nt on the Geniom biochip, the probes were first filtered to select only those with a length equal to or less than 60 nt.

In order to validate our results, a number of regions of known copy number were included on the arrays. Nimblegen control sequences corresponding to a single copy region of chromosome 13q12 and ‘random’ probes which contained no match in the human genome were included. Other control probes were taken from the male-specific SRY region on Yp11, a number of loci which had previously been used as single-copy controls for MAPH experiments by Ed Hollox, and three regions of the X chromosome, two from Xp11 and one from Xp22, as female-specific controls. The sets of X chromosome probes each correspond to DNA sequence from a single BAC clone used previously in an array CGH study (Snijders *et al*, 2001) (Figure 2.1). The three clones were chosen due to the fact that they showed a consistent increase in \log_2 ratio in female compared to male DNA samples in this study, making them likely to be good

control probes for our investigations. Probes for the defensin region on chromosome 8p23 were also included, as this locus is known to be highly copy number variable, as well as probes from sequence distal and proximal of this region which should be single copy. Probes for these regions were created by splitting the required sequence into adjacent 60 nt fragments.

Table 2.4: Probes included on Geniom Biochip (position is according to reference genome build NCBI36/hg18, March 2006)

Name	Chrm	Position	Description	No. Probes
CHR12CSDA	12p13	8160266-8500001	Segmental Duplication A	2053
CHR12LR	12p13	7880003-7915299 & 7990555-8020052	Long Region	797
Def B Region	8p23	7666629-7828449	Defensins	1372
<i>ST19F2</i>	5q35	180620496-180620665	MAPH Controls	49
<i>ST1G9</i>	19q13	63772964-63773090	MAPH Controls	55
<i>PLUNC</i>	20q11	31294502-31295012	MAPH Controls	39
<i>GATA4</i>	8p23	11653239-11653641	MAPH Controls	58
<i>BPI</i>	20q11	36373657-36373761	MAPH Controls	31
<i>TBP</i>	6q27	170706114-170706362	MAPH Controls	46
<i>PX3F11</i>	Xp21	32547543-32547885	MAPH Controls	30
XKR5teldefB	8p23	6638244-6689964	Distal to Def B	550
CendefB1	8p23	8159498-8213078	Proximal to Def B	473
CHR13	13q12	30088007-30090798	Controls	271
Randoms	-	-	No match in human genome	99
<i>SRY</i>	Yp11	2714577-2716247	Male Specific	100
XA	Xp11	43692969-43693923	Female Specific	61
XB	Xp11	49717771-49811276	Female Specific	204
XC	Xp22	20442179-20597564	Female Specific	438
Febit Longmer Controls	-	-	Febit control probes	50
TOTAL:				6776

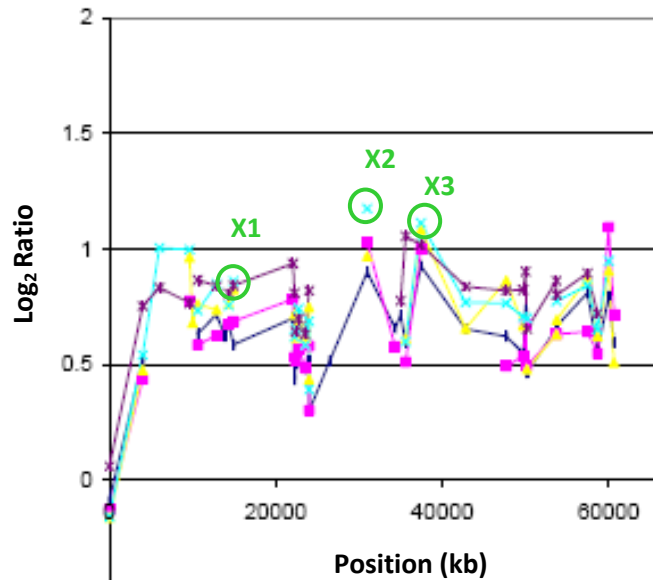


Figure 2.1: Selection of X Chromosome Clones

X chromosome probes were selected using data from Snijders *et al* 2001 (supplementary information). The graph shows data from hybridisation to BAC clones corresponding to regions of sequence on chromosome Xp. Each point on the graph represents the mean \log_2 ratio of the triplicate spots for each of the clones from a male/female hybridisation. A different coloured line represents each of five separate hybridisations. Three clones which showed a consistent increase in \log_2 ratio between male/female samples were selected, these are circled in green and numbered. The chosen clones were then identified and fragmented into short sequences for inclusion on the Geniom biochip.

2.2.2 Array Synthesis

Synthesis was carried out following the manufacturer's instructions. Once synthesis is complete, the array can be stored at 4°C for a few days until required.

2.2.3 Preparation of DNA

2.2.3.1 Whole Genome Amplification

Whole genome amplification (WGA) was carried out using a REPLI-g mini kit (Qiagen) following the manufacturers protocol. Briefly, 2.5 μ l denaturation buffer (D1) was added to 2.5 μ l of 8 ng/ μ l DNA, mixed and incubated at room temperature for 3

minutes. 5 µl neutralisation buffer (N1) was added and the solution mixed. 10 µl H₂O, 29 µl REPLI-g mini reaction buffer and 1 µl REPLI-g mini DNA polymerase were added to each reaction, the mixture was spun in a benchtop minicentrifuge and incubated in an MBS 0.2G thermal cycler at 30°C for 16 hours, followed by 65°C for 3 minutes to inactivate the DNA polymerase. Amplified DNA was stored at -20°C.

2.2.3.2 *Restriction Digests*

Amplified DNA was fragmented using the restriction enzyme *DpnII*. In some instances a subsequent digest with *AluI* was also carried out. For both enzymes, REPLI-g products were digested in a 60 µl reaction containing 47.5 µl amplified DNA, 1.2 µl spermidine (2 mM), 3 U *DpnII* or *AluI*, 6 µl *DpnII* buffer (or NEB Buffer 2 where *AluI* was used), 2.3 µl H₂O. The reaction was incubated at 37°C for 18 hours in an MBS 0.2G thermal cycler, followed by inactivation at 65°C for 20 minutes.

2.2.3.3 *Purification of Digest Products*

Inactivated digests were cooled on ice and the solution transferred to a 1.5 ml centrifuge tube. 240 µl TE (100 mM Tris pH8, 10 mM EDTA pH8) and 300 µl Phenol:Chloroform:Isoamyl alcohol (25:24:1) were added. Vessels containing Phenol were only opened in a fume hood. The solution was spun in an Eppendorf 5415-D microcentrifuge at room temperature for 5 mins at 11000 g. The top aqueous layer containing the DNA was aspirated using a pipette, and placed into a fresh microtube. The lower layer, which contained the phenol, was then discarded.

2.2.3.4 *Precipitation of DNA*

0.1 volume 3M Sodium Acetate pH 5.2 and between 2.5 and 3 volumes 100% ethanol were added to each sample. Samples were mixed thoroughly and incubated at either -20°C for 90 minutes or at -80°C for 30 mins and then centrifuged at 4°C for 20 mins in an Eppendorf 5415-R refrigerated microcentrifuge at a speed of 16100 g. The supernatant was carefully removed and discarded. The pellet was dried by placing the microcentrifuge tube in a Driblock DB 2A heating block on a low heat. When all traces of ethanol had evaporated, the pellet was resuspended in 30 µl sterile distilled water. The amount of DNA contained in the purified digest was estimated on an agarose gel, using the λ DNA *Hind*III marker ladder (NEB) as a reference. Digested DNA was then stored at -20°C.

2.2.3.5 *Biotin Labelling*

The initial labelling protocol was as follows. 2 µg digested genomic DNA was diluted to a total volume of 26.4 µl with sterile water in a microtube. The solution was denatured at 99°C for 5 mins in a waterbath, spun briefly in a minicentrifuge and put on ice. 3 µg/µl random hexamers were diluted 1:5 with buffer (1 mM Tris pH8 and 1 mM EDTA). 1.5 µl of this dilution was added to the denatured DNA solution on ice, along with 8 µl 5x Random Primer reaction buffer (2.5 M HEPES pH6, 2.5 M Tris pH8, 1 M MgCl₂, 1 M 2-mercaptoethanol), 1 µl Oligonucleotide mix (100 mM dATP, dCTP and dGTP, 10 mM dTTP, 1 mM Biotin-dUTP (Roche)), 1.6 µl 10 mg/ml BSA (Sigma) solution and 1.5 µl 5 U/µl Klenow Fragment (Fermentas). The solution was mixed and incubated for 3 hours at 37°C in an MBS 0.2G thermal cycler. The polymerase was inactivated by incubating at 75°C for 10 mins. The solution was then spun in a

minicentrifuge and cooled on ice. DNA was then precipitated, as described in Section 2.2.1.4, with the addition of a second wash step using 1.5 ml 70% cold ethanol. Samples were spun in a 5415-R refrigerated centrifuge at 4°C for 5 mins at a speed of 16100 g. The supernatant was removed and discarded. The pellet was then resuspended in hybridisation solution (see Section 2.2.5).

2.2.4 Pre-Hybridisation

The synthesised Geniom biochip chip was first denatured to break any bonds that may have formed between the oligonucleotide probes prior to hybridisation. This denaturing step takes place inside the Geniom. The biochip was then removed from the Geniom and inserted into the external hybridisation holder. 15 µl pre-hybridisation solution (Febit) was introduced into each array. The solution was pipetted into the wells and then drawn through each array using a syringe. The array was then incubated at room temperature for 15 minutes, after which the pre-hybridisation solution was removed from the arrays using a syringe.

2.2.5 Hybridisation

The pellet of labelled genomic DNA (see section 2.2.3.5) was resuspended in 15 µl hybridisation solution (5x SCC, 40% Formamide, 0.01% Tween-20, 0.1x TE, 0.5 mg/ml BSA, 0.1 mg/ml salmon sperm DNA (Sigma)) by pipetting. In some instances this mixture was incubated at 4°C overnight to allow for complete resuspension of the DNA. Before hybridisation, samples were denatured at 95°C for 3 minutes in a waterbath and cooled on ice. They were then loaded into wells on the external hybridisation holder and drawn into the microchannels of the Geniom biochip using a

syringe. Hybridisation was carried out at 40°C for the allotted time, which ranged from 3 to 96 hours.

2.2.6 Wash Steps

2.2.6.1 *Machine Washes*

Hybridisation solution was removed from the microchannels using a syringe and the arrays were then flushed with 6 x SSPE (Saline-Sodium Phosphate-EDTA) (Fluka) to remove all traces of DNA. The biochip was inserted back into the Geniom for the remaining washes. Buffers were loaded into the Geniom machine (Buffer 1 = 0.5 x SSPE (stringent wash), Buffer 2 = 6 x SSPE (non-stringent wash)) along with a Streptavidin Phycoerythrin marker solution (SAPE) (44 µl SAPE + 9 ml 6 x SSPE) which was prepared freshly each time. Standard washes briefly consisted of a non-stringent (6 x SSPE) wash at 25°C, followed by a stringent wash (0.5 x SSPE) at 40°C, a 15 minute incubation with SAPE and then a final non-stringent wash at 25°C. For detailed workflows see Appendix A.

2.2.6.2 *Manual Washes*

For some biochips, a number of washes were performed outside of the Geniom to allow different formamide concentrations to be used for subarrays of the same biochip, and also for washes using a peristaltic pump. Manual washes were performed with the biochip still in the hybridisation chamber. It was then removed and inserted back into the Geniom machine for the SAPE incubation and final wash. In most cases a non-stringent wash preceding the SAPE incubation was also performed within the Geniom.

Washes performed using a syringe were carried out with buffers at 40°C, and included a 10 minute incubation at this temperature.

2.2.7 Detection

Detection was carried out inside the Geniom using a Cy3 filter, which was also suitable for visualising SAPE. The biochip was first imaged using the autoexposure setting and the exposure was then adjusted manually if necessary, depending on the quality of the first image.

2.2.8 Signal Amplification

After the initial detection step, a signal amplification procedure could be carried out. This process exploits the high affinity of streptavidin for biotin. Fresh antibody solution was prepared (2x Stain buffer (12x MES (64.61 g MES hydrate, 193.3 g MES sodium salt, pH 6.5-6.7), 5 M NaCl, 10% Tween-20), 50 mg/ml BSA, 10 mg/ml Goat IgG stock (Sigma-Aldrich), 0.5 mg/ml biotinylated anti-streptavidin antibody (Vector Laboratories)) and loaded into the Geniom. The program 'Febit Signal Amplification (CSE)' was run and then the detection process carried out as before. Afterwards the Geniom was washed with freshly diluted 0.5x Sodium Hypochlorite, followed by water, to clean the system.

2.2.9 Analysis

The signal intensity (fluorescence) for each feature on the biochip was extracted using software provided by Febit. The median signal intensity for each feature was taken, and

normalisation and data analysis carried out in Microsoft Excel. The normalisation procedure is described in detail in Chapter 3.

2.3 Polymerase Chain Reaction (PCR)

2.3.1 Primer design

Primers were designed manually from the DNA sequence of interest, with the aid of UCSC genome browser (Kent *et al*, 2002). Where possible, primer length was restricted to 20-24 bp, with no complementary sequence within primers or between primers in a pair, and a difference in $T_m < 5^{\circ}\text{C}$ for each primer pair. Ideally, repetitive elements and SNPs were avoided. All PCR primers used in research for this thesis are listed in Appendix B.

2.3.2 Reaction Conditions

2.3.2.1 *Standard PCR Conditions*

Target DNA sequence was amplified using PCR (Mullis *et al*, 1986; Saiki *et al*, 1985). The *Taq* DNA polymerase employed was Kapa *Taq* (Kapa Biosystems), used in conjunction with 10x buffer B (100 mM Tris-HCl pH 8.3, 500 mM KCl, 15 mM MgCl₂ (Kapa Biosystems)). Amplifications were carried out in a 10 μl reaction containing 10 ng genomic DNA, 1x Buffer B, 0.2 mM combined dNTPs, 0.3 μM each primer and 0.03 U/ μl *Taq* DNA polymerase. Standard cycling conditions were performed in an MBS 0.2G thermal cycler as follows: 96°C for 5 min; 35 cycles of 96°C for 30 sec, annealing temperature for 20 sec, 72°C for 1 min. For products >1 kb, the extension time was increased by 1 minute for each extra kb.

2.3.2.2 *Betaine Supplemented PCR Conditions*

10 μl PCR contained 10 ng DNA, 1x Kapa buffer B, 1-2 M betaine (Sigma), 0.2 mM dNTP mixture, 0.3 μM each primer, 0.1 U/ μl *Taq* DNA polymerase. PCRs were

performed in an MBS 0.2G thermal cycler as follows: 98°C for 1 min; 35 cycles of 98°C for 15 sec, annealing temperature for 15 sec and 72°C for 1 min, followed by a final extension of 72°C for 5 min. For amplifications with a product length of > 1kb, the extension temperature was increased by 1 minute per kb.

2.3.2.3 *PCR using 11.1 x buffer*

Each 20 µl reaction contained 20 ng DNA, 1x 11.1x buffer (0.49 M Tris-HCl pH8.8, 0.12 M (NH₄)₂SO₄), 0.05 M MgCl₂, 77 mM β-Mercaptoethanol, 5 µM EDTA pH8.0, 11.1 mM each nucleotide (dATP, dCTP, dGTP & dTTP), 1.3 mg/ml Bovine Serum Albumin (BSA)), 0.3 µM each primer and 0.1 U/µl *Taq* DNA polymerase. The cycling conditions were the same as for the standard PCRs.

2.3.2.4 *Touchdown PCR*

The standard PCR reaction mix was used, but with altered cycling conditions. These consisted of 94°C for 5 min; 15 cycles of 94°C for 20 sec, 68°C for 20 sec in the first cycle and then a reduction in annealing temperature of -0.6°C per cycle, 72°C for 2 min; followed by 20 cycles of 94°C for 20 sec, 60°C for 20 sec, 72°C for 2 min.

2.3.2.5 *FastStart High Fidelity PCR*

The FastStart High Fidelity PCR system is manufactured by Roche. Each 20 µl reaction mix consisted of 20 ng DNA, 1x FastStart High Fidelity Reaction Buffer (0.2 mM dNTPs, 0.4 µM each primer, 0.05 U/µl FastStart High Fidelity Enzyme. Cycling

conditions were 95°C for 2 min, followed by 35 cycles at 95°C for 30 sec, annealing temperature for 30 sec, 72°C for 4 min.

2.3.3 Agarose Gel Electrophoresis

PCR products were separated on low electroendosmosis (LE) agarose gels by electrophoresis. DNA fragments <1 kb were fractionated on 2 % (w/v) gels whereas those >1 kb were separated on 0.7% (w/v) or 0.8% (w/v) gels. Gels were prepared by melting the required amount of LE agarose powder (SeaKem) in 1x Tris–Borate EDTA (TBE) buffer. Ethidium bromide (stock 10 mg/ml), a DNA intercalator, was added to the solution at a final concentration of 2 µg/ml to enable visualisation of the DNA on the gel.

Gels were run in 1 x TBE running buffer, with Ethidium Bromide (stock 10 mg/ml) added at a final concentration of 2 µg/ml. Fragment size was confirmed by running samples alongside an appropriate DNA molecular weight marker, for example λ DNA *Hind*III ladder, 1 kb DNA ladder or 100 bp DNA ladder (all NEB). In the case of the λ DNA *Hind*III ladder, this was first denatured; 10 µl of 20 ng/µl DNA ladder was heated at 65°C in a water bath for 2 minutes to separate cohesive ends of fragments and then quenched on ice for a few minutes. 3 µl Orange G loading buffer (0.3 g OrangeG powder, 5x TBE, 10% Glycerol) was added to 10 µl of each DNA sample as well as the molecular weight marker before loading onto the agarose gel. Gels were photographed under UV light to visualise the DNA bands.

2.3.4 Alkaline Gel Electrophoresis

The required amount of agarose was added to 90% of the final gel volume of water, weighed prior to melting. The mixture was made up to the pre melting weight with distilled water, and left to cool to 60°C for 1 hour in a water bath (Grant Instruments). Prior to casting, 10% of the gel volume of 10x alkaline buffer (500mM NaOH, 10mM EDTA) was added and swirled to mix thoroughly.

During electrophoresis, the running buffer was maintained at 4°C to prevent the gel from overheating. 1x alkaline buffer was used as the running buffer. Samples were mixed with 3 µl/10 µl Orange G loading dye and loaded into the wells of the gel, and left for 5 min prior to electrophoresis. Electrophoresis was carried out at 40 V for 5.5 hours. The gel was then neutralised and stained in 1x TBE with Ethidium bromide (stock concentration 10 mg/ml) added at a concentration of 2 µg/ml, for 45 minutes. This was carried out at room temperature and on a rocking platform (VWR), to allow constant mixing of the solution. The agarose gel was photographed under UV light as for conventional gel electrophoresis (Gel Doc, Syngene).

2.3.5 DNA Sequencing

The PCR products to be sequenced were separated on an agarose gel, run for an extended period of time in order to achieve a good separation. Each band was cut out from the gel with a scalpel, under a Dark Reader (Clare Chemical Research). DNA was extracted and purified using a gel extraction kit (Qiagen) following the manufacturers protocol. Extracted DNA was quantified by separating the fragments by agarose gel electrophoresis and comparing the band intensity to that of bands of known intensity from a suitable molecular weight marker (e.g. 100 bp DNA ladder (NEB)).

A separate sequencing reaction was set up for each of the forward and reverse primers for the PCR product to be sequenced. Each sequencing reaction contained: 1 μ l of Big Dye terminator v3.1; 3.5 μ l 5 x Big Dye Buffer; 1 μ l of 3.2 μ M primer; 20 to 30 ng DNA and water to a final volume of 20 μ l. The sequencing reaction was carried out in a thermal cycler using 25 cycles of 96°C for 10 seconds, 50°C for 5 seconds and 60°C for 4 minutes. The reaction was cleaned up by the addition of 2 μ l 2.2% (w/v) SDS to the reaction mix, followed by incubation at 98°C for 5 minutes in an MBS 0.2G thermal cycler followed by 25°C for 10 minutes. Immediately after this step, dye terminators were removed using Performa Dye-Ex Gel Filtration Columns (Performa), following the manufacturers protocol. At this point the product was submitted for sequencing to the PNACL (Protein and Nucleic Acid Chemistry Laboratory, University of Leicester, UK) where samples were analysed using a 3730 sequencer (Applied Biosystems). Sequencing data was visualised as a chromatogram using the software FinchTV.

2.4 Parologue Ratio Test (PRT)-based Assays

A number of assays were designed based on the Parologue Ratio Test (Armour *et al*, 2007). This method enables the detection of copy number variation by comparing the amount of product produced from a sequence of interest and a paralogous single copy reference sequence, which are amplified simultaneously using the same pair of primers. Since our region of interest is a tandem duplication, we designed assays to compare the relative copy number of the two units of this duplication, rather than using a single copy reference region.

2.4.1 Series A Assays

A assays were designed to exploit the different sized gaps which are present between regions of sequence found in both units of the tandem duplication. Two different-sized products are amplified by each assay, one from each unit, which can be separated on an agarose gel.

Samples were either diluted to 10 ng/μl or if too dilute, 10 ng was air-dried in a PCR plate overnight. All PCRs contained 10 ng DNA, 0.2 mM dNTP mixture, 0.15 μM each forward and reverse primer, 1x Kapa Buffer B, 2 M Betaine and 0.02 U *Taq* DNA polymerase. Amplifications were performed in an MBS 0.2G thermal cycler at 98°C for 1 min followed by 35 cycles of 98°C for 15 seconds, annealing temperature for 15 sec, 72°C for 1 minute, and a final extension of 72°C for 1 minute. Products were separated by agarose gel electrophoresis as described in section 2.3.3.

2.4.1.1 *Data Analysis*

The signal intensity of each product was extracted from agarose gels using the Genetools software (Syngene). Simple data analysis was then carried out in Microsoft Excel, as follows. The ratio of product from each unit of the tandem duplication was calculated by dividing the signal intensity of the product from unit B by the signal intensity of the product from unit A. Data was normalised so that the ratio of a ‘normal’ sample, with equal copy number of both units of the tandem duplication, was equal to 1. To do this, the median signal intensity of each row of twelve samples (corresponding to a single row on a PCR plate) was calculated. 1 was then divided by this value to work out the multiplication factor required to bring the median to 1. Each of the twelve samples was then multiplied by this number. This process was repeated for each set of twelve samples. The \log_2 ratio for each sample was then calculated.

2.4.1.2 *Identifying Classes of Copy Number Variation*

Four distinct classes of variant genotype were detected. The data tend to cluster around the expected values as determined from the ratios, as expected, although there is some degree of variation between each PCR plate. This is discussed in more detail in Chapter 5. Samples were categorised according to cut off points determined using the statistical package R. Expected data points were determined using the \log_2 ratios and the normal distributions of each plate were used to refine these categories on real samples. Individual samples with \log_2 values between -0.45 and -0.5, as well as 0.45 and 0.5 were excluded from this analysis in the Swedish samples as they were thought to be too close to the cluster of data from ‘normal’ samples to be accurately placed in one category or another. The number of samples in each category was counted in cases and

controls and frequencies compared for each class of copy number variation. The significance of any differences in frequency between case and control samples cohorts was determined using a chi-squared test in most incidences, either one- or two- tailed depending on the situation. Where there was a low number of samples in both case and control categories, a Fishers Exact Test was used instead.

2.4.2 Series B Assays

B assays were designed, each of which employed a set of three primers. These include a single forward primer, common to both units of the tandem duplication, and two reverse primers, one specific to each unit of the tandem duplication.

Initially all reactions were set up with a reaction mix consisting of 10 ng DNA, 0.2 mM dNTPs, 0.3 μ M forward primer and 0.15 μ M each reverse primer, 1x Kapa Buffer B, 2 M Betaine and 0.02 U Kapa Taq. Amplifications were performed in an MBS 0.2G thermal cycler using an initial denaturation step at 98°C for 1 minute followed by 35 cycles of 98°C for 15 seconds, annealing temperature for 15 seconds, 72°C for 1 minute and a final extension of 72°C for 5 minutes.

Products were separated by agarose gel electrophoresis as described in section 2.3.3. Analysis and determination of copy number was carried out as described in Section 2.4.1 for the A assays.

2.5 *In Silico* Sequence Studies

2.5.1 *De Novo* Sequence Assembly

All public domain sequencing trace files available from this region as of November 2008 (including the Venter traces obtained using the Sanger sequencing method and the Watson traces from Roche 454 sequencing) were downloaded from the NCBI Trace Archive (<http://www.ncbi.nlm.nih.gov/Traces/>). *De novo* sequence assembly was attempted using several different algorithms from programs including DNABaser (www.dnabaser.com) and Mira (Chevreux *et al*, 1999). The final *de novo* assembly was performed using the Newbler Assembler software (Roche) since this proved to be the quickest and most effective assembly program in producing the most complete contigs, as well as being the most user-friendly. For this purpose the default settings for sequence trimming (to exclude traces of low quality) and alignment thresholds were used. Only the Sanger sequencing reads were included in the final analysis, since the shorter 454 reads did not assemble well *de novo*. The resulting contigs were aligned to the reference genome (build NCBI36/hg18, March 2006). Although most regions assembled without difficulty, there were some incidences where adjacent contigs mapping to the region failed to join together. The most likely reason for these gaps is due to near identical or repetitive sequence that the assembly software is unable to assemble across. All but one of the gaps in the *de novo* sequence assembly were subsequently amplified across using PCR (Chapter 4.3.3).

2.5.2 Dot Plots

Dot plots were created using the online software Gepard (Krumdiek *et al*, 2007). The DNA sequence of each unit of the tandem duplication was obtained from UCSC

genome browser and entered into Gepard. Unit A was used as Sequence A, and unit B as Sequence B. The window size was adjusted to either 100bp or 400bp, and the stringency set so that only exact matches were plotted on the resulting graph. In the case of the 400bp plot, where only three regions fit the required criteria, the genomic location of these sequence matched in each unit was identified by adding the location of the dot to the genomic location of the first base of the relevant unit of the tandem duplication.

2.5.3 BLAT Alignments

BLAT alignments were used to investigate the structure of sequence identity shared between the two units of the tandem duplication. The DNA sequence of each of the two units of the tandem duplication was first split into fragments in a tiling path across each unit using a Perl script (Owen Lancaster, personal communication). Several sets were created, with fragment lengths of 100bp, 500bp or 1kb. Separate sets were made both with and without repeat masking prior to fragmenting. In cases where repeat-masking was carried out, any fragments which contained repeat elements (represented by runs of N residues in the sequence) were discarded. The sequences were then aligned to the reference sequence (build NCBI36/hg18, March 2006) using the stand-alone BLAT software (Kent, 2002) to run the search locally, since high volume BLAT searches are not permitted on the UCSC BLAT server.

The results of the alignment were filtered to remove hits which did not correspond to the tandem duplication under investigation, as well as hits with a sequence identity below 80%, to leave the longest sequence matches in each category (For example, those equal to 100bp in the 100bp category and those >400bp in the 500bp category). For

each set of fragments, separate alignments were performed with and without prior repeat masking. In the case of the 100bp fragments, the short length meant that some fragments contained only repetitive element sequence and therefore produced a vast number of uninformative matches, therefore only the repeat masked fragments were used for analysis. However, since the majority of repetitive elements are less than 400bp in size, this was not as much of a problem for the 500bp fragments. These were likely to contain enough non-repeat sequence to produce unique matches, and so in this case, the non-repeat masked sequences were used for analysis. For each match, the position, length of the match and % sequence identity was plotted on a scatter graph to allow visual comparison of the data.

Chapter 3

Development of Oligo-Array CGH on the Geniom One Microarray Platform

3.1 Introduction

Over recent years there has been an increasing interest in copy number variation. To enable the study of this type of variation, a number of methods have been developed (discussed in Chapter 1.4). One such technique is Comparative Genomic Hybridisation (CGH), which is able to reveal differences in copy number between two DNA samples by comparing their relative degree of hybridisation to specific DNA sequences. Initially, this involved the hybridisation of fragmented DNA samples to metaphase chromosomes (Kallioniemi *et al*, 1992). Further developments of CGH followed in which the DNA samples were hybridised to genetic material fixed onto microarrays (e.g. isolated from BAC clones) (Pinkel *et al*, 1998) and, more recently, arrays of oligonucleotides (Carvalho *et al*, 2004). The latter technique, known as oligonucleotide-array Comparative Genomic Hybridisation (oaCGH), is the focus of method development investigations described in this chapter.

OaCGH involves the hybridisation of fluorescently labelled fragmented genomic DNA from two samples, typically a reference sample of known copy number and a test sample for which the copy number is to be determined, to arrays of target oligonucleotides. These arrays are produced either by spotting oligonucleotides onto a glass slide, or synthesising them *in situ*. The copy number of sequences from each of the two samples hybridised to each oligonucleotide probe is assessed by measuring their

relative fluorescence, and the ratio used to determine differences in copy number between the test and reference genomes over the targeted regions.

OaCGH systems may either be single or dual colour, in terms of the number of wavelengths at which they can simultaneously detect fluorescence. In a dual colour system, the two samples are labelled with different fluorochromes which can be detected at different wavelengths, for example Cy3 and Cy5. Samples are co-hybridised to the same microarray and the relative amount of fluorescence from each fluorochrome is compared. However, not all systems are able to distinguish between fluorescence at different wavelengths in a single experiment, and so samples may instead be labelled with the same fluorochrome and hybridised to two different arrays which are then compared.

OaCGH services are provided by companies including Affymetrix and Nimblegen. These services generally require that the synthesis of the arrays, and in some cases also the hybridisation and detection steps, is carried out in the manufacturer's facilities. This allows the customer little control over most stages of the experiment, and tends to be expensive. The development of in-house systems is desirable as this would allow the user a greater level of control and flexibility over many aspects of the process. However, there is currently a paucity of reliable and cost effective in-house microarray technologies. Therefore the first part of my research for this thesis focussed on the development of an optimised protocol for oaCGH on an in-house customisable microarray platform. Using the commercially available Nimblegen oaCGH platform, investigations within our research group had previously identified a putative region of structural variation on Chromosome 12p13.31 (Chapter 1.7). Once developed, our aim was to use our optimised oaCGH protocol for in-house studies, to further investigate variation in this region.

3.1.1 The Geniom One Microarray System

The platform we selected to use for development of an in-house oaCGH protocol is a bench-top microarray system called Geniom One, which was manufactured by Febit (Baum *et al*, 2003). This system can be employed for a range of applications, including gene expression studies and microRNA profiling. Each stage of the experiment, including synthesis, hybridisation, washing and detection, is carried out by the user, and all except hybridisation take place within the Geniom machine. Whereas most commercially available platforms contain a single array per chip, each Geniom biochip is made up of eight subarrays, each containing 6776 features (Figure 3.1). This design allows eight independent hybridisations to be carried out simultaneously. Another unique characteristic of the Geniom technology is that these features are located within microchannels; each subarray comprises a single microchannel. During hybridisation or wash steps, liquid (for example buffer) enters the array at one end of the microchannel, is drawn through and then leaves from the other end.

Oligonucleotide probes are synthesised *in situ* within the microchannels, up to a maximum length of 60 nt. Synthesis uses a light-activated deprotection method, which employs virtual masks relayed to a digital array of micromirrors (Singh-Gasson *et al*, 1999). These masks are used to define the sequence of each oligonucleotide feature.

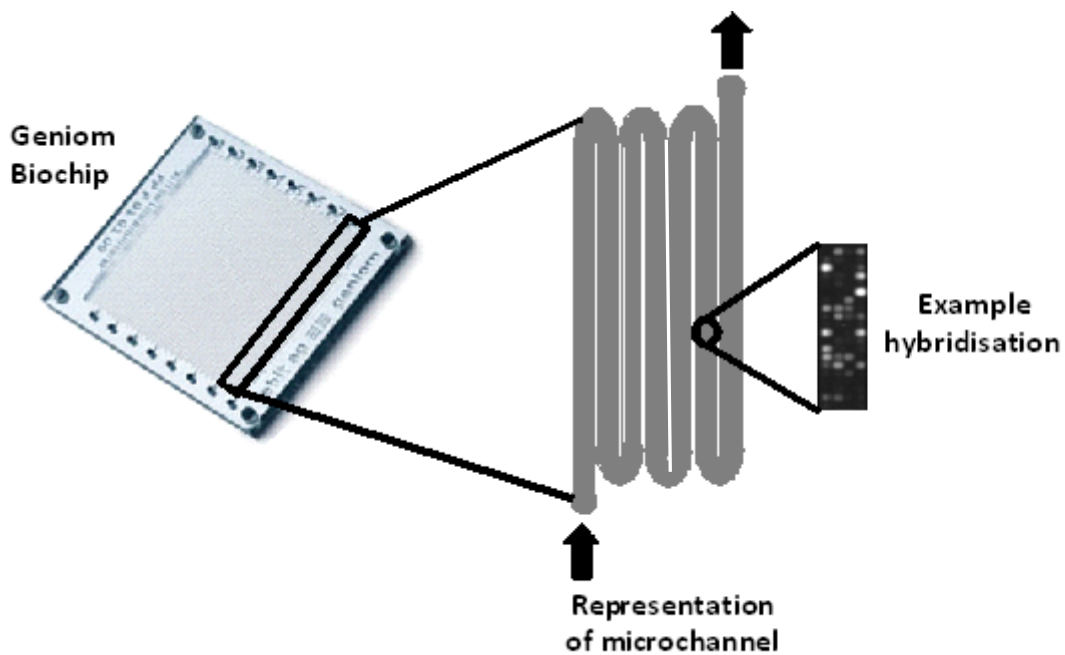


Figure 3.1: Structure of a Geniom Biochip

The diagram illustrates the structure of a Geniom biochip. Within each biochip are eight individual microarrays, one of which is highlighted above, each made up of a single microchannel. Oligonucleotide features are organised into four columns within the microchannel. Black arrows show the direction of liquid flow through the microchannel, for example buffers. Liquid is introduced to each microchannel at alternating sides of the biochip (i.e. in subarrays 2, 4, 6 & 8 liquid is introduced on the left hand side and for the others it enters from the right), drawn through and leaves from the opposite side of the biochip.

The 5' end of the growing oligonucleotide chain is blocked by an MeNPOC ((R,S)-1-(3,4-(methylenedioxy)-6-nitrophenyl)ethyl chloroformate) protected phosphoramidite which prevents the addition of further nucleotides. Exposure to UV light results in the removal of MeNPOC, revealing a free hydroxyl group which is able to bind the next wave of nucleotides that are washed over the array. Each feature is selectively masked during UV light exposure, enabling the synthesis of each oligonucleotide to be carefully controlled.

The Geniom is a single colour system, meaning that it is not able to differentiate between fluorescence at different wavelengths on a single array. Therefore both

samples are labelled with the same fluorochrome, and each different sample hybridised to a separate subarray. During the detection process, the level of fluorescence from each feature on the array is assessed. These signal intensities represent the amount of labelled target DNA which has hybridised to each set of probes. By comparing the signal intensity from each region of the genome represented on the array between the two samples, differences in copy number can be determined.

The Geniom platform has several advantages over standard commercial oaCGH services. Firstly, it is an in-house system, which affords the user greater control over all aspects of the experiment than most commercial platforms can allow. Also, since array synthesis is carried out by the user, there is no minimum order size, unlike for many commercial arrays, which makes this a cost effective system that can accommodate alterations to array design at short notice. The entire experiment, from design to detection, takes about 3 days, making this a time efficient system. However, as the Geniom platform has not previously been used for oligo-array CGH, one of the aims of my research was to develop an optimised protocol for this technique. The following chapter describes the development of an optimised protocol for oaCGH on the Geniom platform.

3.2 Optimisation of oaCGH on the Geniom Platform

The stages involved in an oaCGH experiment can be briefly summarised as array synthesis, DNA preparation, hybridisation, washes and detection. Since the Geniom platform had not previously been used for oaCGH, optimisation of many stages of the process was required. An initial protocol for oaCGH was drawn together, largely based on the Febit protocol for gene expression studies on the Geniom platform, but also using information gained from review of the available literature and also from experience in our lab of using other microarray platforms including Nimblegen and Combimatrix arrays. For more detail on this protocol see Chapter 2 (Materials and Methods). This initial protocol was built upon as investigations progressed; so after each stage was optimised, the protocol was modified and the improved conditions used for subsequent experiments.

Initially the goal of our investigations was to achieve a visible level of hybridisation to the Geniom microarrays. Therefore, in most cases, progress towards an optimal protocol was evaluated by visually assessing the quantity of fluorescently labelled product hybridised to each array. Once the protocol was optimised, signal intensity data from different DNA samples was compared to determine whether the optimised protocol was able to reveal regions of CNV (described in Section 3.3).

3.2.1 Array Design

Previous investigations within our lab using Nimblegen oaCGH detected a region of putative structural variation on chromosome 12p13.31. To validate our optimisations on the Geniom microarray platform, we wished to use a subset of these probes which had demonstrated the successful detection of copy number variation. However, due to

significant differences between the two platforms in terms of probe length and feature capacity, it was not possible to use all of the same oligonucleotide probes. Nimblegen microchips contain ~132000 features with a maximum probe length of 80 nt, compared to 6776 features up to 60 nt which can be synthesised onto the Geniom biochip. However, analysis of the Nimblegen probes revealed that 92% were less than or equal to 60 nt in length, and so the probe list was filtered to leave only those within this size range. Within this subset of probes, particular regions of interest were chosen for inclusion on the Geniom arrays, for example a region of segmental duplication on Chromosome 12. Also included on the arrays were a number of control regions to enable us to evaluate the progress of our optimisations. These included probes from the defensin region on chromosome 8 which is known to be copy number variable, 'random' probes which contained no match in the human genome, Nimblegen probes for a 'normal' region of chromosome 13, a Y specific SRY region and three regions of the X chromosome corresponding to clones previously used as controls in a BAC arrayCGH experiment (Snijders *et al*, 2001). For a more information on the oligonucleotide probes included on the biochip see Chapter 2 (Materials and Methods).

3.2.2 DNA Quantity and Concentration

The Febit protocol recommends using at least 6µg of DNA for hybridisation to each subarray. This is a large amount, especially when considering that a whole biochip (consisting of eight subarrays) requires 48 µg of DNA. Therefore, we investigated the effect of using genomic DNA at concentrations of 2, 4 and 6 µg, to evaluate whether it was possible to achieve effective hybridisation using a lower concentration of DNA.

To evaluate and quantify our investigations, the overall spread of signal intensities was compared between arrays containing different concentrations of target DNA (Figure

3.2). In order to allow accurate comparison of the results from different arrays, background fluorescence was first removed. This was done by calculating the average signal intensity of the set of random probes (which do not contain a match in the human genome) for each array, and subtracting this value from the signal intensities of all other probes in the data set. Those with negative values were removed, leaving only those which showed positive signal intensities (likely to be a result of true hybridisations). The range of signal intensities produced from hybridisations using different concentrations of target DNA was compared.

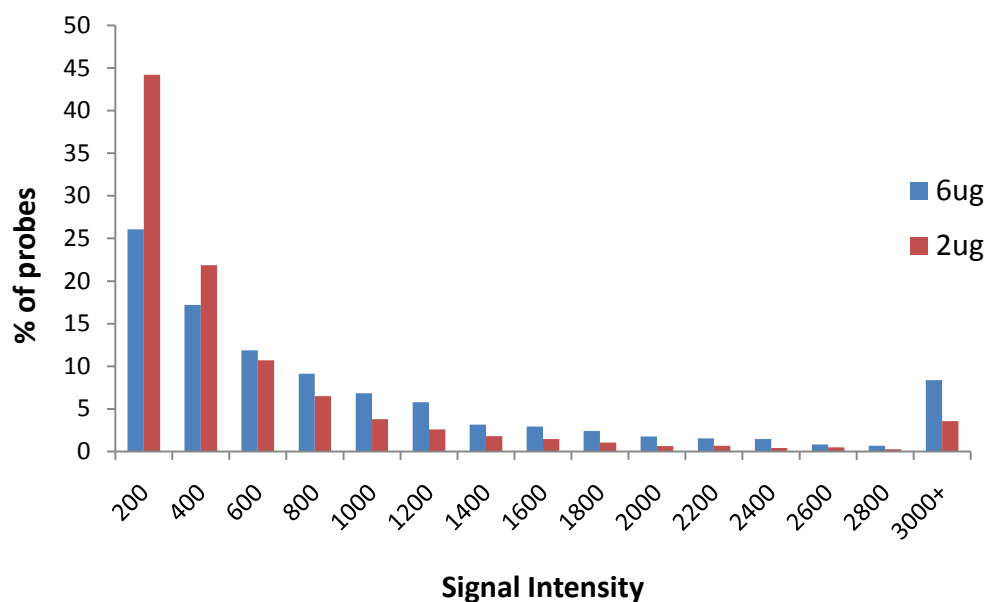


Figure 3.2: Relationship Between Target DNA concentration and Spread of Signal Intensities

Signal intensity data from hybridisations carried out using different target DNA concentrations were compared. This figure compares 6 µg and 2 µg target DNA concentration as an example. To evaluate the spread of signal intensities from each hybridisation, the frequency of data points falling into each of a range of signal intensity categories was calculated. This data was plotted as a bar chart, with different coloured bars representing each concentration of DNA.

Results show that there is a greater range of signal intensities when higher concentrations of target DNA are used, in particular there are over twice the number of probes with signal intensities over 3000 when the concentration of DNA is increased from 2 μg to 6 μg . This indicates that, as Febit suggest, hybridisation is more efficient when a higher concentration of DNA is used. Therefore, all subsequent experiments described in this chapter use 6 μg of target DNA.

3.2.3 Optimisation of Labelling Conditions

The efficiency of target DNA labelling is a significant limiting factor to the success of an oaCGH experiment. Initially a labelling protocol was developed based on a frequently cited method first described by Snidjers *et al* (2001) (See Chapter 2). Optimisations were carried out to increase labelling efficiency by altering variables including reaction time and reagent concentration, as we postulated that these had the potential to significantly limit reaction efficiency.

Labelling efficiency was assessed using a Biotin Chromogenic Detection Kit (Fermentas Life Sciences). A series of dilutions of the labelled genomic DNA were made, based on the concentration of DNA in solution before labelling, ranging from 1000 pg/ μl to 0.01 pg/ μl , and an aliquot of each dilution was spotted onto chromatography paper. After a series of washes, one of which contains streptavidin to bind the biotin, the blot was left to develop overnight. The streptavidin is coupled to alkaline phosphatase and the developing solution contains a chromogenic substrate for this molecule which produces a purple precipitate where biotin-labelled DNA is present, allowing visualisation and quantification of labelled DNA molecules. The manufacturer suggests that the reaction efficiency is acceptable if the 0.1 pg/ μl spot is clearly visible after an overnight development, so this was used as a guide to assess the

progress of our optimisations. Image J analysis software was used to quantify the intensity of each spot on the chromatogram, to allow for a more accurate comparison.

3.2.3.1 *Reaction Time*

The length of time provided for a DNA labelling reaction to take place must be sufficient to ensure there is enough time for all of the DNA molecules to be successfully labelled. Our initial protocol involved a 3 hour incubation step. To investigate whether this was the optimum reaction time, we compared the yield of labelled product produced from 3, 6, 8 and 16 hour incubations (Figure 3.3). Results show that a 3 hour incubation step is not sufficient; when the incubation time is increased this results in a considerably higher yield of labelled product. An incubation time of 16 hours increases the lowest dilution visible on a chromogenic blot from 10 pg/μl (as a result of a 3 hour incubation) to 0.1 pg/μl

3.2.3.2 *Concentration of Nucleotides*

It is essential that a DNA labelling reaction contains a sufficient concentration of each reagent to label the full amount of DNA present. In particular, the concentration of nucleotides and enzyme (discussed in Section 3.2.3.3) were identified as potential limiting factors to labelling efficiency.

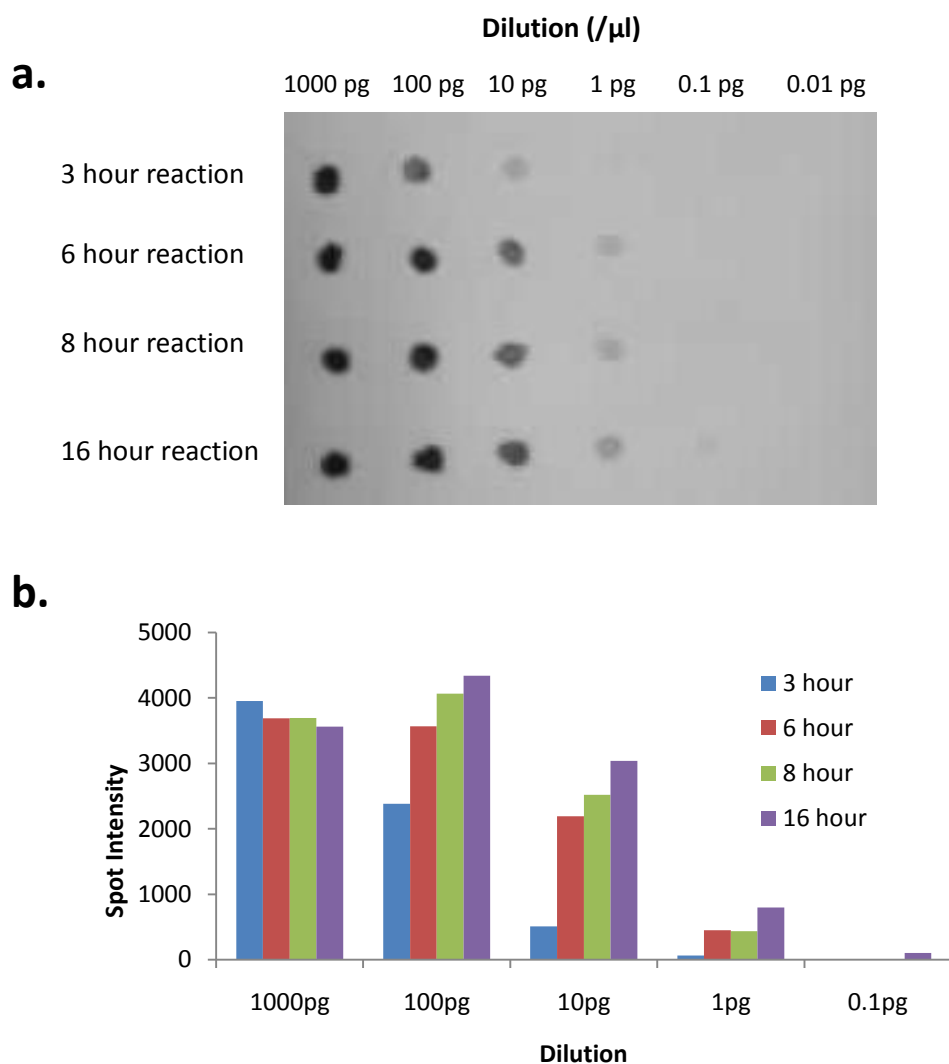


Figure 3.3: Relationship Between Length of Reaction and Yield of Labelled Product

The effect of changing reaction time on yield of labelled product was visualised using a biotin chromogenic detection method (Fermentas) and quantified using ImageJ. a.) Photograph of a chromogenic blot. Each dot represents a different dilution of DNA, based on the concentration of DNA before labelling. The dilutions become 10x more dilute as the spots move from left to right on the blot. The quantity of labelled product is visualised after leaving the blot to develop overnight. b.) The software ImageJ was used to quantify the intensity of each spot on the chromatogram, the results of which are shown here on a bar chart. For each of the four reaction times, the intensity of the spot for each dilution factor is calculated, to represent the amount of labelled product present. Each coloured bar represents a different reaction time, as shown by the key on the graph.

The initial labelling protocol requires a nucleotide concentration of 350 nmol. To investigate whether an increase in nucleotide concentration would increase the yield of labelled product, we doubled the concentration to 700 nmol (Figure 3.4). Results show that this change in nucleotide concentration resulted in a considerable improvement in labelling efficiency, with an increase in spot intensity in all visible dilutions.

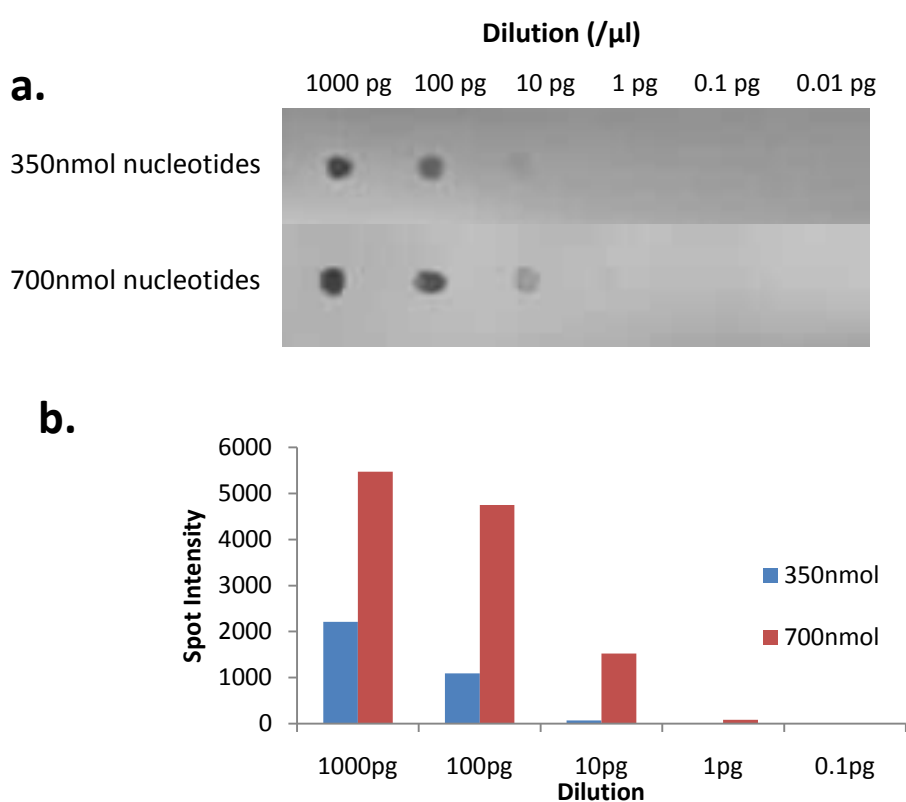


Figure 3.4: Measuring the Effect of Nucleotide Concentration on Labelling Efficiency

The effect of increasing the concentration of nucleotides present in a labelling reaction from 350 nmol to 700 nmol was compared. a.) Photograph of a chromatogram used to determine the yield of labelled product for each nucleotide concentration and b.) the amount of labelled product present in each spot was quantified using the software ImageJ. For more details see the legend for Figure 3.3.

3.2.3.3 *Addition of extra enzyme*

The amount of enzyme available for a DNA labelling reaction decreases as the reaction progresses, due to denaturation of the enzyme. Therefore the rate of labelling is higher at the start of a reaction than toward the end, which can be a significant limiting factor to reaction efficiency. In an attempt to overcome this, the effectiveness of adding extra enzyme to the labelling reaction midway through the incubation step was assessed. This investigation was carried out using the current optimal protocol, which had been adjusted to incorporate the results from previous studies; so at this point it involved a 16 hour incubation step and a nucleotide concentration of 700 nmol.

An additional aliquot of enzyme was added to the reaction mix either once or twice during the incubation step. Results showed that the biggest increase in labelling efficiency occurs when extra enzyme is added twice during the labelling process (Figure 3.5). For convenience, the extra enzyme was added 3 hours after the start of the reaction and then again after 6 hours.

3.2.3.4 *Optimal Labelling Protocol*

An optimal protocol for labelling DNA with biotin was produced using the results of our investigations described above. This method involves an excess of reagents, a 16 hour incubation step and addition of extra enzyme at two points during the reaction. Initially the 10 pg/ μ l spot was the lowest dilution visible on the chromogenic detection blot, however after optimisation of the reaction conditions, it was possible to obtain a clear and reproducible spot from the 1 pg/ μ l dilution (Figure 3.6). In some cases it was also possible to see a spot from the 0.1 pg/ μ l dilution (for example, Figure 3.3), though this was not reproducible.

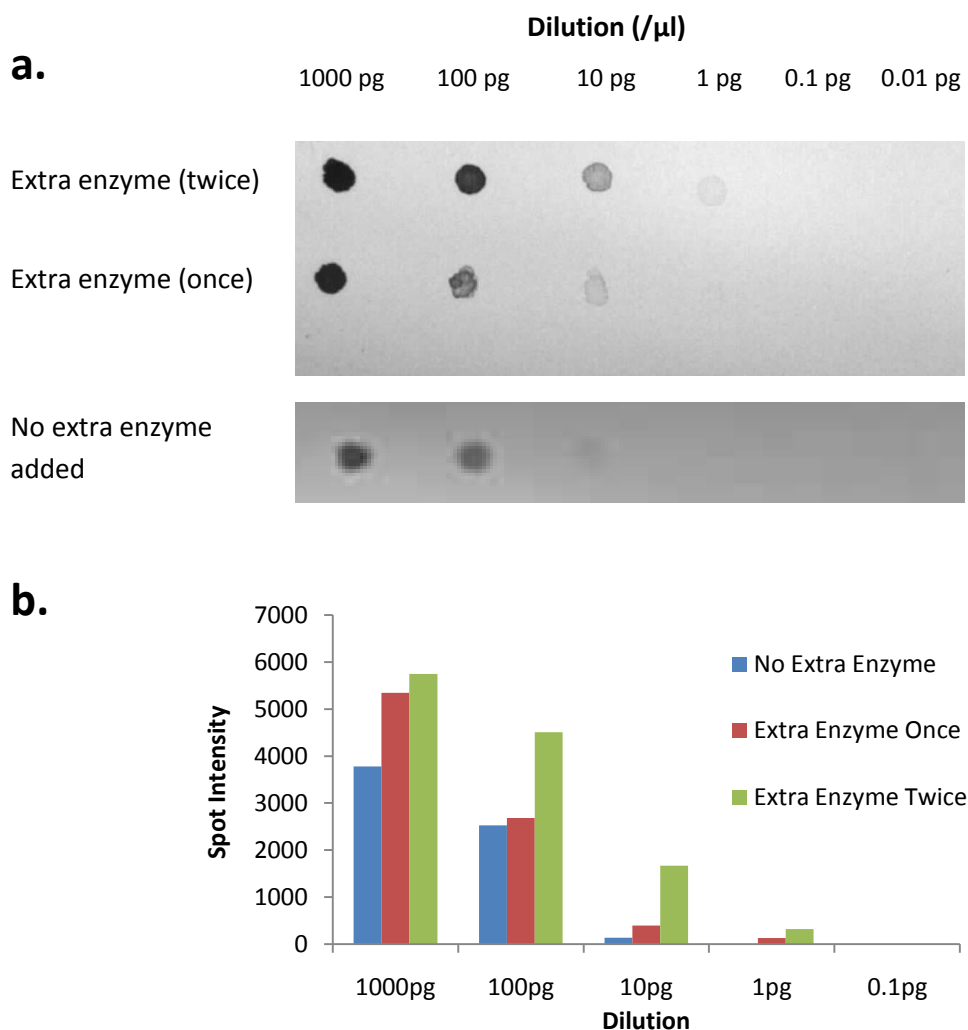


Figure 3.5: Addition of Extra Enzyme During Labelling Reaction

The effect of adding an additional aliquot of enzyme to the labelling reaction during the incubation period was investigated. a.) Photograph of a chromatogram used to determine the yield of labelled product for each reaction and b.) the amount of labelled product present in each spot was quantified using the software ImageJ. For more details see the legend for Figure 3.3.

This improvement represents an increase in labelling efficiency which we estimate to be around 10 fold. It is not possible to be more accurate than this, due to the fact that the chromatographic detection method used here is not truly quantifiable. This improvement in labelling efficiency should lead to a considerable increase in visible hybridisation on the Geniom biochip.

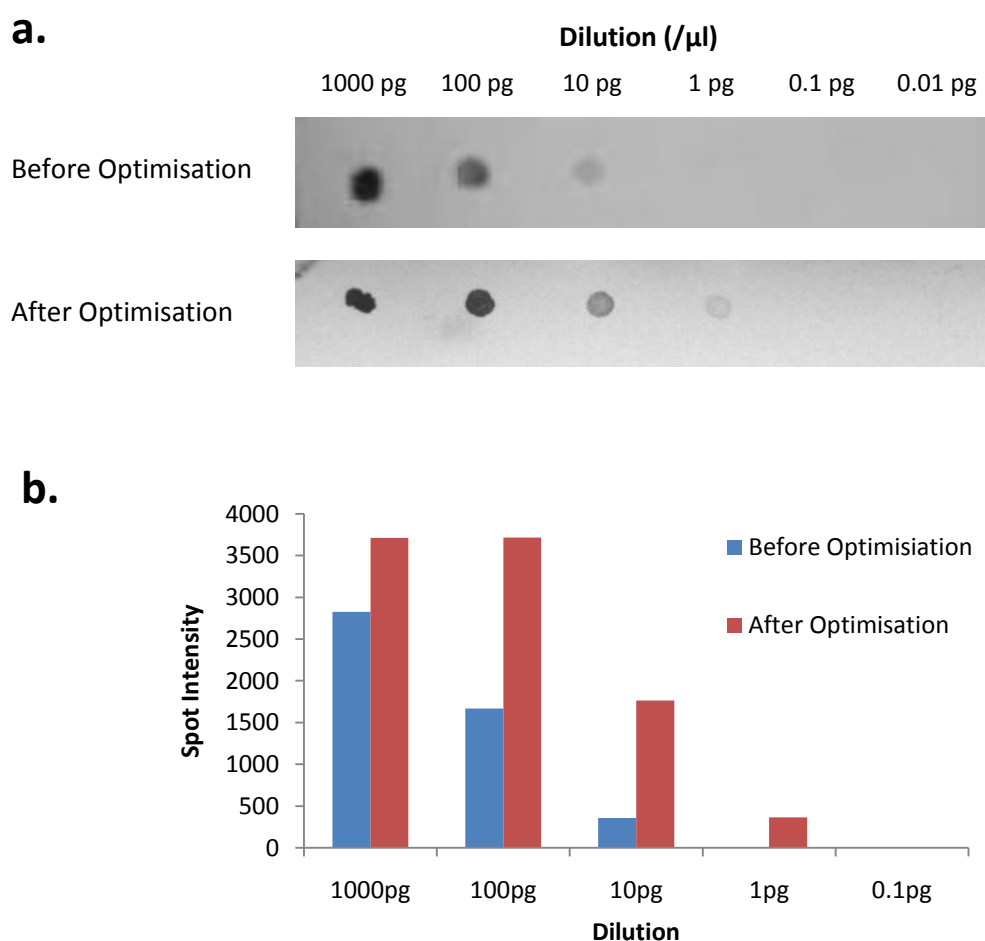


Figure 3.6: Results of Labelling Optimisation

A labelling blot from a reaction using the optimum conditions described in this section is compared with one from a reaction using the initial protocol (see introduction for more details). a.) Photograph of a chromatogram used to determine the yield of labelled product shown for labelling reactions before and after optimisation and b.) the amount of labelled product present in each spot was quantified using the software ImageJ. For more details see the legend for Figure 3.3.

3.2.4 Hybridisation Time

The length of time provided for the target DNA to hybridise to the oligonucleotide probes on the biochip is directly related to the level of hybridisation which will occur. A short hybridisation time (for example 3 hours) may be enough for high copy number sequences, or those that happen to be near their target sequence, to hybridise; however,

due to the microchannel format, we would not expect this to be long enough for most target sequences to hybridise to their specific probes. Longer hybridisations (for example 92 hours) may result in more specific hybridisations, but will also allow more time for non-specific hybridisations to occur. It is possible that a long hybridisation could also result in denaturation of hybridised sequences, and therefore reduce the level of detectable hybridisation.

The effect of a range of different hybridisation times was investigated (Figure 3.7). Results suggest that 3 and 24 hour hybridisation times allow for the highest levels of observable hybridisation. Longer hybridisations, for example 92 hours, were more likely to contain smearing within the microchannels. The array with a hybridisation time of 16 hours shows less hybridisation than either 3 or 24 hours.

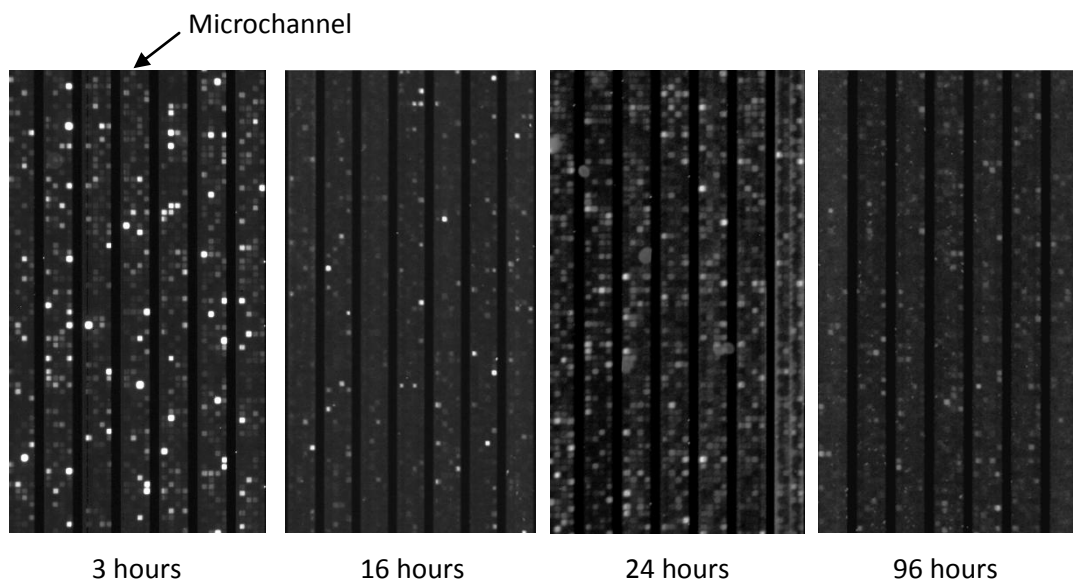


Figure 3.7: Comparison of a Range of Different Hybridisation Times

Detection images from four subarrays are shown, representing a section of each subarray, each of which used a different hybridisation time. Hybridisation is visible as white spots on the arrays.

It is clear that there is a high degree of variation between different hybridisation times, and that increasing the length of a hybridisation does not necessarily increase the level of hybridisation by an equivalent amount.

Comparing the results from the 3 hour and 24 hour hybridisations it can be seen that the 24 hour array contains a higher level of background fluorescence and also some smearing within the microchannels. Therefore, somewhat surprisingly, we concluded that in the case of the Geniom platform, 3 hours appears to be an adequate time to achieve a visible level of hybridisation.

3.2.5 Adjusting the Stringency of Hybridisation

The stringency of hybridisation determines the specificity, rate and percentage of single-stranded DNA molecules which hybridise to each other. A balance is required to achieve optimum stringency, where non-specific hybridisation is slowed but specific oligonucleotide binding is retained. Stringency can be altered by a number of factors; for our optimisations we have concentrated on adjusting stringency by changing the concentration of formamide in the hybridisation solution. Formamide is a solvent which favours the denaturation of DNA by lowering the melting temperature of bonds between strands. Febit recommend using formamide at a concentration of 40% in the hybridisation solution only.

We investigated the effect of including formamide at a range of concentrations in the hybridisation buffer (Figure 3.8). Results show that reducing the concentration of formamide in the hybridisation solution from 40% to 25% appears to increase the presence of a 'smearing' effect within the microchannels. In contrast, a concentration of 50% formamide seems to be too stringent, as it considerably reduces the level of visible

hybridisations. In order to achieve a compromise between these two scenarios, we decided that the concentration of formamide included in the hybridisation solution should remain at 40% as recommended by Febit. In the future, it may be desirable to carry out further optimisation of hybridisation stringency in order to refine the specificity of oaCGH hybridisations.

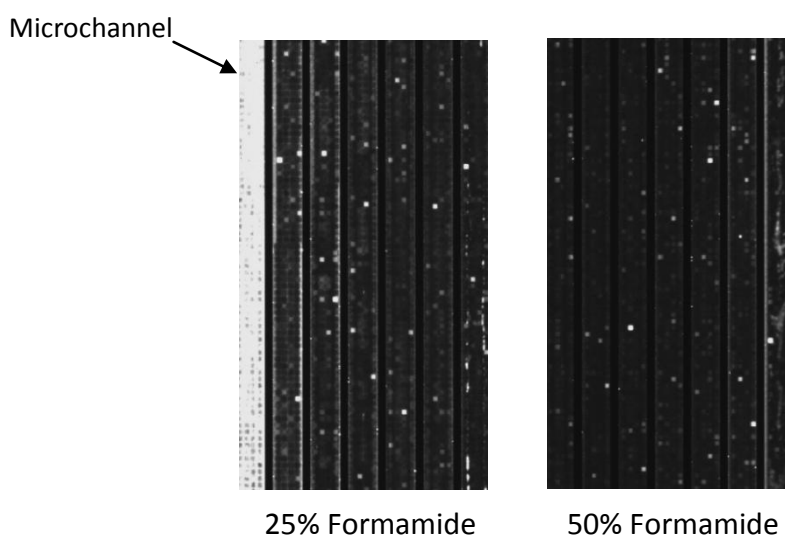


Figure 3.8: Inclusion of Formamide in Hybridisation Solution

Detection images for arrays in which formamide was included in the hybridisation solution at a concentration of 25% and 50% are shown. Each image shows a section taken from different hybridisations.

3.2.6 Mixing During Hybridisation

One of the novel features of the Geniom microarray system is that the oligonucleotide features are located within microchannels. In other microarray platforms, for example CombiMatrix arrays, the features are located on the surface of the microchip and therefore solution-phase reactants are able to easily come into direct contact with the entire surface of the array. In order to further aid hybridisation, the solution may be

mixed by rocking during hybridisation (www.combimatrix.com). In the case of the Geniom platform, we were concerned that the microchannel format could prevent the target solution from having access to all the probes on the array, thereby hindering the target DNA from accessing its complimentary probes, and reducing the likelihood of successful hybridisations.

In an attempt to overcome this problem, mixing was carried out during hybridisation by pumping the solution back and forth within the microchannels using a number of different methods. The aim of this was to allow the target solution to come into contact with as much of the array surface as possible. In order for this to occur, we determined that the most important mixing parameters to consider were how often mixing should take place, and how to achieve maximum mixing without causing the arrays to dry out.

The effectiveness of using both manual and mechanical methods of mixing buffer within the microchannels was evaluated. Manual mixing was carried out using a syringe to pull the liquid out of the array and then back through the hybridisation chamber, taking care not to remove it too far as this could dry out the array. However, it was difficult to obtain consistency using this method, and it was not practical to carry out mixing at regular intervals throughout an overnight hybridisation. In an attempt to overcome these difficulties, we modified a rocking platform (Grant-bio) to create a custom-made pump, which pushed down on a series of pipette bulbs and used the changes in air pressure this created to move liquid through the microchannels. This allowed for a slower and more consistent pumping speed than was possible from manual mixing. However, mixing (both with a syringe and with the modified rocking platform) also brought about further problems by increasing the appearance of a ‘smearing’ effect within the microchannels (Figure 3.9). The more times the solution was mixed, the higher the level of smearing which occurred.

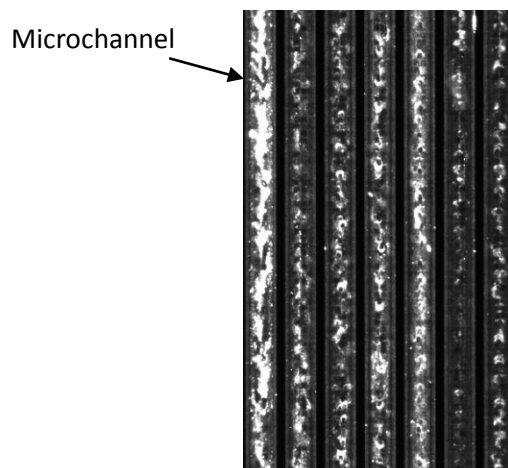


Figure 3.9: Mixing Hybridisation Solution During Hybridisation

A detection image is shown from a section of a subarray which was mixed at twenty intervals during hybridisation, using a syringe. This resulted in the appearance of a smearing effect within the microchannels.

3.2.7 Smearing within Microchannels

Throughout optimisation, one of the most significant challenges we faced was achieving consistent levels of hybridisation. This varied substantially, between sub-arrays on the same biochip as well as between biochips. As optimisations progressed, it became apparent that on many occasions a build-up of a DNA aggregate occurred during hybridisation, leading to the appearance of a smearing effect within the microchannels. On many occasions the presence of smearing meant that any hybridisations were masked or undetectable, invalidating the experiment. This proved to be a considerable limitation to the success of our hybridisation efforts. We noticed that adjusting a number of the reaction variables had an effect on the degree of smearing which was present within the microchannels (Table 3.1).

Table 3.1: Conditions which enhance Network Formation

Variable	Effect	Possible Explanation
Concentration of DNA	Increasing DNA concentration increases smearing	More DNA present to form networks
Hybridisation Time	Longer hybridisations increase smearing	Longer hybridisations allow more time for networks to form
Formamide Concentration	Higher formamide concentrations reduce smearing	Formamide encourages denaturation so may cause networks to break down
Mixing	Mixing increases smearing	More DNA comes into contact with building networks

These results lead us to hypothesise that the smearing effect could be a result of the formation of networks, made up of multiple fragments of DNA which non-specifically hybridised to each other within the microchannels (Figure 3.10). During the wash steps it is possible that these networks could be pulled through the microchannels, the force of which could lead to the disruption of the networks as well as specific hybridisations.

3.2.8 Digestion of DNA with Restriction Enzymes

Our hypothesis is that smearing within the microchannel of each array occurs as a result of networks forming between multiple molecules of DNA. Therefore, reducing the size of the DNA fragments should lead to a decrease in the frequency and size of network formation. Up until this point, all DNA samples had been digested with the restriction enzyme *DpnII* prior to labelling. To test whether the size of the DNA fragments effects the level of network formation, the effect of incorporating a double digest into the DNA preparation stage was investigated.

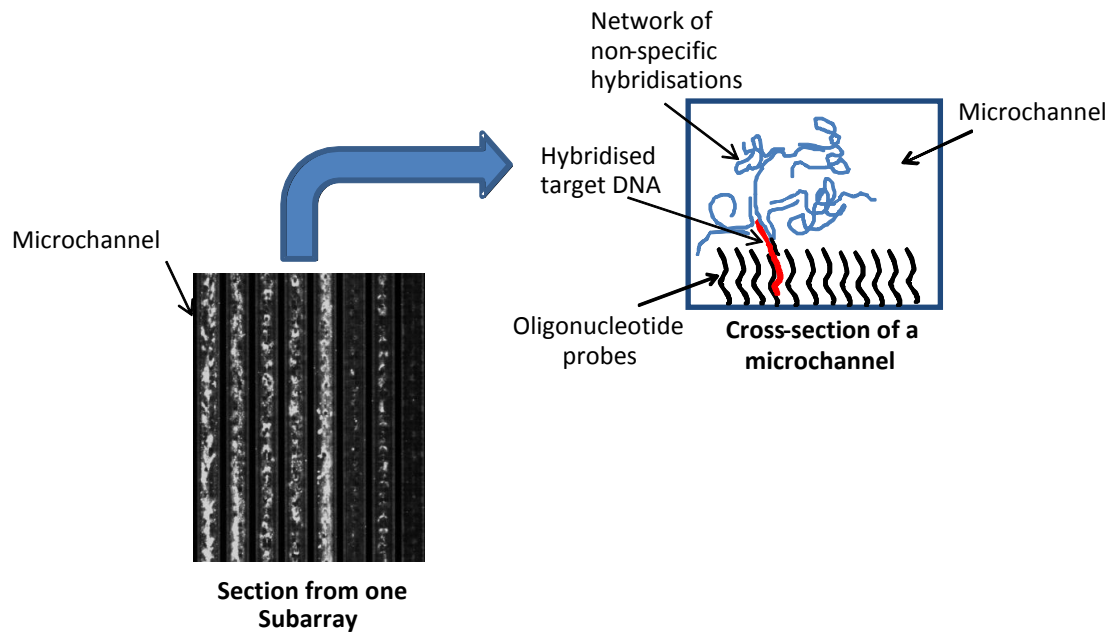


Figure 3.10: Network Formation May Lead to 'Smearing' within Microchannels

A detection image of a section of a subarray is shown, within which a smearing effect can be seen. The cartoon shows a cross-section of a microchannel, illustrating a possible cause of this effect. Black lines represent oligonucleotide probes, whereas the red line shows a fragment of target DNA which has formed a specific hybridisation to one of the probes. The blue lines show a network of non-specific hybridisation, which has built up as an extension of the specific hybridisation. This diagram shows a network built from one hybridisation only; we hypothesise that networks may build up from any number of separate hybridisations.

All samples were first digested with *DpnII* using the usual protocol and then some were taken forward for a second round of digestion with *AluI* (For more detail of protocols see Chapter 2, Materials and Methods). After hybridisation there appeared to be significantly less smearing in arrays where DNA had been subject to a double digest compared to a single digest (Figure 3.11). This suggests that using smaller fragments of target DNA is effective at reducing network formation within the microchannels, thereby reducing the appearance of smearing. Further investigations are required to assess the relative merits of using different combinations of restriction enzymes for the double digest; however, such studies were outside of the scope of this project.

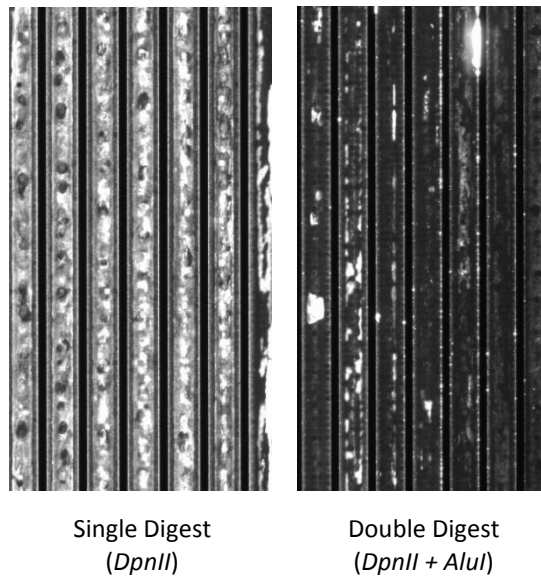


Figure 3.11: Fragmenting DNA with a Double Rather than Single Digest

Sections of detection images from two subarrays are shown. In addition to *DpnII*, target DNA hybridised to the second array underwent a secondary digestion with *AluI*.

3.2.9 Summary of Development of oaCGH on the Geniom Platform

In this section, progress towards an optimum protocol for oaCGH on the Geniom microarray platform has been described. The developments so far are summarised below in Table 3.2.

3.2.10 Probe Optimisation

Optimal design of the oligonucleotide probes which are synthesised onto the Geniom biochip is important to allow accurate detection of copy number changes. Factors to consider in regards to probe design which can affect hybridisation include length, GC content and whether the probe matches (or is highly similar to) other regions of the genome. These variables can affect the degree of specific and non-specific hybridisations which occur.

Table 3.2: Summary of Conditions used for oaCGH on the Geniom Platform

Variable	Optimum
DNA Concentration	6 µg
Labelling Conditions	700nmol nucleotides, 16 hr reaction, addition of extra enzyme after 3 and 6 hours
Hybridisation Time	3 hrs
Formamide Conc. in Hybridisation Solution	40%
Formamide Conc. in Wash Buffers	0%
Mixing During Hybridisation	No
Digest of DNA	Double digest

Initially the majority of the probes used on the Geniom arrays were taken from the Nimblegen microarrays which had been used for preliminary oaCGH experiments. These were filtered for length, since the Geniom biochips can synthesise up to a maximum length of 60bp, compared to 80bp on a Nimblegen array. Other probes, for example those within the many control regions, were created by splitting the sequence of interest into fragments of 60bp. These were repeat-masked prior to fragmentation and filtered to remove any repeat regions. Other than this, no filtering was carried out initially, since the aim was to subsequently assess probe performance and use the results of this analysis to identify the optimal factors for probe design.

3.2.10.1 *GC Content*

The guanine and cytosine base content (GC content) of the oligonucleotide probes is an important factor for consideration since it influences properties of a probe such as melting temperature. Base pairings formed between G and C residues are held together

by three hydrogen bonds rather than the two formed between the other DNA bases (A and T), making them stronger. Probes with a low GC content are likely to show little hybridisation, whereas probes with a high GC content may bind strongly to even non-specific sequences. In an ideal situation, all probes will be within a specified GC range, to allow for optimal specific hybridisation. However, this is not always possible as some regions of interest may have an unusually high or low GC content.

In order to assess the affect of GC content, the percentage of GC bases within a subset of chromosome 12 probes on one subarray was compared with the natural log of the signal intensity for the same probe (Figure 3.12).

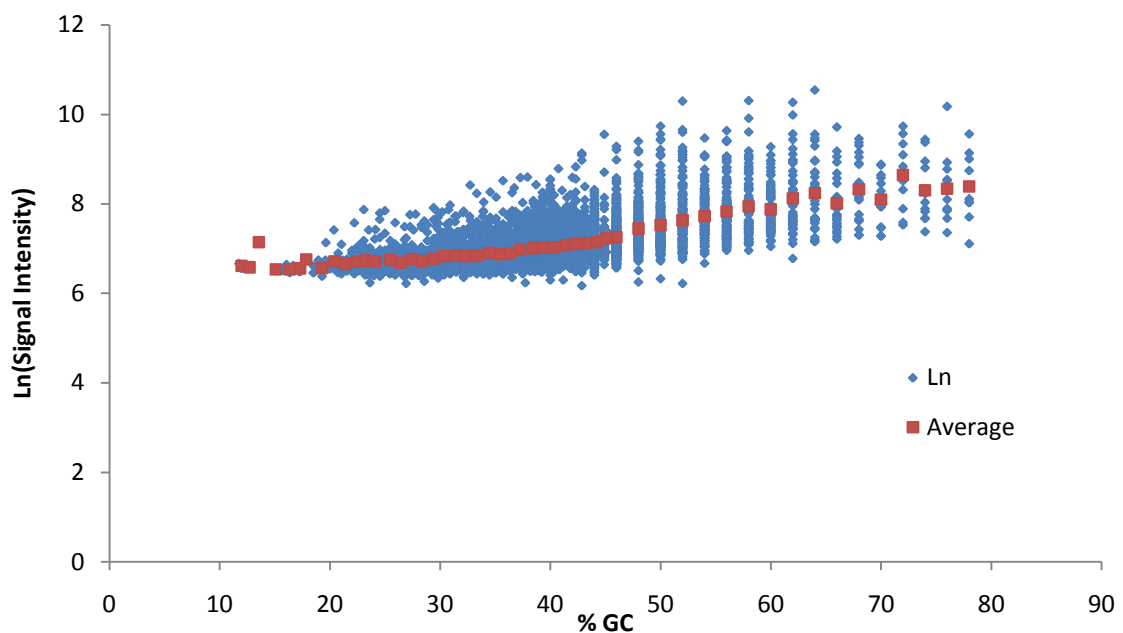


Figure 3.12: Relationship Between Probe GC Content and Signal Intensity

The percentage GC content (number of nucleotide bases in each oligonucleotide which are either G or C) of each of a set of Chromosome 12 probes was calculated and plotted against the natural log (Ln) of the signal intensity for each probe from a single hybridisation experiment (data shown in blue. For each GC value, the average Ln(signal intensity) was also calculated, this is shown in red on the same graph.

There appears to be a correlation between GC content and signal intensity. As expected, at a low GC there is little hybridisation. The average signal intensity increases with GC content. However, what is interesting is that even at a higher GC content, there is a wide range of signal intensities. This indicates that even if probes were designed to be within a certain GC range, there may still be considerable variation in signal intensities between probes of the same GC content. This could be a result of other factors, for example whether the GC bases are organised into clusters, or spread across the probe. Further analysis, for example close examination of the probes sequences, could be carried out to investigate this further.

3.2.10.2 *Number of BLAT hits*

Another factor which was investigated was the ‘uniqueness’ of the probe. Although each 60bp probe was in its entirety unique, most contained short sequence matches which aligned elsewhere in the genome. The degree to which this occurred for each probe was assessed by performing BLAT alignments (after repeat masking had been carried out to remove repetitive elements) and recording the number of short sequence matches returned for each oligonucleotide. The greater the number of matches, the more chance there is that the probe would bind to sequences other than the sequence of interest. Therefore ideally probes should be unique, or have as few matches as possible, in order to avoid non-specific hybridisations. Preliminary analysis of the probes included on the Geniom biochip revealed that although repeat-masked probes typically contained between 0 and 2 matches equal to or greater than 30bp in length, there was a wide range in the number of shorter sequence alignments (less than 30bp). In order to minimise the effect of non-specific hybridisations, probes were filtered during data analysis to leave only those with fewer than 5 short (between 11 and 30 bp) sequence

alignments (Section 3.3.3). Further analysis of the probes would be required to determine to what degree those with a high number of short sequence alignments are affected by non-specific hybridisation.

3.3 Data Analysis and Identification of Structural Variation using the Geniom Platform

Evaluation of our progress towards an optimal protocol for oaCGH using the Geniom microarray system has so far concentrated on assessing the overall signal intensity levels produced from hybridisations carried out using a range of different conditions. However, in order to assess whether our protocol for oaCGH is able to successfully detect copy number variation, it is also necessary to compare signal intensity data from hybridisations carried out using the same reaction conditions, but different DNA samples.

In the example of this analysis given below, a single biochip was selected on which oaCGH had been carried out using our optimum conditions, with the exception that a single rather than double digest was used to fragment the DNA samples. This biochip was also deemed to be the best in terms of a high level of hybridisation and lack of smearing. Within this biochip, subarrays 1 and 2 were chosen for comparison (Figure 3.13). These were selected because the DNA samples hybridised to these subarrays were known to contain a different number of copies of the defensin region on chromosome 8 (7 and 4 respectively), probes for which were included on the array to act as controls. Before two data sets could be directly compared, they were first corrected for background fluorescence and normalised (see section 3.3.2 below). The optimised parameters for these are described below.

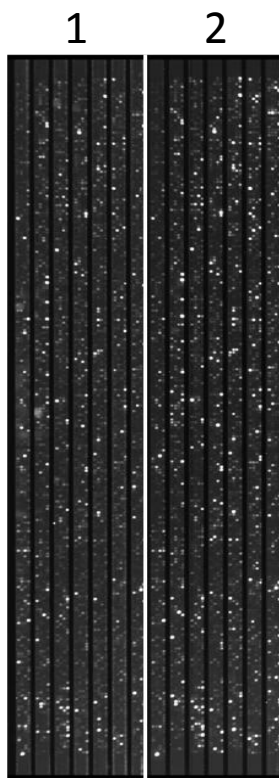


Figure 3.13: Detection Image of Geniom Biochip Used for Analysis

The above image is the resulting picture after the detection of fluorescence from Geniom biochip 16. The two subarrays are separated in this image by white lines and numbered. Out of all the hybridisations carried out during our optimisations, biochip 16 showed the highest level of hybridisation with the lowest levels of background fluorescence and smearing. This hybridisation was carried out the conditions laid out in Table 3.2.

3.3.1 Removal of Background

Background fluorescence originates from sources other than the specific hybridisations of interest, for example non-specific hybridisations of labelled DNA, or substances other than the fluorophore of interest. The level of background fluorescence may differ between arrays and also in different areas of the same array. It is important that background is removed before analysis, as it prevents determination of the true value of fluorescence (signal intensity) which occurs as a result of specific hybridisations. The level of background can be calculated a number of ways, for example using the level of

fluorescence emitted from an area of the array outside a feature, a feature containing no DNA, or a reference probe for which no true match is included in the target DNA. Since there is likely to be variation in background between different areas of the array, it is advisable to determine the background fluorescence level using several points and then calculate the mean background value of these. For analysis of the Geniom data, background was calculated using 'random' probes which were included on the arrays but had no matching sequence in the human genome. Since no specific hybridisation should occur to these probes, any fluorescence detected can be considered background fluorescence. The mean signal intensity of the random probes was subtracted from the signal intensities of the rest of the features.

3.3.2 Normalisation

Before the data sets from two different arrays can be compared, normalisation of the data is carried out to remove any signal variability. This may occur due to differences in factors such as labelling efficiency or probe position. There is no single method of normalisation. The simplest methods involve adjusting for differences between the average signal intensity of the two sets of data. For example, the ratio between the signal intensity means of two arrays may be calculated and the data sets then adjusted accordingly to bring the means to the same level. This makes some allowance for the presence of hybridisation differences and allows the data to be compared without significant bias towards the data set with the highest mean of signal intensities. However, such approaches correct each data point in the same way and by the same proportion. These methods are only suitable for use with single colour systems where it is not necessary to take into account dye bias. Other normalisation techniques have been developed for dual colour systems, for example Lowess (Yang *et al*, 2002), which

corrects the data points independently by assuming that the degree of bias which occurs is dependent upon the signal intensity.

In the case of the Geniom, the system is single colour and therefore is not affected by dye bias. The main issue to consider was differences in hybridisation efficiency between the arrays. A simple global approach to normalisation was applied. The standard deviation of signal intensities from each array was calculated to determine differences between the spread of signal intensities in the two arrays. To correct for this difference, the signal intensity data in the array with the lowest standard deviation (in this case array 2) was multiplied by the ratio of standard deviations. Also, in an attempt to bring the base value of the data to zero, the minimum value in each array was subtracted from the rest of the signal intensity data.

3.3.3 Displaying oaCGH Data Using a Logarithmic Scale

oaCGH data comparing two DNA samples is typically displayed as a \log_2 ratio on a scatter plot, where the vertical axis represents the \log_2 value and the horizontal axis shows the relative position of the oligonucleotide features. \log_2 ratio is used as a convenient scale because it means that an increase and decrease in copy number by 1 (i.e. a doubling and a halving of relative intensity) are indicated by values of 1 and -1 respectively. The \log_2 ratio of the signal intensity data from the two arrays compared in this example was calculated and viewed on a scatter plot (Figure 3.14). However, due to a high level of noise in the data, it was difficult to locate regions of normal copy number. Even after normalisation, the base value of the data is still below zero. Therefore it was not possible to accurately distinguish regions of copy number variation, even in the defensin region where the two samples were known to vary.

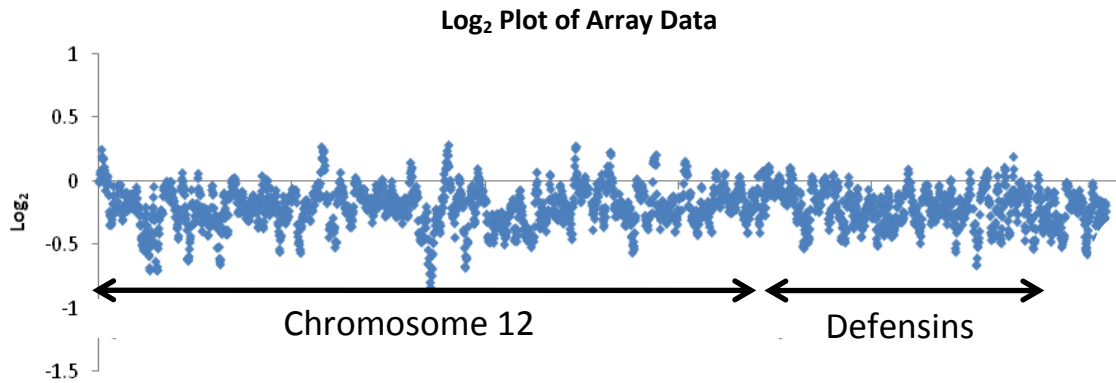


Figure 3.14: Log₂ Plot Comparing Relative Hybridisations from Two DNA Samples

The signal intensity data from two DNA samples hybridised to two different subarrays on the same Geniom biochip was compared. The log₂ ratio of signal intensities from each oligonucleotide feature was calculated, and displayed on a scatter graph. The Y axis shows the log₂ ratio, and the X axis displays the relative location of the probes. Due to the level of noise in the data, it is not possible to determine the 'normal' level of hybridisation, which makes it impossible to reliably distinguish regions of copy number variation. Even after normalisation, the base value of the data is still below 0.

In an attempt to overcome these difficulties, data were compared using a different approach. The natural log of the raw signal intensities from each array was calculated and then the data normalised as before. The chromosome 12 probes were also filtered to exclude probes which contained short sequence matches to more than four regions of the genome, as identified using a BLAT alignment, in an attempt to remove probes which may be susceptible to non-specific binding. When the signal intensity data from each array were viewed individually on a scatter plot, it was observed that results appeared to follow the same pattern for both arrays (data not shown). Therefore, rather than calculating a ratio of the signal intensities between the two arrays, the data from each array was represented individually as two different tracks on the same graph (Figure 3.15).

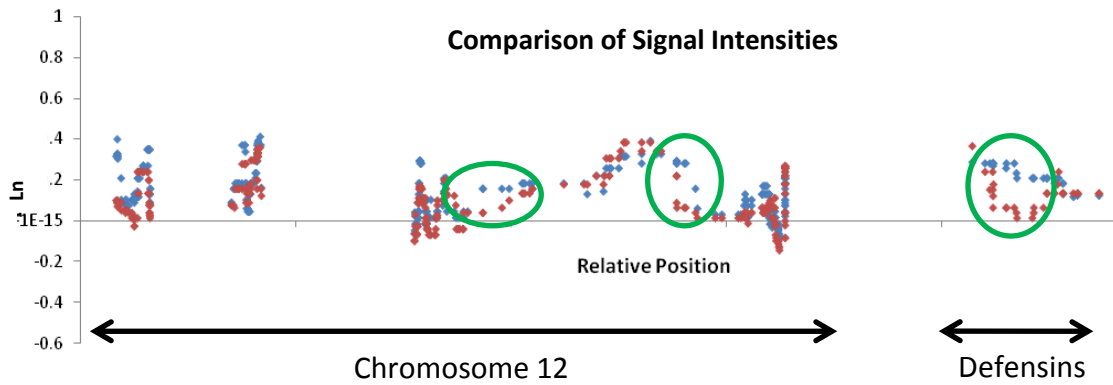


Figure 3.15: Identifying Structural Variation Using the Geniom Platform

Signal intensities from two DNA samples hybridised to subarrays on the same Geniom biochip were compared. For each sample, the natural log of the signal intensity from each probe was calculated and plotted individually onto a scatter plot. Probes were also filtered to exclude those which matched to more than four regions in the genome, as identified using a BLAT alignment. Blue dots indicate data from array 1 and red dots array 2. Results from probes corresponding to sequence in the Chromosome 12 and defensin regions are organised according to the relative genomic position of the oligonucleotide probes. It can be seen that for the most part the data from both samples closely follows the same pattern, except for three regions, which are highlighted by green circles. These represent differences in copy number between the two DNA samples.

It can be seen that, for the most part, the two data sets closely follow the same pattern, except for three regions which are highlighted. These represent regions which differ in copy number between these two samples. One of these regions is located within the defensin genes, which are known to be highly copy number variable.

The samples used here have previously been genotyped, and the sample in array 1 is known to contain an increased copy number in the defensin region, which is reflected here in the oacGH data. The two other regions which appear to vary in copy number are located on chromosome 12. The fact that we have been able to identify regions of known copy number using our optimised oacGH protocol is a validation of our technique.

3.4 Summary & Discussion

This chapter describes work carried out in order to develop an optimised protocol for oaCGH on the Geniom microarray platform. After extensive optimisation of a number of variables including labelling efficiency, DNA concentration and hybridisation conditions, we were able to use the modified protocol to detect regions of copy number variation.

3.4.1 DNA Concentration and Labelling

We concluded that the optimum amount of input DNA for each sub-array is 6 µg. This is a considerable quantity of DNA, and may be a limiting factor as the amount of DNA available for such experiments is often limited, for example when the source is tissue samples. In an attempt to overcome this issue, it was considered practical to incorporate a pre-amplification step prior to labelling. Several studies (including Barker *et al*, 2004 and Pinard *et al*, 2006) have suggested that any amplification bias that this process may have is reproducible, therefore providing both samples are treated in the same way any bias should not affect the results. It is also interesting to note that a number of commercial systems, including Agilent arrayCGH and Illumina genotyping arrays, already incorporate a whole-genome amplification step into their protocols (Brueck *et al*, 2007; www.illumina.com/support/faqs.ilmn).

Given the large quantity of input DNA required for microarray analysis, it is surprising that the two labelling protocols we drew (from Febit's protocol and Snijders *et al*, 2001) were both relatively inefficient. We have described an optimal labelling protocol which achieved an increase in efficiency of around 10 fold compared to these methods, and therefore prevents wastage of precious and often limited DNA samples.

3.4.2 Identification of Structural Variation

Using our optimised protocol, it was possible to obtain enough clean arrays (without the presence of smearing) to allow data analysis to be carried out. For this process two subarrays were chosen. The conventional method used to compare oaCGH data from two DNA samples is to calculate the ratio of signal intensities for each probe between the two samples, and plot this on a scatter diagram. However, when we tried this approach it was not possible to identify regions of structural variation, even in the defensin region where we knew there to be a difference, due to the level of noise present. Therefore a different approach was taken, and the signal intensities from each data set were individually plotted on a scatter graph. For the most part the two data sets closely followed the same pattern, however at a number of points they differed, indicating the presence of structural variation. It is interesting to note that the conventional approach used to analyse and visualise oaCGH data was unable to reveal copy number variations in our data. This may warrant further investigation, as it could be that some regions of CNV are being routinely missed in other analyses which use this method.

3.4.3 Network Formation

A number of the unique features of the Geniom platform meant that optimisation of oaCGH on this system was more challenging than expected. For example, the microchannel format makes it difficult for the target solution to access all the probes on the surface of the array freely. In an attempt to overcome this problem, mixing of the solution during hybridisation was attempted. However, this had the adverse effect of increasing the formation of DNA networks within the microarray channels, resulting in

visible smearing within the microarrays after detection and inhibiting the identification of true hybridisations.

We hypothesise that this smearing effect, which has been seen often throughout the optimisation process, occurs as a result of networks made up of largely non-specific hybridisations which form within the channels of the microarray. This may be especially likely to happen around areas of ‘sticky’ probes e.g. those with a high GC content, which is another reason that further investigation into optimal probe design is necessary. Mixing may either aid formation of these networks, causing them to block the microchannels, or disrupt them leading to the accumulation of aggregate DNA within the arrays. This could also disrupt the initial specific hybridisation. Since the diameter of the microchannels is so small, movement of liquid through the channels is likely to be damaging to the network structures. It is likely that a combination of a number of these factors leads to the smearing that is seen in the microarray channels after detection.

We have shown that it is possible to reduce this network formation to some extent by performing a double digest on the DNA solution prior to labelling. Further work could be done to identify the optimum length of target probes using different combinations of restriction enzymes and looking at the optimum length of fragments that would result in successful hybridisation but minimal network formation and non-specific hybridisation. Theoretically it should be possible to reduce the length of the genomic DNA fragments to as small as 60 bp, as this is the size of the oligonucleotides on the array, although in practice this would depend on where the enzymes cut and fragments would probably have to be longer to avoid disrupting the probe sequence.

3.4.4 Summary

We have managed to develop a protocol for oaCGH on the Geniom platform, using which it has been possible to detect regions of known copy number variation. Unfortunately, due to the time constraints of this project, we were not able to continue with further optimisation of this method and were unable to use our optimal protocol to replicate and investigate copy number variations within the chromosome 12p13.31 region. This was largely due to limitations concerning probe length, which meant that it was not possible to include many of the oligonucleotides from this region which had been on the Nimblegen array. However, other investigations into structural variation in this region, carried out in parallel with oaCGH development, have succeeded in developing robust direct assays for copy number variation within 12p13.31. Therefore, the remainder of this thesis will concentrate on studying this locus in greater detail, using a different approach.

Chapter 4

Investigating Structural Variation within a Novel Tandem Duplication on Chromosome 12p13.31

4.1 Introduction

Oligo-arrayCGH experiments carried out within our research group suggested the presence of a region of structural variation located on chromosome 12p13.31 (Figure 4.1). To investigate these results further, members of the group performed sequence analysis of the publically available second generation sequencing trace data from within this region. These data were visualised using a trace viewing tool developed within our group. A highly similar repeated pattern of sequence trace depth was revealed, which suggested the presence of a previously unidentified tandem duplication (Figure 4.2). This pattern of sequencing trace depth can be hypothesised to occur as a result of the high level of sequence identity between two units of a tandem duplication, which means that sequencing trace data from this region will align to both units. We have termed the two units of this apparent duplication ‘A’ and ‘B’ (Figure 4.3). Studying the trace data reveals that in unit B, the pattern of traces is spread out over a larger distance. This occurs as a result of an increased number of repetitive elements in this unit, as shown via the UCSC repeat element data track (Figure 4.2).

Within this tandem duplication there are a number of genes, specifically *NANOG*, *NANOGPI* (an untranslated pseudogene which resulted from duplication of *NANOG*) and two sequence-related glucose transporter genes, *SLC2A3* and *SLC2A14*.

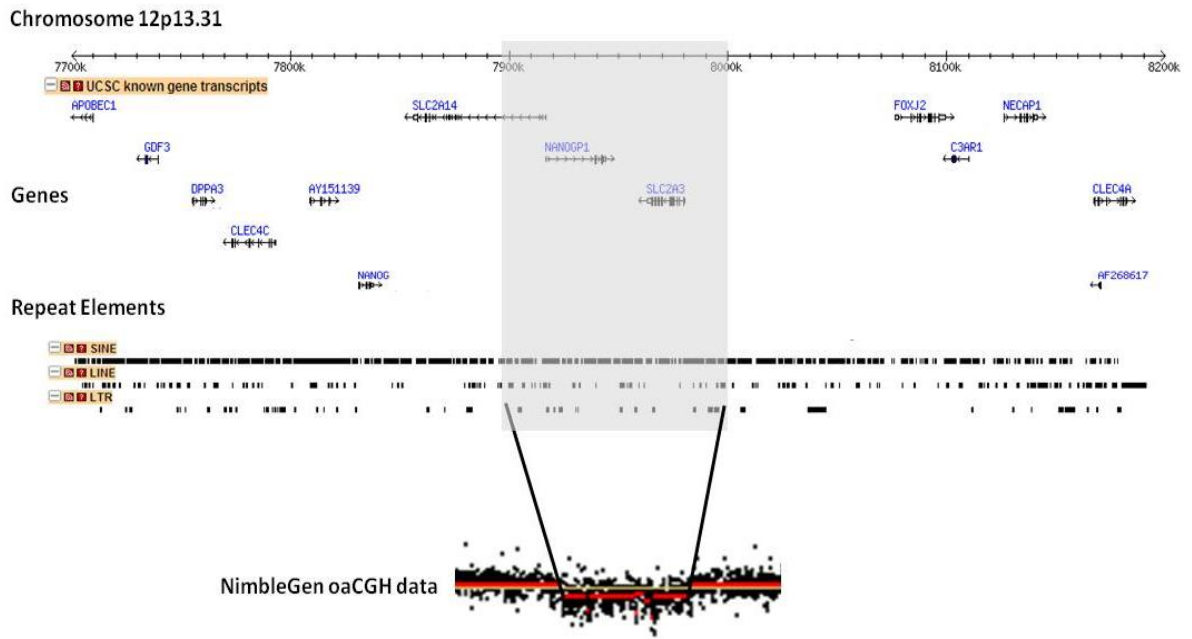


Figure 4.1: Structural Variation at 12p13.31

Chromosome 12p13.31 displayed in the UCSC Genome Browser. Genes and repeat elements present at this locus are shown. The putative area of structural variation identified using oaCGH on the Nimblegen platform is indicated by grey shading and the oaCGH data is represented as a scatter plot of the relative log₂ signal intensity ratios from two DNA samples. It can be seen that there is a decrease in ratio over this region, suggesting the presence of copy number variation.

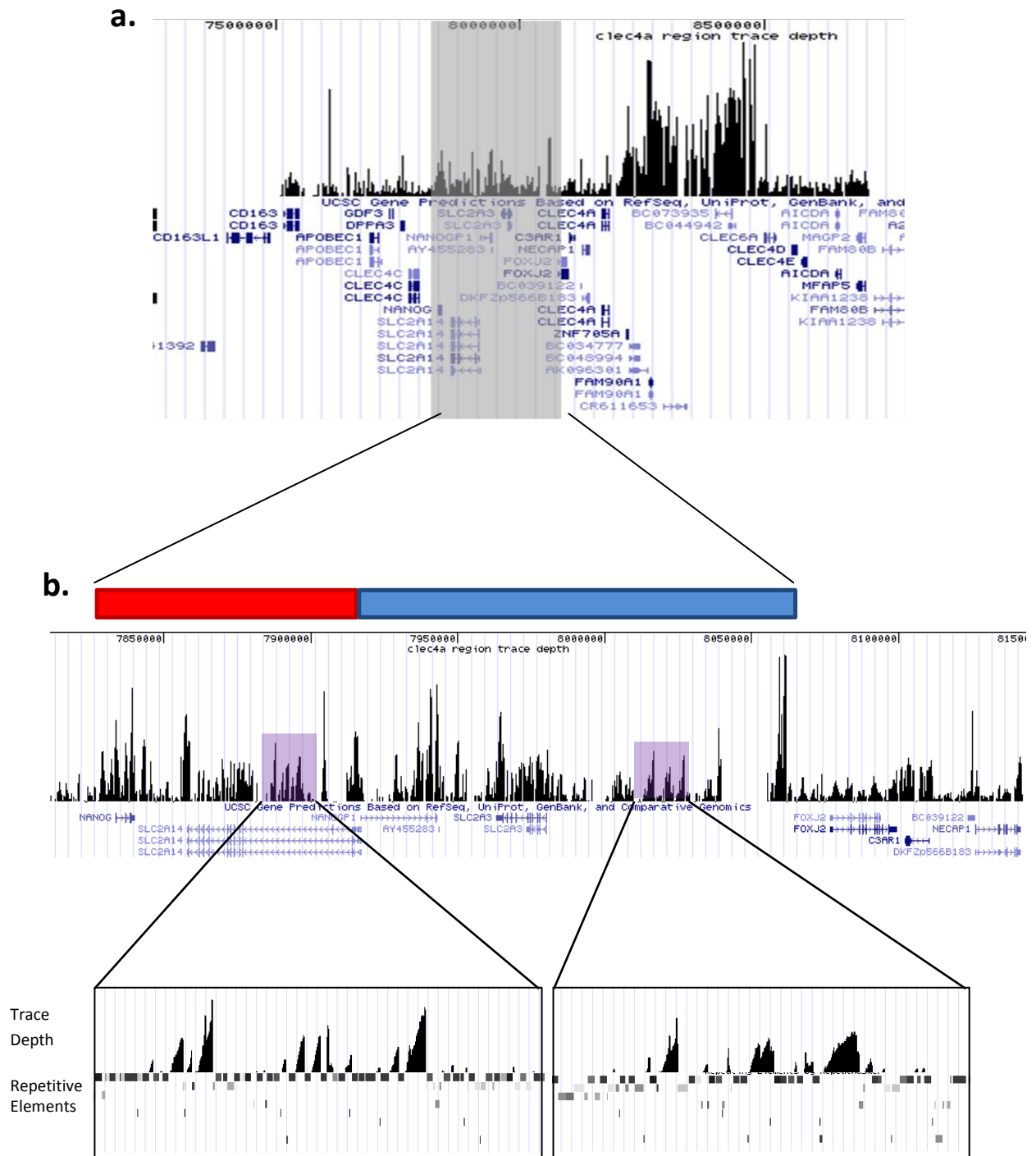


Figure 4.2: Trace Data Reveals a Tandem Duplication at 12p13.31

a.) Second Generation sequencing trace data from within this region, shown here in UCSC Genome Browser, revealed a highly similar pattern of trace depth between two tandem elements. The location of an apparent novel tandem duplication suggested by these data is indicated by grey shading. b.) Closer examination of sub-sections of the two regions, shown shaded in purple, reveals that in unit B the pattern is broken up by the presence of additional regions of sequence.

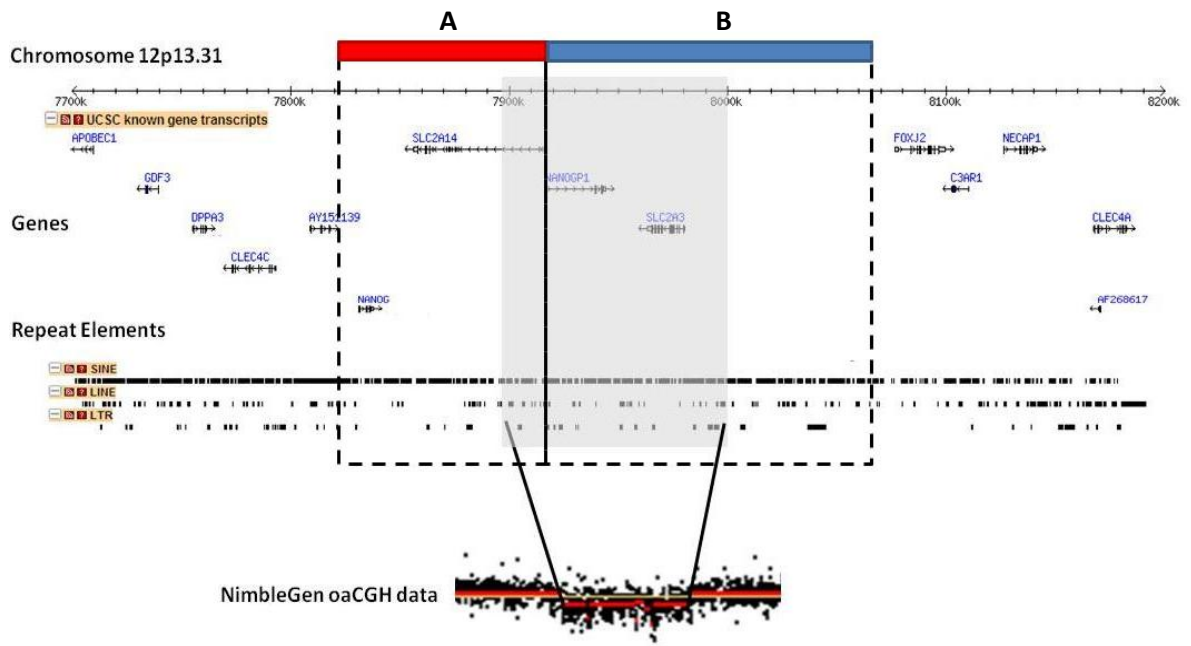


Figure 4.3: Tandem Duplication at 12p13.31

The likely presence of a tandem duplication within 12p13.31 was revealed by examination of the DNA sequence and sequencing trace files from within this region. The location of the tandem duplication relative to the oaCGH data which first suggested the presence of copy number variation is shown here. The two units of the tandem duplication are labelled as “A” and “B”. A red bar represents the location of unit A, whereas a blue bar shows unit B. It can be seen that the putative region of structural variation spans a section from each unit.

NANOG encodes a transcription factor expressed in embryonic stem cells, which is thought to play a key role in the maintenance of pluripotency (Mitsui *et al*, 2003). *SLC2A3* codes for the glucose transporter GLUT3, which is expressed in the brain as well as a number of other tissues, and has been shown to play an essential role in embryonic development (Schmidt *et al*, 2009). *SLC2A14*, the other glucose transporter gene within this region, codes for GLUT14 which is thought to be expressed only in the testes (Wu & Freeze, 2002). This gene shares a high level of sequence similarity with *SLC2A3*, which is to be expected if it resulted from a duplication of this gene. Due to the presence of a number of genes with important biological functions within the tandem duplication, structural variation at this locus could potentially have a

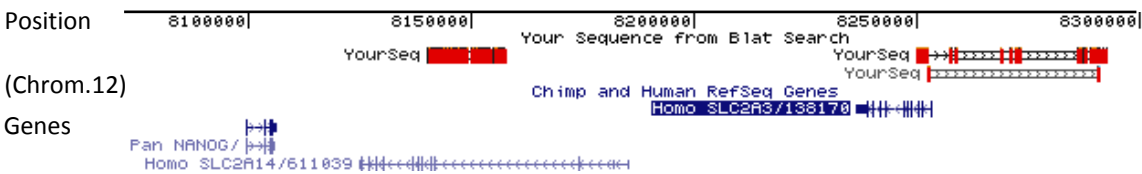
considerable effect on physiological processes including cell development and glucose transport.

Having identified a putative region of structural variation on chromosome 12p13.31, the next step was to carry out further investigations in order to confirm its presence, learn more about the frequency at which this variant occurs and to study its potential role in rheumatoid arthritis. This chapter describes further investigations of variation within this novel tandem duplication, through sequence analysis and development of direct assays for copy number variation.

4.2 Sequence Analysis

As described above, publically available second generation sequencing trace data from 12p13.31 suggests the presence of a previously undescribed tandem duplication within this region. To investigate at what point in time the tandem duplication was formed, homologous sequences from a number of mammals were examined using the UCSC genome browser. This revealed that, whilst the same duplication event can be seen in higher primates such as the chimpanzee, it is absent from other mammals including the horse and mouse (Figure 4.4).

a. Chimpanzee



b. Horse

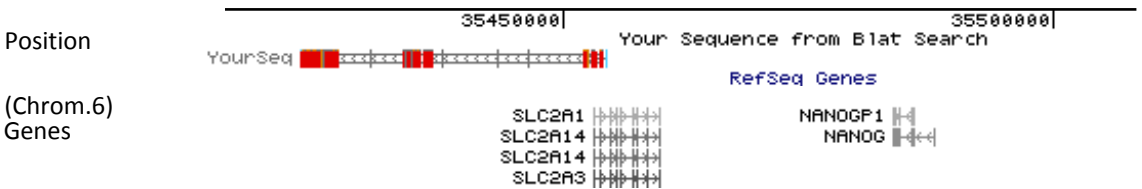


Figure 4.4: Investigating the Evolutionary Age of the Tandem Duplication

Sections of sequence from the human tandem duplication, labelled as ‘YourSeq’ and coloured red in the figure, were aligned with the reference genome of a number of other mammals in UCSC genome browser. Blue and grey lines show genes mapped to each location. a.)The human sequence matches two locations in the chimpanzee genome, indicating that the tandem duplication is present in this species. *SLC2A14* and *SLC2A3* are marked in different locations, suggesting that both genes are present. Since the track only illustrates chimp and human genes, the pseudogene *NANOGP1* is not shown in this alignment. b.) The same sequence only aligns to one position in the horse genome, indicating that the tandem duplication is absent. It can also be seen that the relative position of *NANOG* and *SLC2A3* is inverted in this genome. The human genes for both *SLC2A3* and *SLC2A14* map to the same position, suggesting that *SLC2A14* is not present as a separate gene in this species, but instead aligns to its closest match (*SLC2A3*). The same is true for *NANOG* and *NANOGP1*.

The closest non-primate human ancestor for which the genome sequence is currently available is the mouse. The primate and rodent ancestors are thought to have diverged around 90 million years ago (Janecka *et al*, 2007). Since the tandem duplication is not present in the mouse, then the initial duplication event must have occurred after this date.

In mammals without this duplication, *SLC2A14* and *NANOGPI* are absent. These results suggest that the duplication event took place after primates branched away from other mammals, and one copy of each gene then diverged to form *NANOGPI* and *SLC2A14*. While *NANOGPI* is an untranslated pseudogene, *SLC2A14* has evolved its own function as a glucose transporter specific to the testes (Wu & Freeze, 2002).

Direct comparison of the two units of the tandem duplication shows that the sequence similarity of the two units is structured as blocks of extremely similar (often identical) sequence with an average identity of 94%, which are broken up by stretches of sequence unique to either unit, much of which is composed of repetitive elements. A stretch of unique sequence covers the boundary between the two units of the duplication, and since this includes *NANOGPI*, we consider this region to be part of unit B.

One of the most obvious differences between the two units is size; Unit A is approximately 100 kb in length whereas unit B is around 140 kb. This is largely due to differences in repetitive elements between each of the two units, especially Alu elements, which have been differentially inserted or deleted. There are very few Alu elements located at the equivalent position in both of the two units. This suggests that the original duplication event took place before the insertion and rapid expansion of Alu elements, which is thought to have occurred around 30 – 50 million years ago (Britten,

1994). We have already ascertained that the initial duplication event cannot have occurred earlier than 90 million years ago, and therefore are now able to date the likely origin of the tandem duplication to between 90 and 30 million years ago.

The tandem duplication contains a particularly high frequency of repetitive elements, which to some extent is surprising; due to the importance of the genes within this region, insertion of sequence such as repetitive elements risks the disruption of the essential genes. However, it is possible that, once inserted into areas outside of genes, repetitive elements in this region have been maintained since their removal through deletion may risk affecting essential genes.

It has been shown that homozygous deletion of either *NANOG* or *SLC2A3* is lethal in mice (Mitsui *et al*, 2003; Ganguly *et al*, 2007). Currently there is no information on whether the same is true for humans; however there is no doubt that both *NANOG* and *SLC2A3* play vital roles in cell growth and development. Since each unit of the tandem duplication contains one of the original essential genes (*NANOG* in unit A and *SLC2A3* in unit B), neither unit can be permanently removed from the population, as a homozygous deletion of either gene is likely to be lethal. It is therefore interesting that we have detected structural variation within this locus, as changes in copy number of *SLC2A3* and *NANOG* which may affect expression levels could have a significant effect on cell development and function.

4.3 Structural Investigations

The simplest hypothesis regarding structural variation within the tandem duplication is that variation of the tandem duplication would take the form of deletion or duplication of either one complete unit, or part of one unit. However, this is not consistent with the oaCGH data, which suggests that the structural variant spans both units of the tandem duplication. This result could represent the real situation, or it could be incorrect due to misassembly of the sequence in this region. For example, it may be that there are regions of inversion or translocation within this duplication, which are not correctly represented in the reference genome. Before carrying out further studies of variation within this region, the tandem duplication was examined in further detail. This was necessary as future investigations, for example assay development, require an accurate knowledge of the sequence.

4.3.1 Analysis of Variant Boundaries

To test whether the current genome build correctly represents the overall structure of the sequence in this region, PCR primers were designed around the perceived boundaries of each unit of the duplication (as determined by analysis of the reference sequence). The aim was to amplify across these regions, and thereby identify any deviation in structure from the reference genome. Initially, three sets of primers were designed to amplify sequence across the proximal and distal boundaries of the tandem duplication, as well as the region of unique sequence located at the proximal end of unit B (Figure 4.5). These primers were located within regions of sequence unique to one or other unit of the tandem duplication, to ensure that any product would be a result of amplification across a specific boundary.

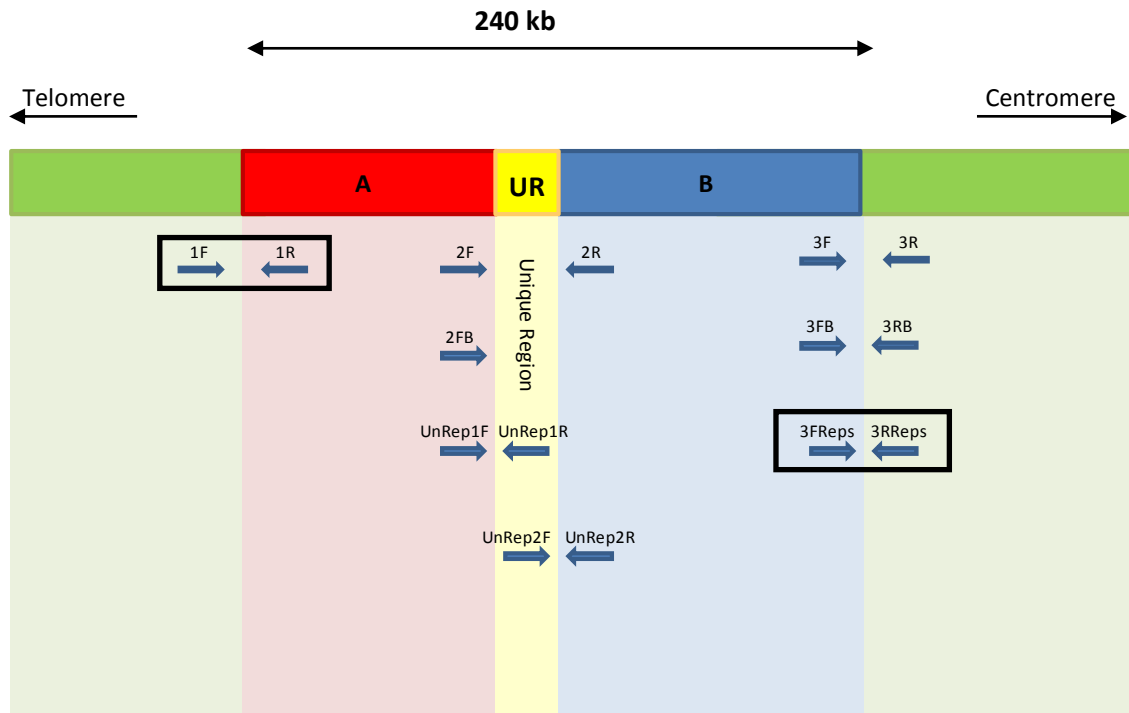


Figure 4.5: Primers Designed Around Theoretical Boundaries

PCR primers were designed to amplify sequences across the predicted boundaries of each unit of the tandem duplication. The two units are represented by shading in red (for A) and blue (for B) with the unique region located at the boundary of the two units indicated in yellow. The location of each primer is indicated, with arrows showing their 5' to 3' orientation. Primer pairs which produced successful amplifications are outlined with black boxes. It can be seen that only primer pairs 1F/1R and 3FReps/3RReps produced products of the correct size.

Although it was possible to amplify across the distal end of the tandem duplication using primers 1F/1R, the other amplifications failed to produce a product even after extensive variation of PCR conditions.

The design of the first set of primers was somewhat restricted by the requirement for them to be located within regions of unique sequence. A second series of primers were designed which were located either entirely or partially within repetitive elements. Since amplification can only take place if the forward and reverse primers are in close proximity to each other, even primers located within repeat regions can be specific

enough to produce a product from the sequence of interest. It was possible to amplify across the proximal end of unit B using primers 3FReps/3RReps. However, the other primers failed to produce a product (UnRep1F/UnRep1F, UnRep2F/UnRep2R, 3FB/3RB, and 2FB used with 2R).

Although it was possible to amplify across both the distal and proximal boundaries of this tandem duplication, none of the primers described in this section which were located within the duplication appeared to work consistently. This may be due to difficulties associated with designing primers in this region, for example the high degree of homology between the two units of the duplication makes it difficult to design primers unique to either unit, as does the high density of repetitive elements.

4.3.2 De Novo Sequence Assembly

We have experienced difficulties with amplifying sequence within the tandem duplication. This, together with oaCGH data suggesting the presence of structural variation spanning both units of the duplication, led us to question whether the current genome build is a true representation of the sequence structure at this locus. To investigate this further, all public domain sequencing trace files available from this region as of November 2008 (including the Venter traces obtained using the Sanger sequencing method and the Watson traces from Roche 454 sequencing) were downloaded and assembled *de novo* (without a reference sequence).

Sequence assembly was attempted using several different algorithms from programs including DNABaser (www.dnabaser.com) and Mira (Chevreux *et al*, 1999). The final assembly was performed using the Newbler Assembler software (Roche) since this

proved to be the quickest and most effective assembly program, as well as being the most user-friendly. Only the long sequencing reads were used in the final analysis, since the short reads did not assemble well *de novo*. For more detail on the assembly process and the settings used see Chapter 2 (Materials and Methods). The resulting contigs were aligned to the reference genome (build NCBI36/hg18, March 2006). Although most regions assembled without difficulty, there were some gaps between contigs which were not covered and also some incidences where adjacent contigs failed to join together (Figures 4.6 & 4.7).

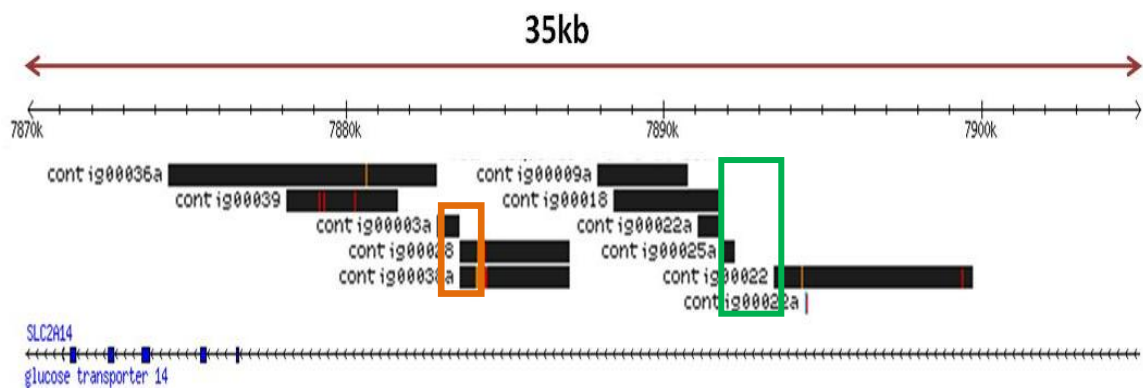


Figure 4.6: Results of a *de novo* Sequence Assembly

Sequencing trace files from the 12p13.31 locus were assembled *de novo* (without a reference sequence). The resulting contigs were aligned to the current genome build (build NCBI36/hg18, March 2006) using BLAT. An example output from this is shown above. Although in many regions the contigs aligned successfully, there were a number of situations where gaps remained between contigs (one example of this is indicated by a green box) and also where adjacent contigs failed to join together (represented by an orange box).

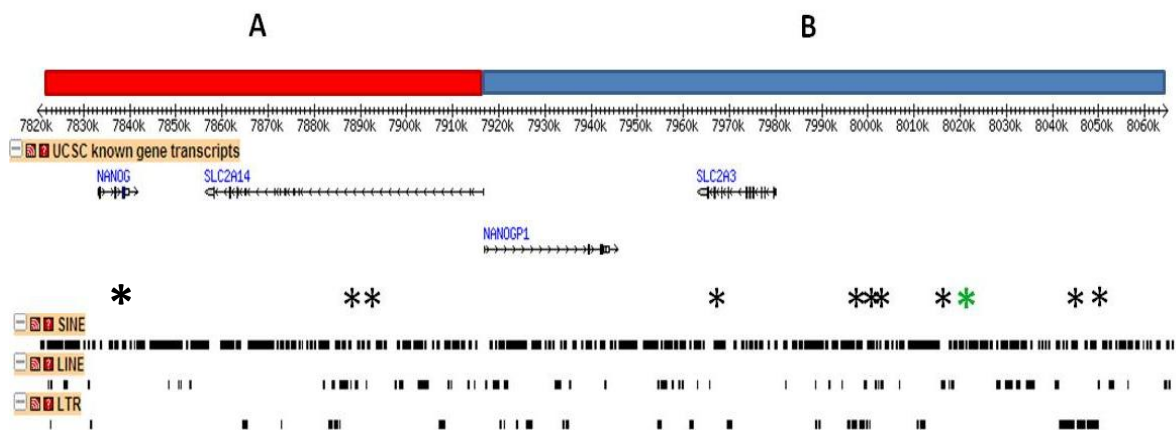


Figure 4.7: Gaps between Contigs Identified using *de novo* Sequence Assembly

Contigs produced from the *de novo* sequence assembly were aligned with the reference genome. In most cases there was complete coverage; however there were a number of gaps between contigs. The position of these gaps in the tandem duplication is indicated by asterisks. All but one of these gaps were subsequently amplified across using PCR, these are represented by black asterisks. A green asterisk indicates the gap which was not closed.

4.3.3 Confirmation of Region Structure Using PCR

Gaps in the sequence assembly could indicate the presence of structural rearrangements such as insertions. This would mean that the two contigs are not adjacent to each other in the genome, and explain why they do not join. To investigate this, primers were designed to amplify across gaps in the *de novo* assembly. Amplification of sequence across most of the gaps was successful, except for at one position (indicated in Figure 4.7 by a green rather than black asterisk) which repeatedly failed to amplify despite trying a range of PCR conditions and redesigning the primers. This may be due to genomic features such as the high number of repetitive elements in this region, which make the design of suitable primers challenging.

4.3.4 Identification of a Polymorphic Alu Element

Whilst amplifying across one of gaps in the sequence alignment, a variant was identified. In one of the DNA samples, amplification with the primers 46/1a produced four products of different sizes, rather than a single product as expected. This is shown in Figure 4.8, where an atypical sample is shown in track number 8. One of these products (corresponding to band 2 in Figure 4.8) is the expected product, whereas the identity of the other three was unknown. It seemed likely that this was the result of a sequence polymorphism. However, for a single locus in a diploid organism this variation would be expected to produce two products rather than four. To investigate the identity of the variant products further, sequencing was carried out.

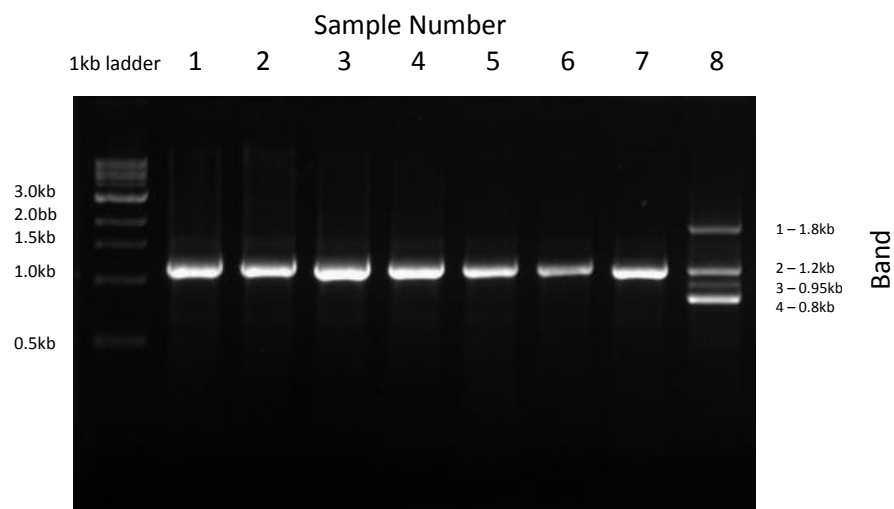


Figure 4.8: Identification of a variant sample

A variant sample was detected, in which amplification with primers 46/1a produced four bands on an agarose gel rather than a single band. The variant sample is seen here in lane 8, the other seven samples produce a single band as expected. The size of each product was determined using a 1kb ladder, and as is marked alongside the gel.

Whilst preparing the four products for sequencing, it was observed upon subsequent gel analysis that the largest band (band 1) had ‘broken down’ to give two products equal in size to bands 2 and 4, and band 3 was not recovered from the purification step (Figure 4.9). This left two products, each of which was sequenced using the dideoxy-sequencing method (For further details see Chapter 2, Materials & Methods). The sequencing data from bands 2 and 4 illustrates that the difference in size of the two products is due to the presence of a polymorphic Alu element in the variant sample (Figure 4.10).

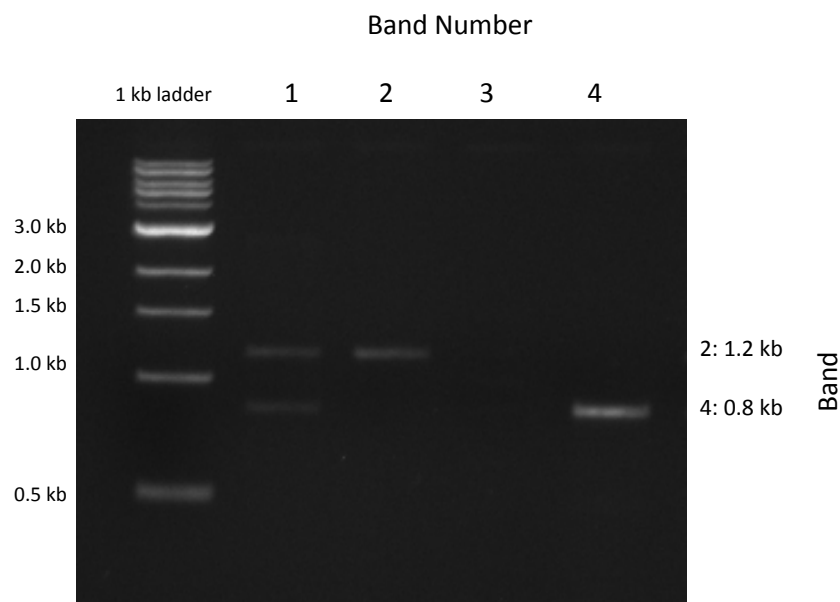


Figure 4.9: Separation and Purification of the Products from a DNA Variant

After extraction and purification, the four products were separated using agarose gel electrophoresis. The product from each band was run in a separate lane, indicated along the top of the gel picture. It can be seen that band 1 has separated to produce products equal in size to bands 2 and 4, suggesting that this may have represented a complex of these two products. No DNA was recovered from band 3.

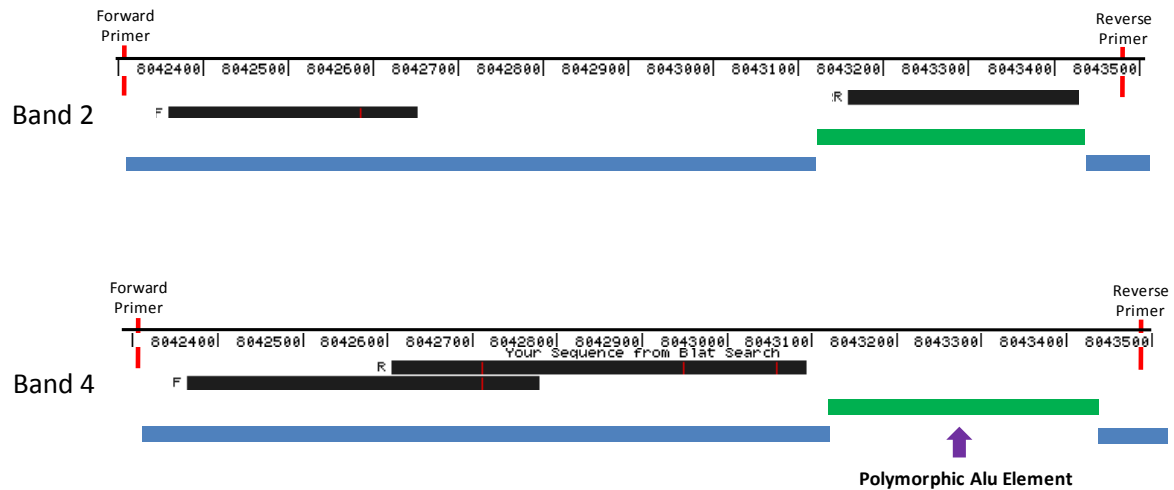


Figure 4.10: Identification of a Polymorphic Alu Element

Amplification with the primers 46 & 1a revealed a sequence variant. Alignment of the sequencing data from two of the products (corresponding to bands 2 and 4) with the current genome build reveals the presence of an unmarked polymorphic Alu element. The position of the forward and reverse primers is indicated, as are repetitive elements within this region (SINEs represented by green bars and LTRs shown as blue bars). The black bars represent the sequencing data; both the forward and reverse products were sequenced, indicated by F and R. A gap is present between the position of the primer and the product due to poor sequencing quality at the start of the trace. Data shows that in the case of band 4 the trace starts after an Alu element, indicating that this element is absent from this DNA sample.

Bands 1 and 3 were not recovered after sequence preparation. Band 1 separated into two products equal in size to bands 2 and 4, which suggests that it may have been a complex formed between two of the smaller products. In order to investigate this, the four variant products were separated using alkaline agarose gel electrophoresis (Figure 4.11). The high pH of this gel denatures the DNA, and as a result the molecules run through the gel as single strands. Two bands are visible on the alkaline gel, indicating that bands 1 and 3 were double stranded complexes of the normal and variant products.

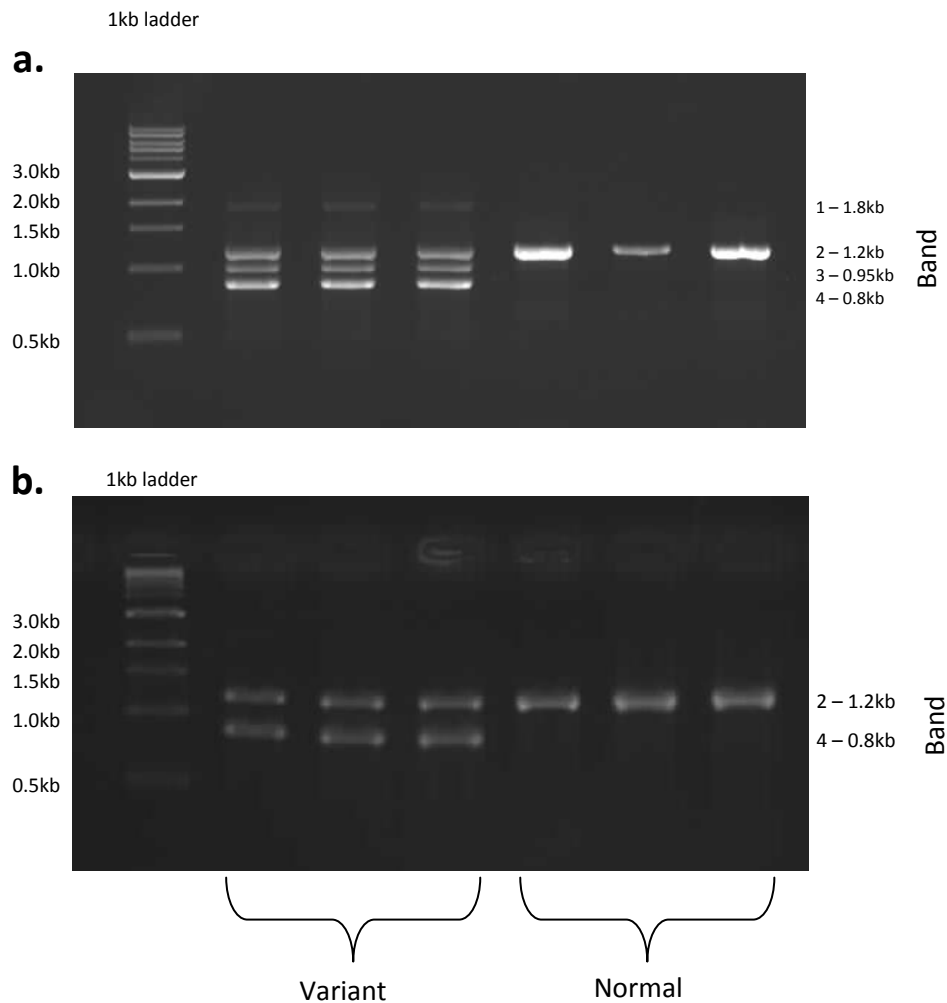


Figure 4.11: Investigating Variant Products Using an Alkaline Gel

a.) The products of the 46/1a PCR were separated on a standard agarose gel. A normal sample gives a single band, whereas the products of a variant sample separate to give four distinct bands. Three duplicates of each sample are shown. b.) When the same products were separated on an alkaline gel, under denaturing conditions, bands 1 and 3 disappear. This suggests that they occur as a result of complexes formed between the normal and variant products.

So far, the polymorphic Alu element described in this section has only been identified in a single sample. Amplification of the same region in an additional 92 samples was unable to identify any additional variants. This suggests that this polymorphism is relatively rare. The presence of a polymorphic Alu element in this region, as well as the

identification of a number of single base sequence anomalies from the sequencing data, suggests that although the assembly largely appears to be correct, there may still be a number of sequence errors within this region. It is likely that some of the SNPs from this region listed in dbSNP are not true polymorphisms, but sequence differences between the two units of the tandem duplication. This could be a contributing factor to difficulties associated with designing PCR primers within this locus. Since primers were designed according to the current sequence assembly, single base discrepancies could influence primer annealing and therefore affect the success of PCR-based reactions in this region.

4.4 Development of PRT-based Assays

In order to further study possible structural variation within the now validated tandem duplication on chromosome 12p13.31, assays were designed based on a version of the Parologue Ratio Test (PRT) (Armour *et al*, 2007). Standard PRT involves the simultaneous amplification of two sequences with one PCR primer pair. One target (the ‘reference’ site) has known copy number and the other is the ‘test’ sequence under investigation. The two products of different sizes are separated on an agarose gel, and the product intensity ratio is used to determine copy number of the test locus.

We designed a number of ‘modified PRT’ assays to investigate copy number variation at 12p13.31. Rather than one test and one reference region, these assays amplify one locus from each unit of the tandem duplication simultaneously. These modified PRT assays are therefore intended to report the ratio of two sites of potential variations in copy number, rather than (as in ‘normal PRT’) the ratio of one copy number varying site relative to a stable single copy locus.

The primers for these assays were designed by Colin Veal around stretches of DNA which are differentially interrupted by sequence insertions, for example repetitive elements, in the two units. This results in the amplification of two regions, one from each unit of the tandem duplication, which are of different lengths. These are then separated on an agarose gel. As for a standard PRT, the product ratio is compared and used to reveal the relative copy number of the two loci. In most cases the product from unit B is the largest, since this unit contains the largest number of inserted regions of sequence. Therefore the product ratio is always stated as the B:A ratio, i.e. calculated by dividing the amount of product from unit B by the amount of product from unit A (determined from the signal intensity of the products on an agarose gel). In theory, in a

situation where the copy number of the two units is equal, the B/A product ratio would be equal to 1. However, in practise this is not always the case, due to variation between reactions, differing fragment sizes, and agarose gel separation. In order to bring the product ratio of non-variant samples to 1, the raw product ratio data is therefore mathematically normalised (see Chapter 2, Methods and Materials for more details).

Since the modified PRT method described here does not employ a single-copy control region, it is only possible to identify relative changes in copy number between the two units, rather than absolute copy number of A or B. Also, each assay scores the relative copy number at a single position in each unit of the tandem duplication; the region within which the amplified sequence is located.

Initially twenty amplifications, termed A assays, were designed at varying points across the tandem duplication. After initial testing, eight were taken on for further optimisation on a panel of 95 Swedish control samples (Figure 4.12). Of these, Assay 9 (A9) produced the best and most reproducible results (Figure 4.13), probably due to the fact that the two products produced by this assay are relatively short (200 bp in A and 285 bp in B) and have a small size difference.

4.4.1 Correction of Variation Between Different Batches of DNA Samples

Whilst carrying out optimisation of the A9 assay, visible differences between the raw signal intensity ratios (before normalisation) in different sections of the control plate were identified (Figure 4.14a). Since this effect often transcended more than one row of an agarose gel, it was considered unlikely to be a gel artifact.

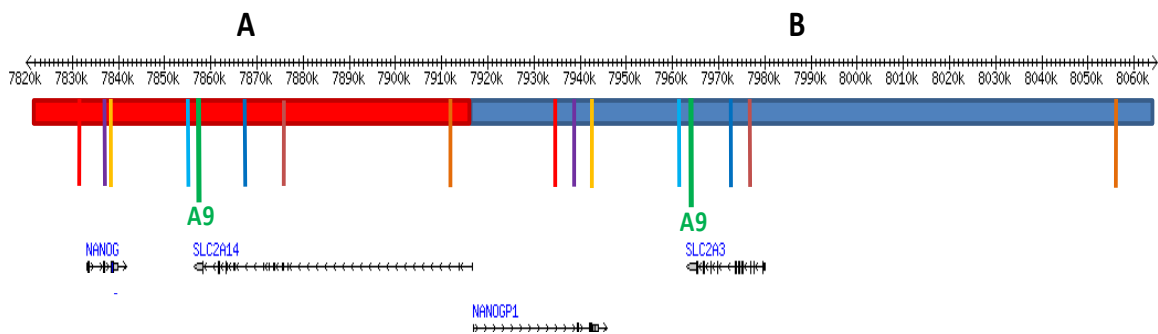


Figure 4.12: The Location of PRT-based Assays Within the Tandem Duplication

Modified PRT assays were designed to reveal relative changes in copy number between the two units of the tandem duplication. Initially twenty assays were designed; of these, the eight selected for initial optimisation are shown on the diagram above. The two units of the tandem duplication are shown as a coloured bar, red for unit A and blue for unit B. The location of each set of primers in the two units is represented by a coloured line. Each colour corresponds to a different assay. Assay A9 is highlighted as this was the most robust and reproducible assay.

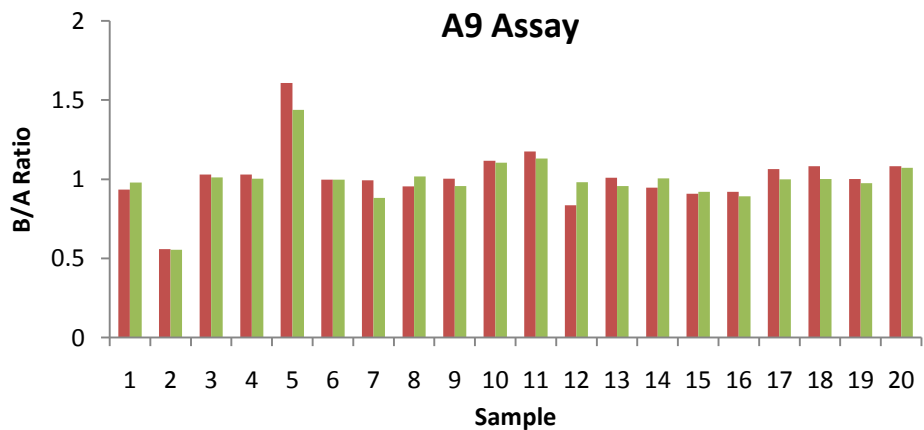


Figure 4.13: A9 Assay

The A9 assay was shown to be the most reproducible of the PRT-based assays. The relative copy number of sequence in the two units of the tandem duplication is determined using the ratio of product from each unit. This ratio is represented graphically as shown above. The coloured bars represent results from two duplicate experiments, showing that this assay is highly reproducible. Two copy number variations can be seen, one observed as an increase in product ratio (sample 5) and the other as a decrease in this ratio (sample 2).

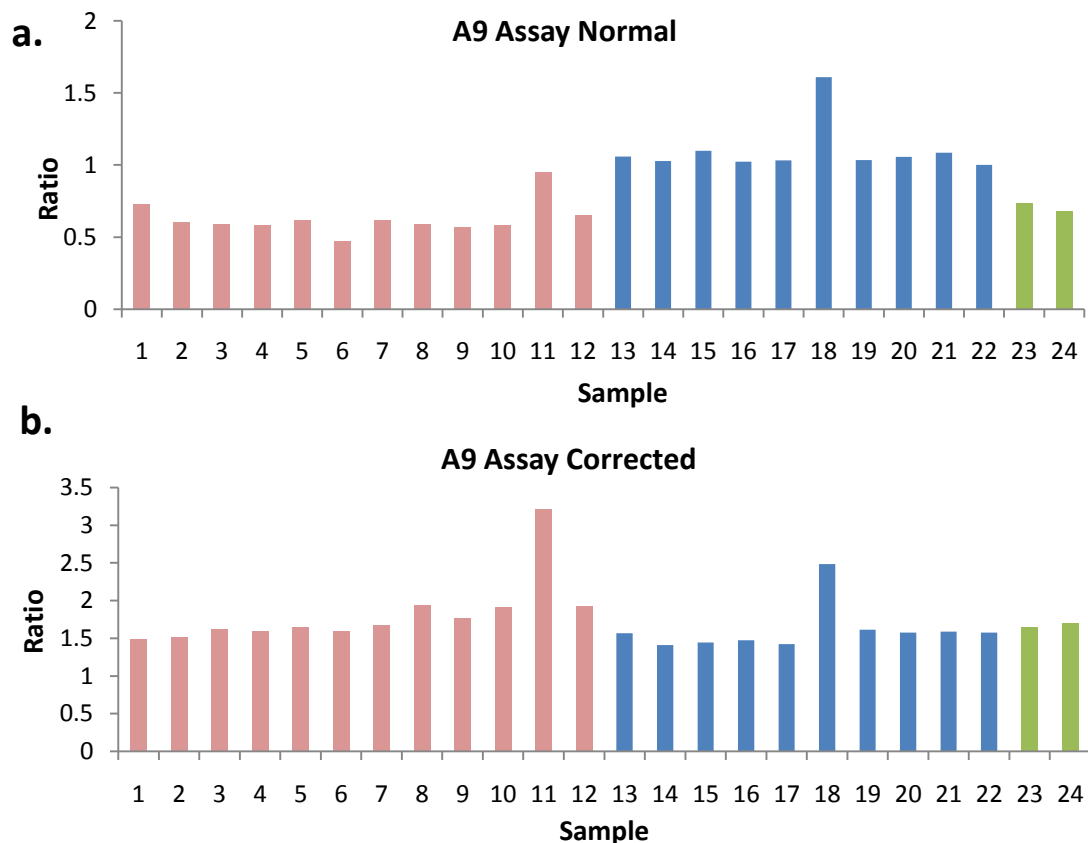


Figure 4.14: Elimination of Batch Effect

It was observed that samples from different DNA extraction batches amplified slightly differently, with a noticeable difference between the ‘normal’ ratio of each batch (a.). Each batch of DNAs extracted at the same time is represented using different coloured bars. Raw product ratio data is shown here (before normalisation) b.) The addition of 2 M Betaine to the reaction appears to correct this effect, and bring the base level to the same level. This also makes the variation in sample 11 more apparent.

Further investigation revealed that the DNA samples had been extracted in several batches, all within a few days of each other and by the same individual; however each batch appeared to be amplifying slightly differently in this assay. This illustrates the sensitivity of this type of genotyping method to sample quality and slight differences in extraction procedure. Optimisation of reaction conditions was carried out and it was

possible to eliminate this batch variation by the addition of between 1.5-2 M Betaine to the reaction (Figure 4.14b) (Anthony Brookes, Personal Communication).

The optimised A9 assay was used to genotype 95 Swedish control DNAs (Figure 4.15). Results show that five of the Swedish samples contain a copy number variant, detected as either an increase or decrease in product ratio. This corresponds to a frequency of 5.26%. Repetition of this genotyping demonstrated that the results of this assay are highly reproducible.

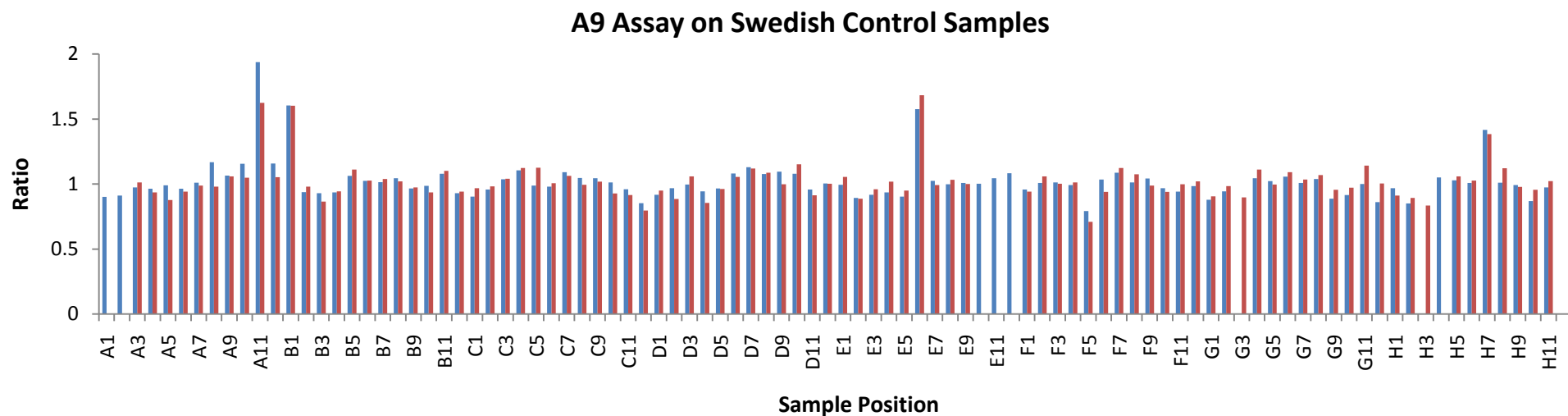


Figure 4.15: Identification of Copy Number Variation in Swedish Control Samples

The A9 assay was used to genotype 95 Swedish control samples. Red and blue bars represent the results of two duplicate experiments. Five samples show the presence of copy number variation, detected as either an increase or decrease in the ratio of signal intensities. Four samples (A11, B1, E6 and H7) show an increase in product ratio, whereas one sample (F5) shows a decrease in this ratio.

4.5 Studying Variation in HapMap Samples

Up to this point, all of the samples genotyped were Swedish control DNAs. To investigate whether the frequency of copy number variation within this tandem duplication differs between populations, the A9 assay was used to investigate variation in the 270 DNA samples used in the HapMap project. These samples are from three geographically distinct populations; Caucasian, African (Yoruba) and Asian (Chinese and Japanese) (International HapMap Consortium, 2003). The Caucasian HapMap samples are part of the Foundation Jean Dausset-Centre d'Etude du Polymorphisme Humain (CEPH) collection, which contains over 40 multi-generation families from Utah, with ancestry from Northern and Western Europe. These will be referred to as CEPH samples in this thesis.

Samples from all three populations were genotyped using the A9 assay. The frequency and type of each variant was calculated for each population, and compared to data previously obtained from the Swedish samples. Copy number variation was identified in all three populations (Table 4.1).

Table 4.1: Frequency of Variation in Four Populations

Population	No. Samples	Increased Ratio	Decreased Ratio	Variation Frequency
Swedish	95	4	1	5.3%
CEPH	59	4	2 (3)	10.2%
Yoruba	59	1	0	1.7%
Chinese/Japanese	95	4	0	4.2%

In the CEPH samples, as for the Swedish samples, two forms of variation were identified, revealed by either an increase or decrease in the product ratio. However, in the Yoruba and Chinese/Japanese samples, only variation resulting in an increase in product ratio was detected. The CEPH and Yoruba sample collections contain family trios, and only DNA from the parents was used for the above analysis. This was necessary to ensure that the CNV status of the samples remained independent of each other, a requirement for accurate determination of the CNV frequency. In the case of the CEPH samples, one sample which initially showed a decreased ratio was later shown (upon replication) to be a false positive. This sample is shown in brackets in the table, and was not included when calculating the population frequency.

The frequency of any variation within the tandem duplication was calculated for each population (Table 4.1). The results show considerable population differences. Variants occur most often within the CEPH samples, with a frequency of 10.2%, whereas the Yoruba have the lowest frequency of variants at 1.7%. It is interesting that we have detected such large differences between populations. To test whether any of these differences are statistically significant, pair-wise Fisher exact tests were carried out on the four sample sets. This test was chosen due to the fact that in all four populations the number of variants detected is very small (less than five). Results of these calculations (not shown) indicate that none of the differences are statistically significant (have a p-value of less than or equal to 0.05) and therefore they may be a result of sampling differences. To gain a more accurate picture of population differences in the frequency of this variant, it would be necessary to study much larger sample collections.

The presence of related individuals within the HapMap collections enabled us to investigate whether copy number variation within the tandem duplication is stably inherited. The results of genotyping family trios from the CEPH and Yoruba collections

were studied (Figure 4.16). We identified three situations where copy number variation was transmitted from parent to child, all involving the stable transfer of an increased ratio of signal intensities (CEPH trios 11 and 24, Yoruba trio 12). There were also four incidences where the same variants were present in a parent only, and not inherited by the child (CEPH trios 2, 5, 18 and 29).

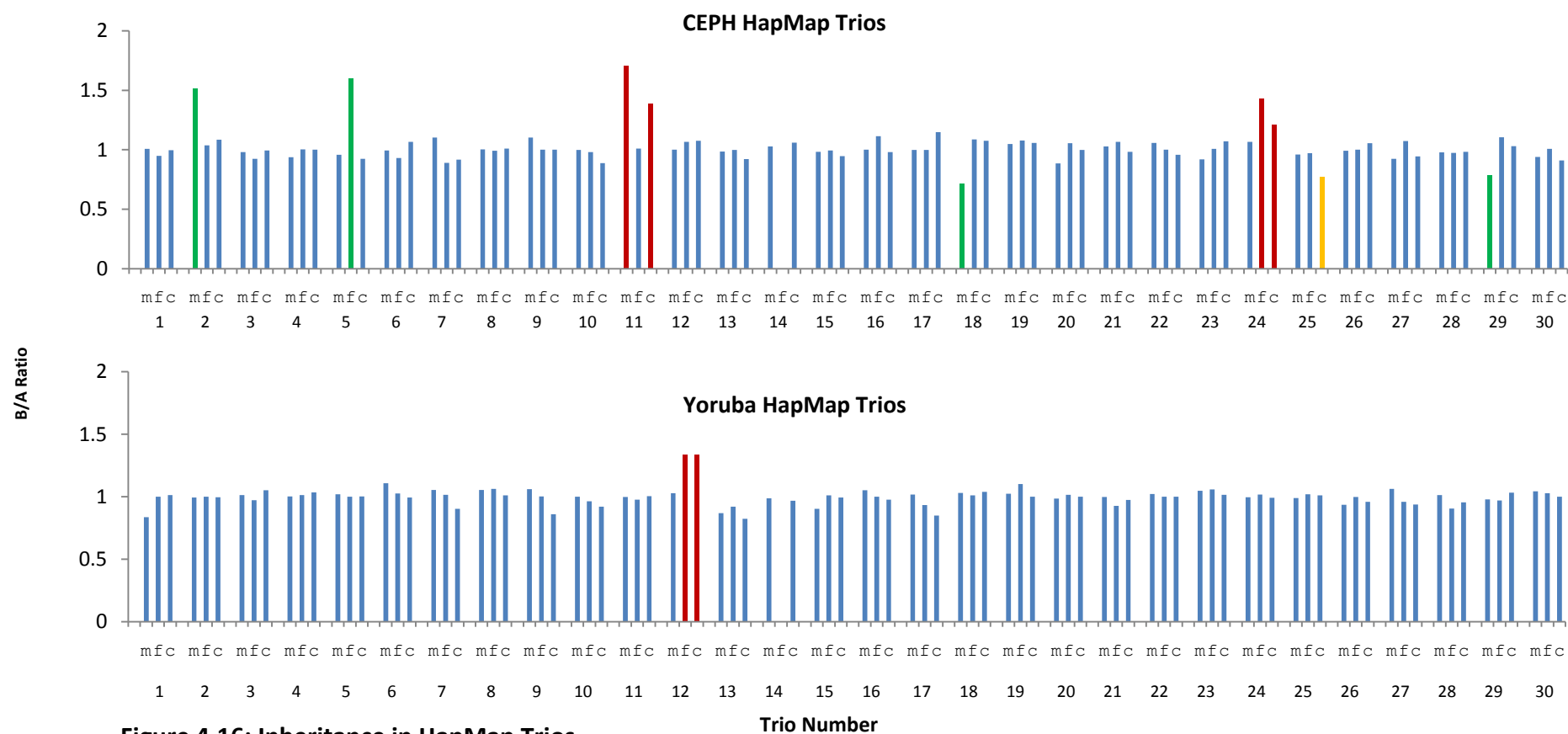


Figure 4.16: Inheritance in HapMap Trios

Family trios from the CEPH and Yoruba HapMap populations were genotyped using the A9 assay. Each set of three results represents one family trio, ordered mother, father, child (m, f, c) (for more details on the family trios used here, see Appendix 2). Blue bars represent samples which do not contain variation within the A9 regions, red bars show inherited variations and green bars indicate variants which are not inherited. The single yellow bar indicates a variation which was initially thought to be *de novo*, but was later shown to be a false positive result.

4.6 Investigating CNV Inheritance

Studying family trios from the CEPH and Yoruba HapMap collections revealed several incidences where copy number variation within the tandem duplication was stably inherited. However, so far this has only been studied in a small number of samples in small family groups (trios). To investigate inheritance of this copy number variant further, larger CEPH families were obtained and genotyped. In total four families were chosen for this study. CEPH family 1362, which had not been identified as variable in the initial genotyping, was used as a control. Family 1341 had shown a variant in one sample which had not been inherited, and families 1349 and 1447 had shown variants which were inherited.

All four families were genotyped using the A9 assay (Figure 4.17). The results showed no variation in any member of family 1362 (data not shown) and in no individuals of family 1341 other than the paternal grandfather. However, copy number variation within the tandem duplication at 12p13.31 was detected in all three generations of families 1349 and 1447. In CEPH family 1349, an allele was transmitted from the maternal grandmother to the mother, and subsequently inherited by five of her eight children (three daughters and two sons). In family 1447, the variation was inherited through the male line from the paternal grandfather to the father, and then to two of his eight children (both of them male). This data shows that variation with the tandem duplication is inherited, in a manner which is consistent with Mendelian patterns of segregation.

Figure 4.17: Patterns of CNV Inheritance in CEPH Families

Four CEPH families were genotyped using the A9 assay. Results of the genotyping are shown for each family, along with a pedigree which displays the inheritance of variation within a tandem duplication at 12p13.31. Our normal method of normalisation was not suitable for this data since in a number of families the majority of samples were variants, which made the process inaccurate. Therefore the raw data is shown. Squares represent males whereas circles represent females. The CEPH code for each individual is shown beneath each sample. This consists of a number, which is different for each family member, and a letter which indicates the family relationships (for example FF, Father's Father; F, Father; M, Mother; D, Daughter; S, Son). A shaded symbol is used to indicate an individual with CNV in the region of interest. A dashed line indicates that the sample was not available for this study. a.) Variation in this region was not inherited in family 1341. b.) & c.) Copy number variation was detected in all three generations of families 1349 and 1447.

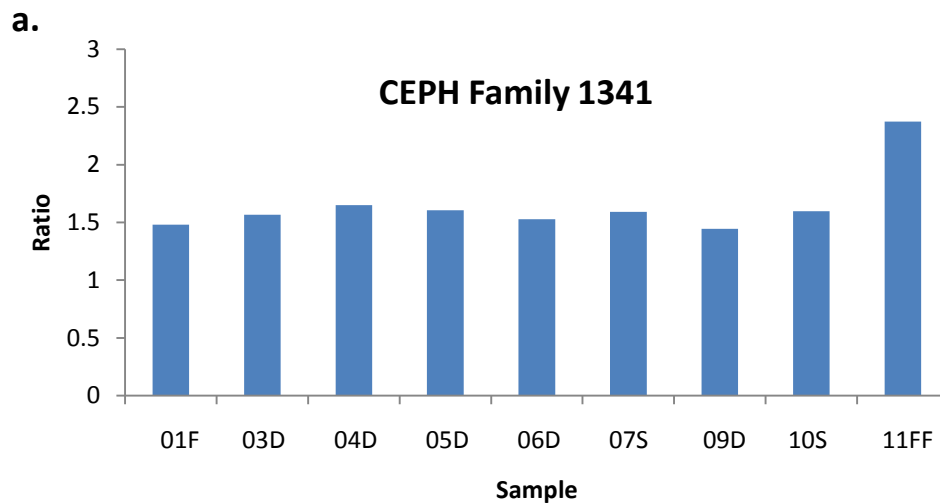
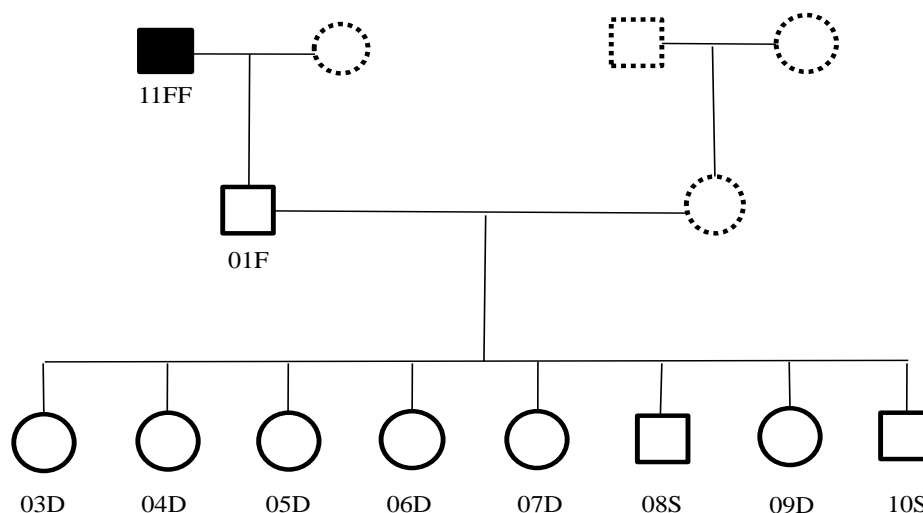
**CEPH Family 1341 Pedigree**

Figure 4.17 continued

b.

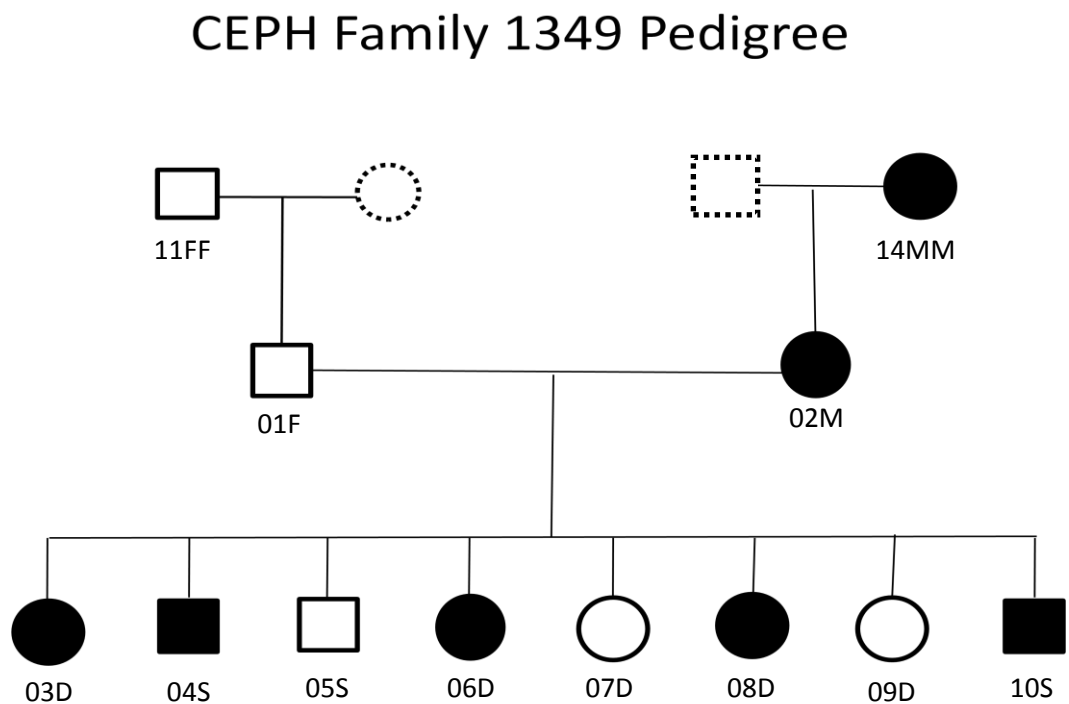
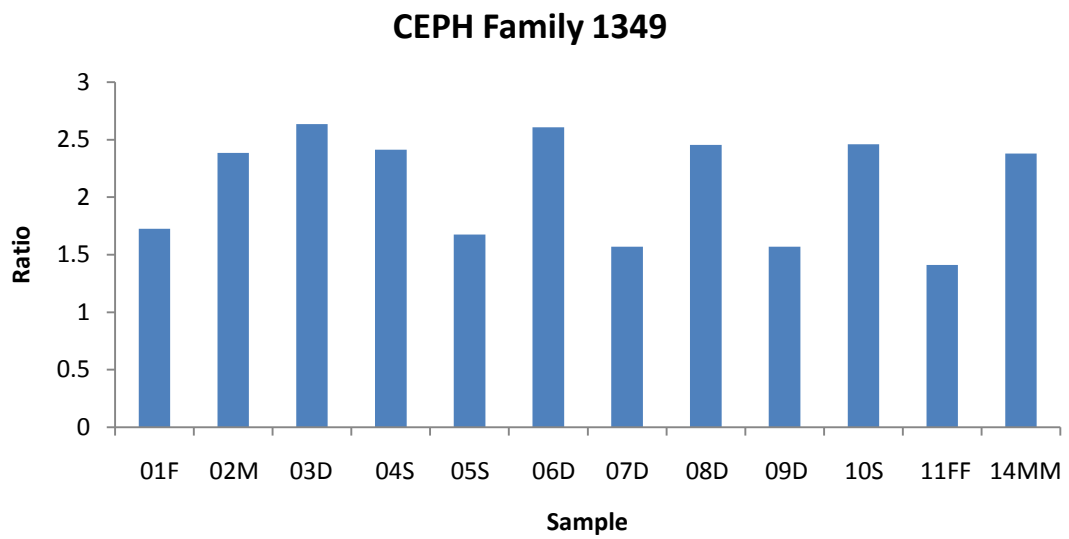
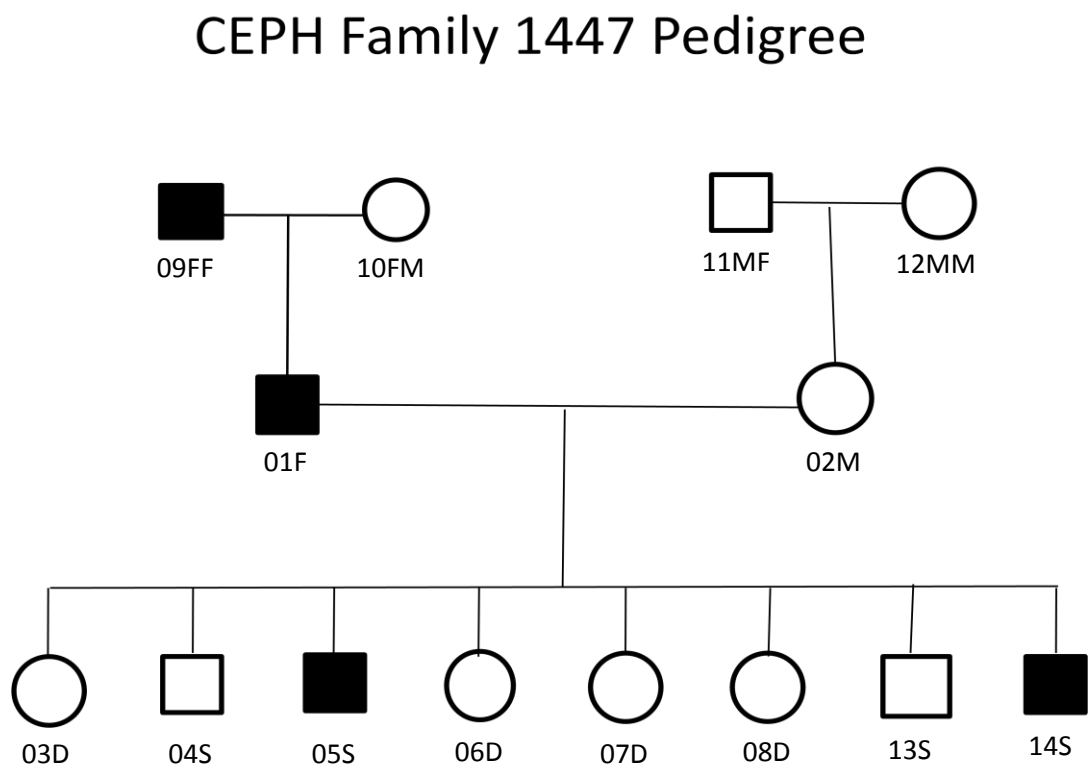
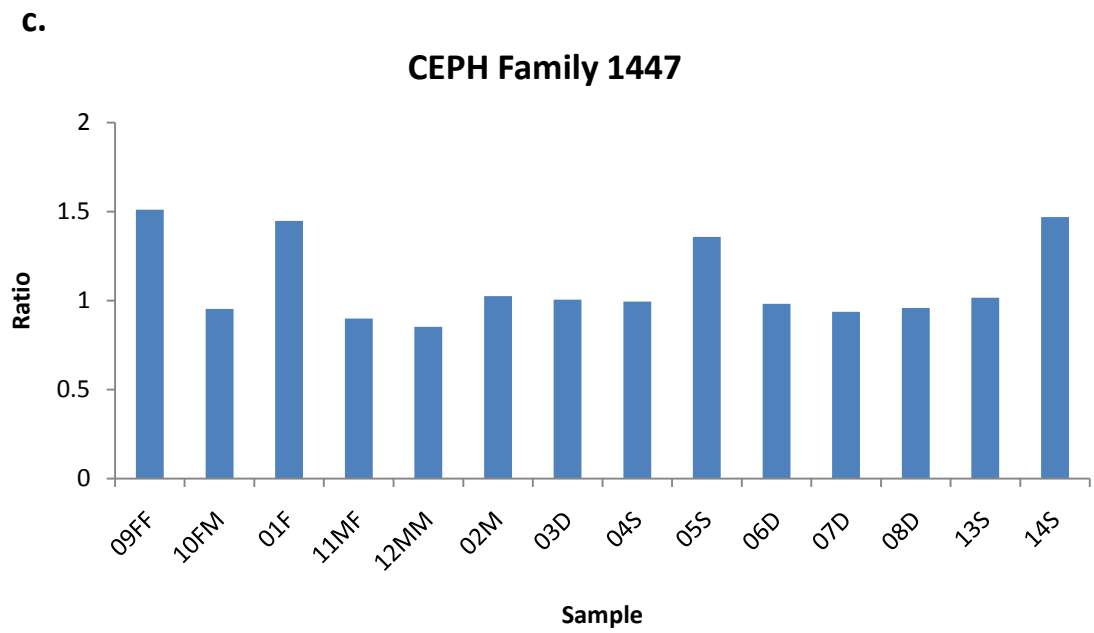


Figure 4.17 continued



4.7 Summary and Discussion

This chapter has described sequence investigations of a novel copy number variable tandem duplication, located on chromosome 12p13.31. Sequence analysis has enabled us to predict that this is an ancient duplication event, which most likely occurred between 90 and 30 million years ago, after the divergence of the primate and rodent lineages and before the rapid expansion of Alu elements.

Investigations into the sequence structure of this tandem duplication showed that, despite difficulties with PCR in this region, the structure of the reference genome appears to be correct. This was shown by a *de novo* assembly of second generation sequencing traces from within this region.

A number of assays were designed to detect copy number variation within this region. One of these, the A9 assay, has been shown to be a highly reproducible and effective method for detecting copy number variation within this region. However, in order to gain a more overall picture of variation across the tandem duplication, and to detect potential sites of recombination, a greater number of assays located at different points across the two units will be required. This has already been attempted with the A assays described in this chapter; however these were of limited success, in part due to difficulties with primer design in this region which have already been discussed. Therefore, in order to investigate recombination events in greater detail, a different approach is required. This issue is addressed in Chapter 6.

So far, using the A9 assay, it has been possible to detect two distinct forms of variation within this region. These are visible either as an increase or decrease in the ratio of sequence at the A9 position between the two repeat copies. Since the product ratio compares the relative amount of product from each unit, each of the two altered ratios

could be a result of a change in copy number of either unit. The ratio is calculated by dividing the amount of product from A9(B) by the amount of product from A9(A). Therefore, an increase in product ratio indicates that the amount of product from A9(B) has increased relative to the amount from A9(A). This could be a result of a duplication of A9(B), or a deletion of A9(A). However, this method is only able to identify changes in copy number which involve sequence at the position of the A9 assay, and says nothing of changes which may involve sequence at other locations within the tandem duplication. Further discussion of the different forms of copy number variation detected in this region using the A9 assay can be found in Chapter 5 (Section 5.2).

The frequency of variation within the tandem duplication has been shown to vary between populations. This ranged from 10.17% in the CEPH samples to 5.26% in Swedish samples, and 1.69% in the Yoruba individuals. However, given the small sample sets, these results are subject to error. A much larger study would be required to confirm whether the population differences implied here are accurate.

Using HapMap family trios, we have shown that at least one form of variation within the tandem duplication appears to be inherited in a Mendelian fashion. This was confirmed through the study of four larger CEPH families. So far it has only been possible to investigate inheritance of variation in this region detected as an increase in product ratio, due to sample availability. To investigate whether other types of variation are inherited in the same way, a larger number of families would need to be genotyped.

In summary, this chapter has described sequence analysis of a novel tandem duplication located on chromosome 12p13.31, and the design of assays to assess structural variation within this region. These assays have confirmed the presence of copy number variation and shown that this variant is present at different frequencies in a number of

populations. Previous investigations have suggested association of DNA variation at 12p13.31 with rheumatoid arthritis (Lorentzen *et al*, 2007; Jawaheer *et al*, 2001). Therefore in the next chapter we will go on to use the assays described here to investigate the involvement of copy number variation in this region with complex disorders including RA.

Chapter 5

Association Analysis of a CNV Located at 12p13.31 with Rheumatoid Arthritis and Other Complex Disorders

5.1 Introduction

Heritability for rheumatoid arthritis (RA) is estimated at around 60% (Macgregor et al, 2000). A large proportion of the genetic susceptibility is thought to be due to the effect of shared epitope alleles in the HLA region, which was the first RA susceptibility locus to be identified (Stastny, 1976; Gregersen et al, 1987). Other risk loci have since been described, not least via GWAS carried out for common complex disorders. However, so far the vast majority of RA risk loci appear to contribute a relatively modest effect to disease susceptibility in comparison to the HLA genes. It has been suggested that alleles of the HLA locus may explain up to 60% of the heritability of RA (Deighton *et al*, 1989) compared to an estimated 8% contributed by the 1858T polymorphism in *PTPN22* (Lee et al, 2007b). *PTPN22* was the second locus for which association with RA was confirmed, and increased risk contributed by common *PTPN22* alleles is thought to be far greater than for any other non-HLA gene. Therefore a significant portion of the genetic basis of RA remains to be accounted for. Some of this remaining heritability may be explained by the involvement of copy number variation. So far just one such variant, an increased copy number of the ligand CCL3L1, is thought to contribute to an increased risk of RA (McKinney et al, 2008).

A number of association studies (Jawaheer *et al*, 2001; Lorentzen *et al*, 2007) have suggested the presence of DNA variations within the chromosome 12p13.31 locus, which may play a role in susceptibility to RA. OaCGH experiments within our lab carried out using the Nimblegen array platform were able to identify a region of structural variation within this region. Detailed sequence analysis of the 12p13.31 locus revealed the presence of a novel tandem duplication, which we have shown to be copy number variable (see Chapter 4). In order to investigate whether this structural variant is associated with RA, we obtained and genotyped a disease cohort using the A9 assay, which was first used to reveal CNV at this locus (Chapter 4).

RA is known to share common susceptibility loci with the inflammatory skin disorder psoriasis, as well as with other autoimmune diseases (Jawaheer *et al*, 2001). Risk loci shared with psoriasis include genes within the HLA region (Mallon *et al*, 1999), *STAT4* (Zervou *et al*, 2009) and *TNFAIP3* (discussed in Li & Begovich, 2009). We therefore went on to investigate whether variation at 12p13.31 is associated with psoriasis.

There is evidence to suggest that, in addition to autoimmune disorders, the 12p13.31 locus may be involved in susceptibility to other forms of complex disease. There are several human and rat Quantitative Trait Loci (QTLs) located within this locus which are associated with phenotypes including blood pressure and glucose, insulin and cholesterol levels. Such factors are involved in susceptibility to disorders such as diabetes and heart disease. To test whether variation in this region is involved with cardiovascular disease, a cardiovascular disease cohort was also genotyped. Results of the association studies described above are presented in this chapter.

5.1.1 Samples

Case and control samples were provided by collaborators (see Section 2.1.2 for details). In an attempt to rule out population stratification, we chose case and control sample sets which had previously been matched. The Swedish case and control RA sets had been used together for previous studies (Lorentzen *et al*, 2007). Analysis of the two control sets used for the UK studies, and justification of their combining, has been described elsewhere, where they were used alongside the same RA and CVD case groups (WTCCC, 2007).

The Swedish samples were provided on 384 well plates with mixed case and control samples, and were genotyped simultaneously. Due to the fact that the rest of the samples were provided as separate cohorts, it was not possible to carry the remaining studies out blind (without knowing which samples were from the case group and which were controls). However, where possible, genotyping of case and control samples was carried out simultaneously and reagents were obtained from the same batch in order to keep any differences to a minimum.

5.2 Swedish Rheumatoid Arthritis Cohort

A Swedish RA cohort made up of 2403 rheumatoid arthritis cases and 1269 control samples was obtained from collaborators at the Karolinska Institutet, Stockholm. These samples were genotyped using the A9 assay described in Chapter 4.4. Previous genotyping using this assay identified two distinct forms of variation within a tandem duplication on chromosome 12p13.31 which were visible as changes in the ratio of product amplified from each unit of the tandem duplication, one as an increase in product ratio and the other as a decrease. We believe that these represent changes in relative copy number between the two units. However, genotyping results from this larger set of Swedish samples suggest that the two original categories of variation can be further divided to give four distinct variant genotype classes, all of which were detected in both case and control samples (Figure 5.1). These four genotypes are seen in addition to the ‘normal’ genotype expected from amplification of the reference sequence. We hypothesised that these four genotype classes occur as a result of a deletion or duplication of sequence from either unit of the tandem duplication at the respective A9 assay site. Each of these events would alter the product ratio in a different and predictable way.

5.2.1 Genotypes Distinguishable using the A9 Assay

The A9 assay product ratios which we would expect to occur as a result of changes to the reference sequence in a diploid individual are shown in Table 5.1. The A9 assay is only able to reveal copy number at single position within each unit, which is indicated by referring to the copy number of each unit as A9(A) and A9(B). In many cases, multiple genotypes result in the same product ratio (Table 5.2).

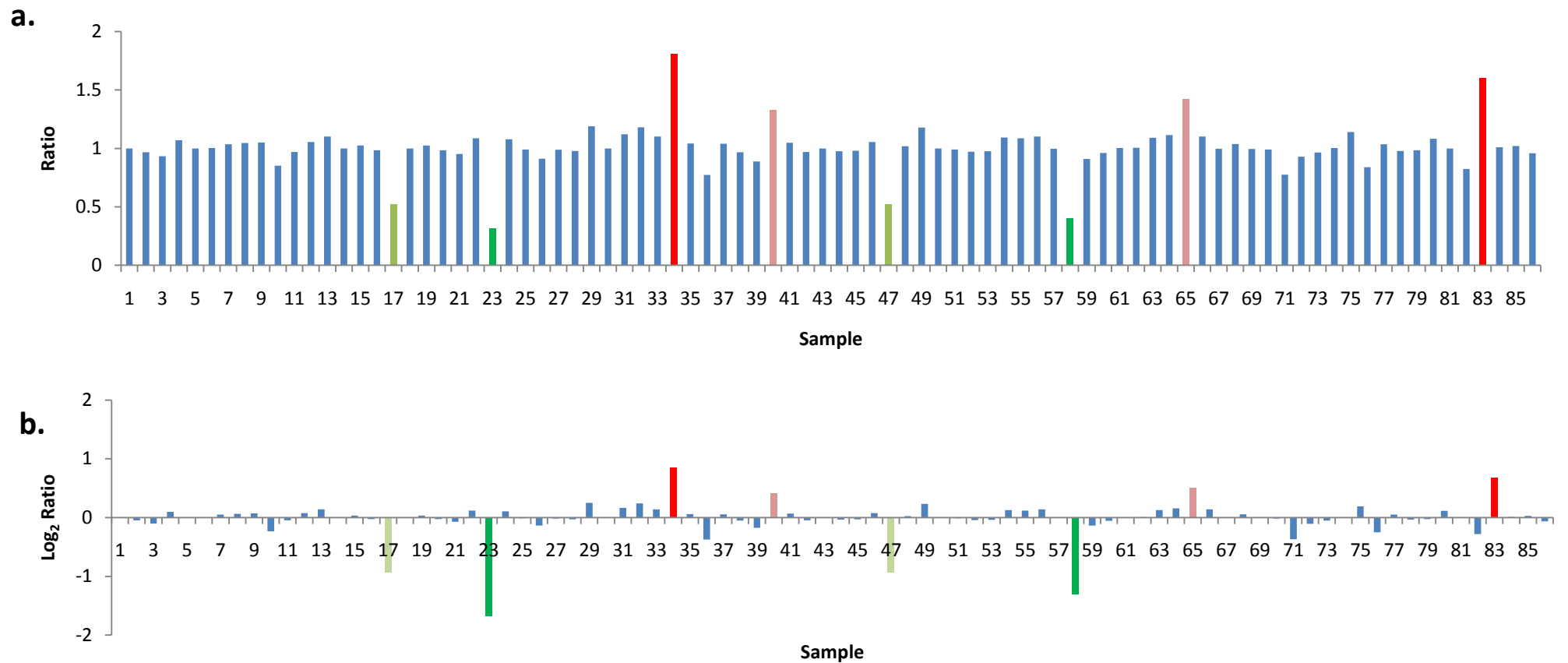


Figure 5.1: Four Types of Structural Variation

A Swedish RA cohort was genotyped using the A9 assay. Product ratio data from a single control plate is shown on the graphs above. Four distinct genotype classes can be seen, which are present in both case and control samples. a.) normalised product ratios for each sample and b.) data displayed using the \log_2 of the product ratio. Blue bars represent samples without variation, red shades indicate an increase in product ratio and green shades a decrease in product ratio.

Table 5.1: Classes of Genotype Detectable With the A9 Assay

		Chromosome 12 (i)				
		Ref Seq	A9(B) Del	A9(B) Dup	A9(A) Del	A9(A) Dup
Chromosome 12 (ii)	Ref Seq	2:2	1:2	3:2	2:1	2:3
	A9(B) Del	1:2	0:2	2:2	1:1	1:3
	A9(B) Dup	3:2	2:2	4:2	3:1	3:3
	A9(A) Del	2:1	1:1	3:1	2:0	2:2
	A9(A) Dup	2:3	1:3	3:3	2:2	2:4

Grey shading represents the four simplest variation classes, which we assume can be assigned to the four variant genotypes we have detected.

Table 5.2: Product Ratios for Detectable Genotypes

Ratio	Genotypes	Ratio Value	Log ₂ Ratio
0:2	0:2	0	-
1:1	1:1, 2:2, 3:3	1	0
1:2	1:2, 2:4	0.5	-1
1:3	1:3	0.3	-1.7
2:0	2:0	0	-
2:1	2:1, 4:2	2	1
2:3	2:3	0.6	-0.7
3:1	3:1	3	1.6
3:2	3:2	1.5	0.6

Grey shading represents the four simplest classes of variation, which we assume can be assigned to the four distinct variant genotypes we have detected.

Few compound variants, which involve more than one change in copy number, can be resolved from simple variants. For example, an A9(B) duplication on one chromosome will be balanced out by an A9(B) deletion on the other, resulting in a 2:2 product ratio, the same value produced from the reference sequence. In the same way, a sample with two A9(B) duplications, either on the same or different chromosomes would give a product ratio of 4:2, which is indistinguishable from the 2:1 ratio which would occur as a result of a single A9(A) deletion. However, based on the frequency of variants in the samples genotyped so far, we have made the assumption that compound variants are likely to be very rare. So far we have been able to detect four distinct categories of ratio change using the A9 assay, which we believe can be attributed to the four simplest forms of structural variation within the tandem duplication, namely a deletion or duplication of sequence from either unit, within which the A9 assay is located. These four categories are indicated by shading in Table 5.1.

Copy number data is often displayed using a \log_2 ratio, as on this scale a relative doubling or halving of copy number appear equivalent on this scale (1 and -1 respectively). The expected \log_2 ratio was also calculated for each genotype (Table 5.2). As before, values corresponding to the four simplest variant genotypes are shaded in grey. These were compared with a subset of the genotyping data displayed in Figure 5.1. It can be seen that the data does not correspond exactly to the predicted values; for example decreases in ratio appear as 0.5 and <0.5 rather than the predicted 0.67 and 0.5, which may be a result of experimental variation.

5.2.2 Determining Boundaries for Each Category of Variation

A density plot of the \log_2 ratios for all the Swedish RA samples (cases and controls) was created in the statistical package R, using the lattice package (Figure 5.2).

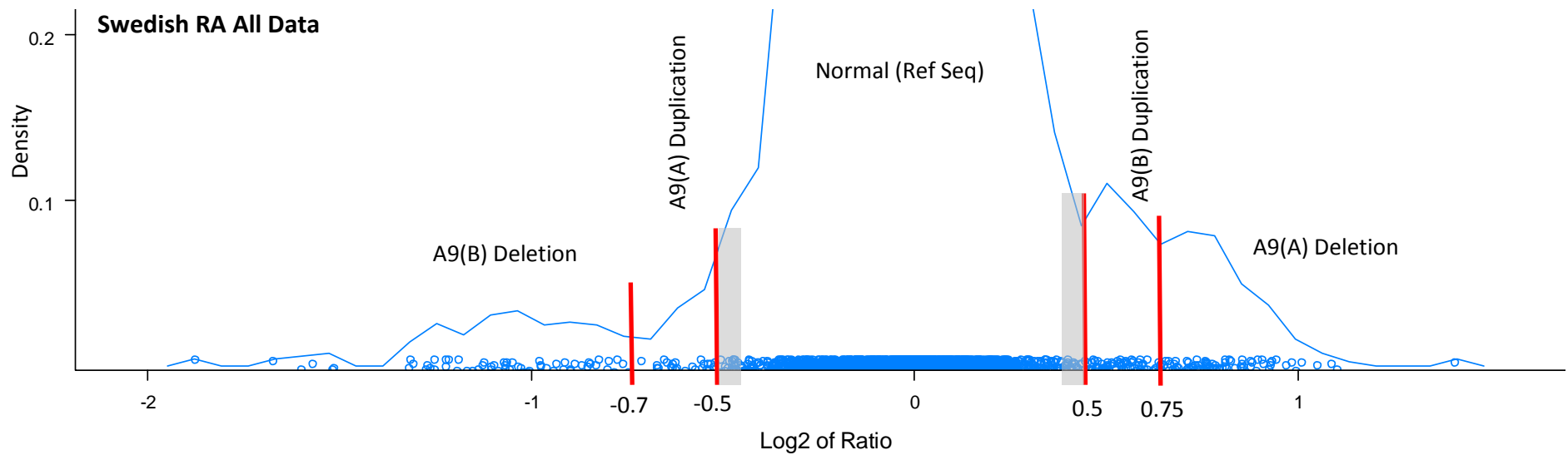


Figure 5.2: Density Plot of Swedish RA Data

A density plot showing the spread of the \log_2 ratios was used, along with the expected \log_2 ratios, to determine boundaries of each category. The red lines indicate the position of boundaries determined using the position of data clusters. The grey shading represents regions which were excluded from the analysis.

As seen previously for the product ratio values, although the data tends to form clusters, in practice these do not correspond to the expected \log_2 values. Using both the expected values and position of data on the density plot as a guide, category boundaries were determined for each of the four classes of variation detected using the A9 assay (Table 5.3). Although these are somewhat arbitrary categories, the margin of error is expected to be the same in both case and control groups, and therefore should not affect the validity of the results overall.

Table 5.3: Categories of Variation

Type of Variation	Ratio	Expected \log_2	Categories from R
A9(A) Deletion	2:1	1	>0.7
A9(B) Duplication	3:2	0.58	0.5 to 0.7
Reference Sequence	2:2	0	-0.45 to 0.45
A9(A) Duplication	2:3	-0.58	-0.5 to -0.75
A9(B) Deletion	1:2	-1	<-0.75

5.2.3 Results of a Swedish RA Association Study

Using these categories, the frequency of each type of variant in the two sample groups (case and control) was compared (Table 5.4). Density plots for each group are also shown to provide a visual representation of the data for comparison (Figure 5.3). Samples with \log_2 values between -0.45 and -0.5, as well as 0.45 and 0.5, were excluded from this analysis since it was not considered possible to determine a clear cut off point from the main data curve, and therefore data points within this interval could not accurately be placed into one category or another.

Table 5.4: Swedish Rheumatoid Arthritis Data**a.**

	<-0.75 A9(B) Deletion	-0.5- -0.75 A9(A) Duplication	-0.45- -0.5	-0.45-0.45 Normal	0.45-0.5	0.5-0.7 A9(B) Duplication	>0.7 A9(A) Deletion	TOTALS
Case	28	20	10	2252	12	38	43	2403
Control	33	8	7	1166	10	22	23	1269
TOTALS:	61	28	17	3418	22	60	66	3672

b.

	<-0.75 A9(B) Deletion	-0.5- -0.75 A9(A) Duplication	0.5-0.7 A9(B) Duplication	>0.7 A9(A) Deletion	Any Change
Case	1.18%	0.84%	1.60%	1.81%	5.42%
Control	2.64%	0.64%	1.76%	1.84%	6.87%
P-value	0.001	0.512	0.740	1	0.078
Odds Ratio	2.3				

A two-tailed chi-squared test was carried out to identify significant differences in the frequency of each variant genotype class between case and control samples. Each category was tested individually against the rest of the data. So, for example, the frequency of samples from each group (case and control) with an A9(B) deletion was compared to those without. The chi-squared test was selected because the data sets are relatively large in size. A two-tailed test was carried out because in this instance we wanted to ensure it would be possible to detect any significant differences in the frequency of variants between the two groups, regardless of the direction of the effect.

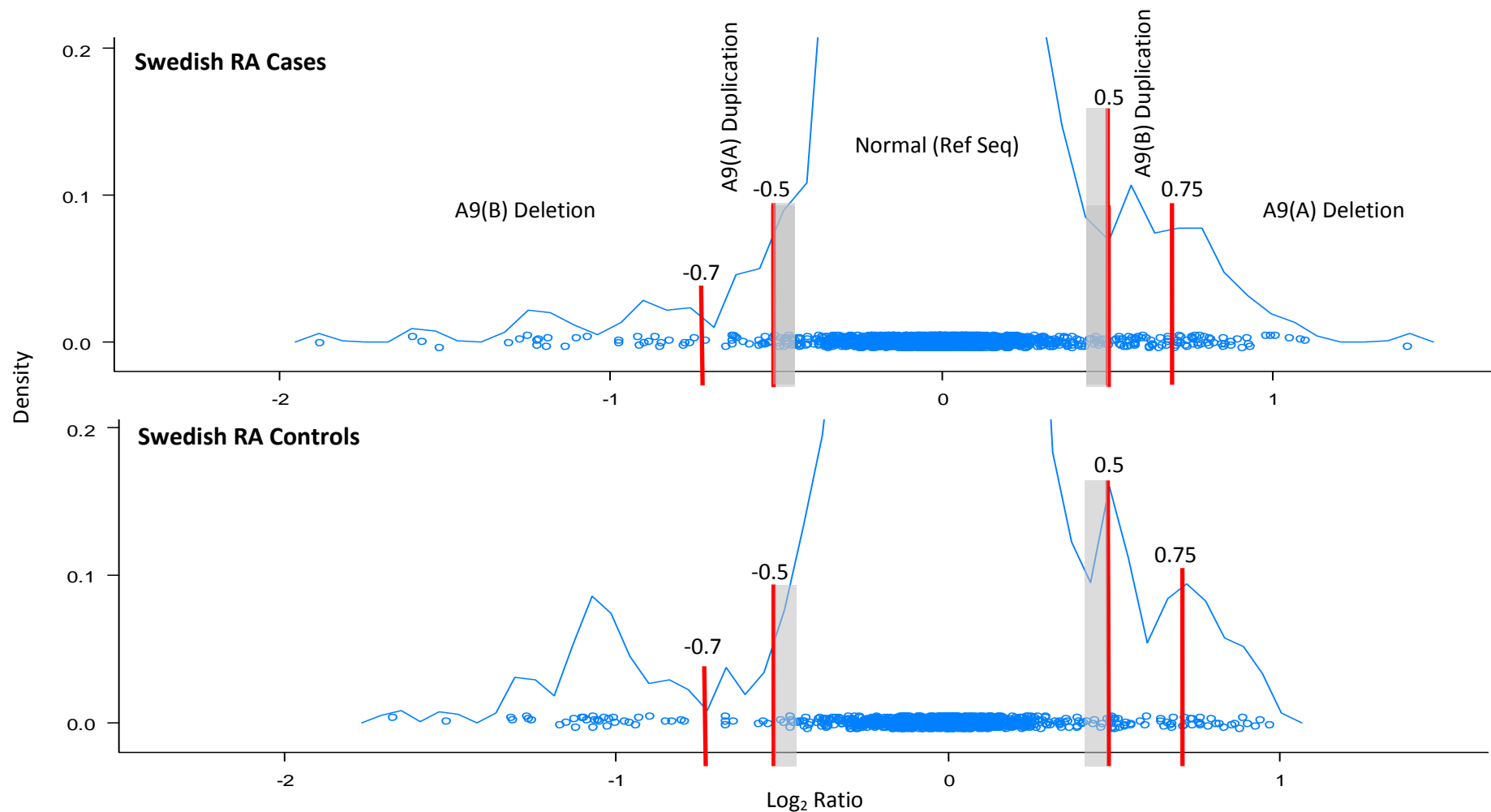


Figure 5.3: Density Plots comparing Swedish RA Case and Control Data

A density plot created in the statistical programme R showing the spread of the log₂ ratios was used, along with the expected log₂ ratios, to determine boundaries of each category. The red lines indicate the position of boundaries previously determined using the position of data clusters. The grey shading represents regions containing samples which were excluded from the analysis.

Results of this genotyping show that the frequency of any form of copy number change within the Swedish control samples is 6.87%. This decreases to 5.42% in case samples. There is a significantly higher frequency of A9(B) deletions in the control group compared to RA cases ($p = 0.001$). This result has an odds ratio of 2.3 (95% CI 1.4-3.9), which suggests that individuals with a deletion in this region are 2-3 times less likely to develop RA. The frequency of samples in the other three categories of variation are highly similar between the two groups; this is best visualised as a graph, shown in Figure 5.4. It can be seen that the frequencies of each type of variant follow the same pattern in both case and control groups, except for where a statistically significant difference has been identified.

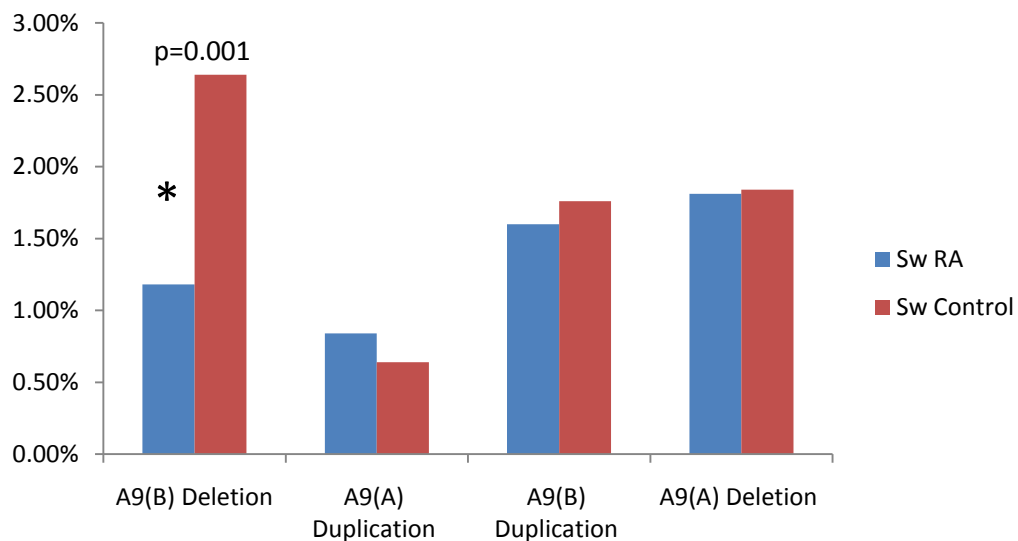


Figure 5.4: Frequency of variants in Swedish RA cases compared to controls

The frequencies of each of four types of structural variation in a Swedish RA case-control cohort were plotted on a graph. The two groups (case and control) are represented by differently coloured lines, as indicated by the key. An asterisk is used to mark a statistically significant difference in the frequency of A9(B) deletions.

5.3 UK RA Replication Cohort

In order to validate the association detected in the Swedish RA cohort, a UK RA replication cohort made up of 2235 case samples was obtained and genotyped. Since these were from a different population, a collection of UK control samples was also genotyped for comparison. This control cohort was made up of two panels of samples; 3797 from the 1958 British Birth Cohort and 2533 from the UK Blood Service collection.

5.3.1 Determining Category Boundaries for UK Samples

Genotyping was carried out using the A9 assay, as described for previous association studies in this chapter. In order to determine whether a difference in the frequency of any type of variation at 12p13.31 exists between Swedish and UK samples, data from the two control populations were compared (Table 5.5). Results show that the frequency of all four types of variation differs between the two cohorts (Figure 5.5).

Table 5.5: Frequency of Variants in UK and Swedish Control Populations

	A9(B) Deletion	A9(A) Duplication	Normal	A9(B) Duplication	A9(A) Deletion	Unclassified	TOTAL
UK	58	13	6016	234	9	0	6330
Swedish	33	8	1166	22	23	17	1269
TOTALS:	91	21	7182	256	32	17	7599

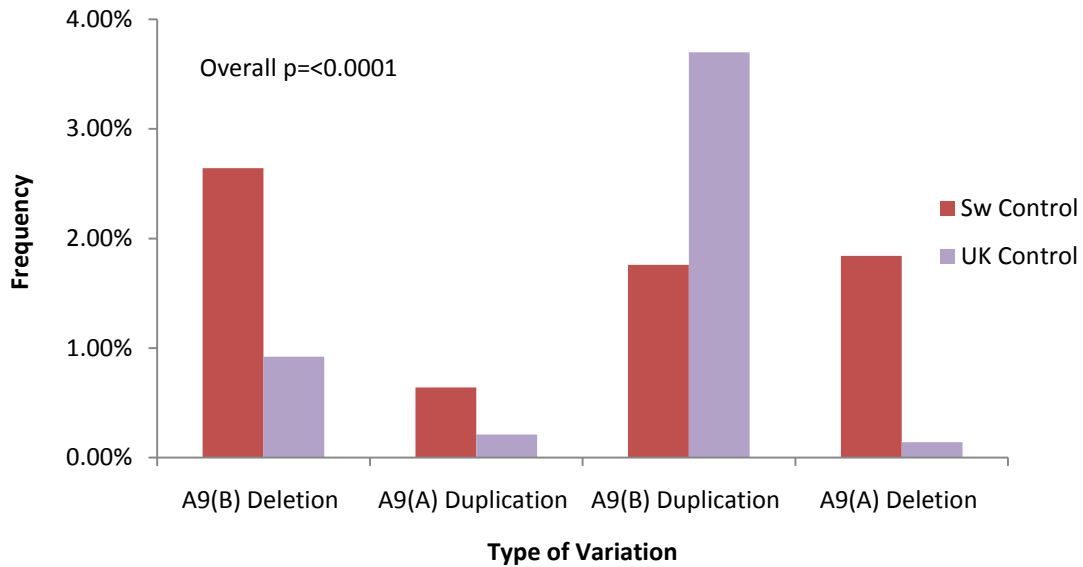


Figure 5.5: Comparing Frequency of Variants in UK and Swedish Control Populations

The frequency of each of four types of structural variation in both Swedish and UK control populations was plotted on a graph. The two populations are represented by differently coloured lines, as indicated by the key.

A chi-squared test was carried out on the 2x6 contingency table shown in Table 5.5. The result of this test gave a p-value of <0.0001 , which suggests that differences between the frequency of variants in these two populations are extremely statistically significant.

A9(B) deletions occur less frequently in the UK samples compared to the Swedish cohort, whereas A9(B) duplications are present at a higher frequency in the UK DNAs. The frequency of A9(B) deletions, the category associated with RA in the Swedish cohort, is reduced from 2.64% in the Swedish samples to 0.92% in the UK samples. Since this is the variant of interest for the replication study, the fact that it occurs at a much lower rate in the UK samples means the power of this study to reveal differences in the frequency of this variant is reduced. In order to detect a difference of the same significance as seen in the Swedish samples, a much larger sample set would be

required; however the UK RA cohort available for this study is actually slightly smaller than the Swedish cohort.

Having shown that there are statistically significant differences in the frequency of variants between the Swedish and UK control cohorts, we decided to investigate whether the categories determined for the Swedish sample sets were also appropriate for the UK samples. The genotyping data from the 1958 UK control panel was viewed on a density plot in R to determine whether the \log_2 ratios were clustering at the same points as for the Swedish samples (Figure 5.6). The density plot suggests that the UK control data have a different distribution to that from the Swedish samples. The data are less spread out from the main curve, and there are fewer points to the left hand side of this curve. There is also less noise in the UK data. This could be confirmed by re-genotyping samples from each population which surround the perceived category boundaries. Unfortunately, it has not yet been possible to carry out this investigation due to limitations concerning sample availability. In the absence of this data, we decided to revise the boundaries for each category of variation according to the position of data clusters on the UK density plot.

Initially the category boundaries were defined using the 1958 UK control panel. These were then compared to a density plot of the UK RA sample data, to determine whether the distribution of data showed the same pattern in both groups, which does appear to be the case (Figure 5.7). The two inner category boundaries were adjusted to -0.45 and 0.45, due to the fact that the main data curve is more compact in the UK samples. Since there is less noise in the UK data, there appears to be less of an overlap between normal and variant samples.

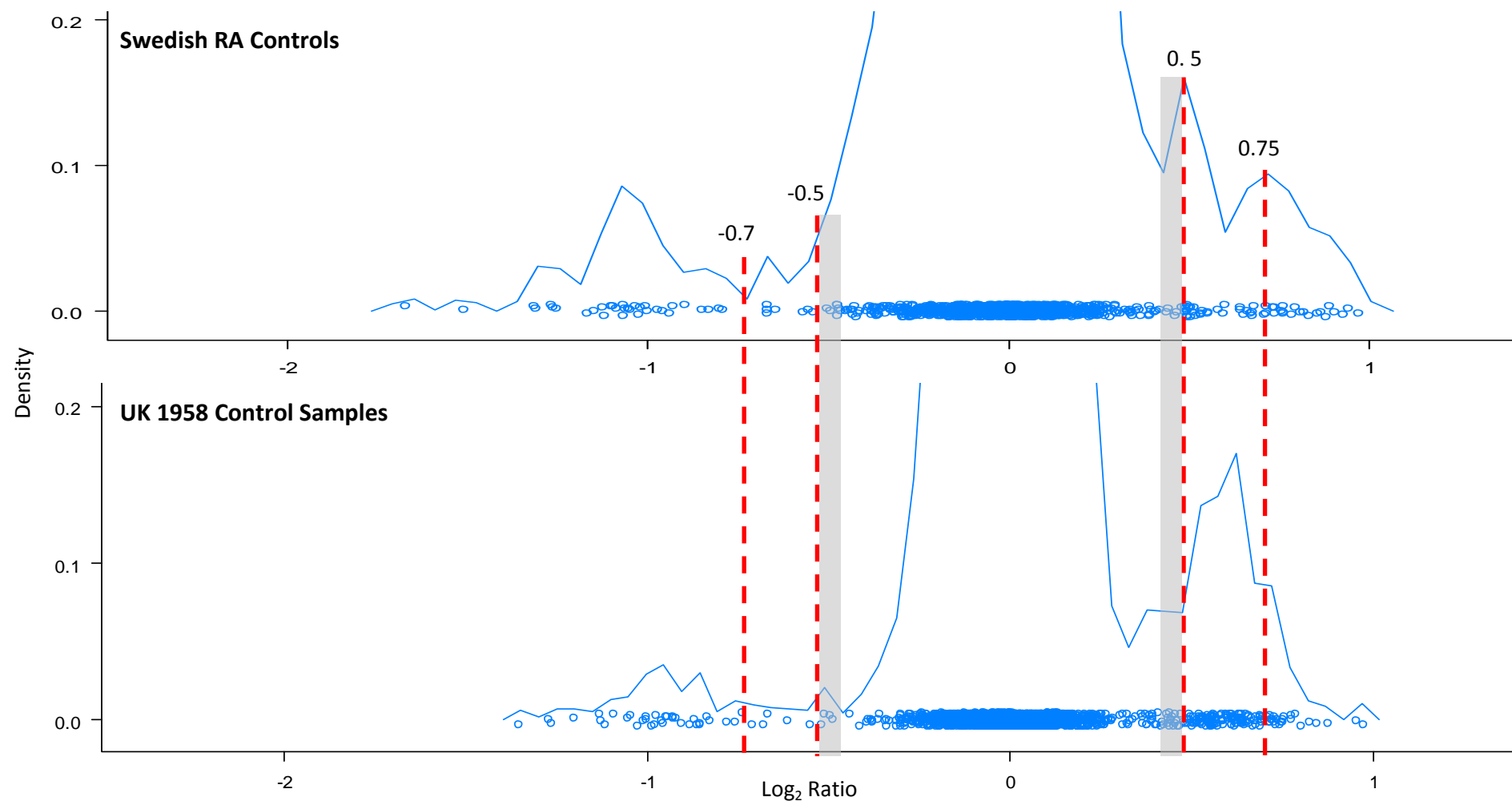


Figure 5.6: Comparing UK and Swedish Control Sample Data to Determine Category Boundaries

Data from the largest UK control panel was viewed as a density plot created in the programme R using the package lattice. The distribution of the UK data is shown above, compared to that for the Swedish RA control data. Boundaries used for the Swedish RA study are indicated by broken red lines, with the grey shading representing samples which were excluded from the analysis.

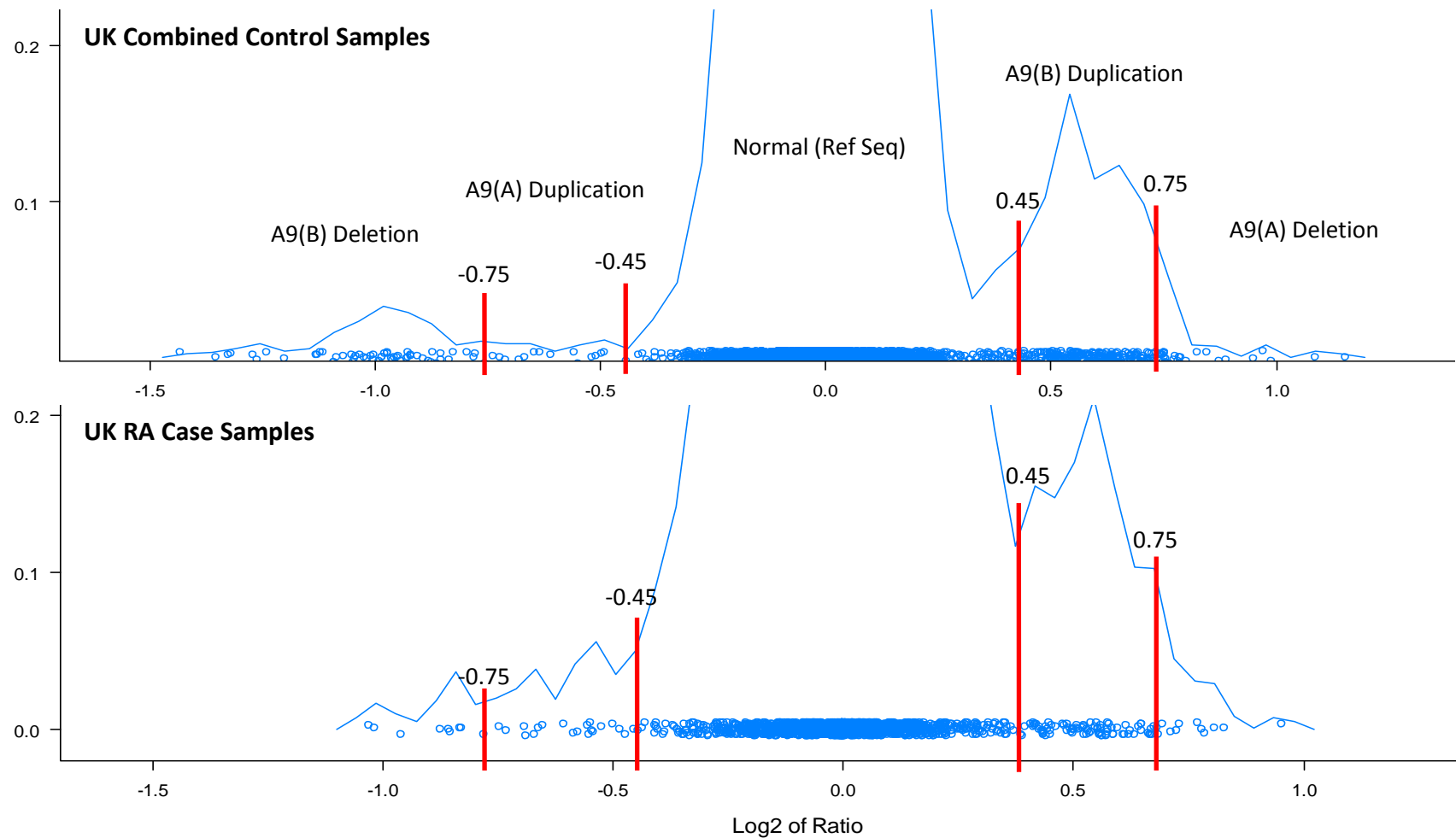


Figure 5.7: Density Plots of UK RA Case and Control Data

Density plots of the genotyping data from each group (case and combined control samples) were created in the statistical program R using the package lattice. Red lines indicate the position of the boundaries for each category, which were determined using the position of data clusters on the density plot.

Due to the lower level of noise in the UK data, we did not consider it necessary to define a ‘grey area’ of samples to be excluded from analysis for this population as had been required for the Swedish data.

The upper category boundary which defines A9(A) deletions was adjusted from 0.7 to 0.75. The lower boundary used to categorise A9(B) deletions (previously -0.75) was more difficult to determine, as data to the left of the curve does not cluster in the same way as for the Swedish samples (Figure 5.2). This is probably due to the fact that there are fewer points on this side of the curve in the UK cohorts. We decided that it was appropriate to retain this boundary at -0.75, since this appears to mark the edge of a region of data clustering.

5.3.2 Replicating the A9(B) Deletion Association with RA

The purpose of the UK RA study was primarily to investigate whether the association of A9(B) deletions with RA detected in the Swedish cohort could be replicated. Therefore, for this purpose we were only interested in determining the frequency of A9(B) deletions in each group. Using the newly defined boundaries, the frequency of this type of variation was calculated for both case and control groups (Table 5.6). Since the frequency of this class of variation is lower in the UK samples than in the Swedish cohort, the power to detect an association is reduced. Therefore, we chose to use a one-tailed chi-squared test to determine the statistical significance of any differences in the frequency of A9(B) deletions between the two groups, as this is a more sensitive test.

Table 5.6: UK RA Data

a.

	<-0.75 A9(B) Deletion	>-0.75 Other	TOTALS
Case	9	1837	1846
1958 Controls	34	3763	3797
UK Blood Controls	24	2509	2533
TOTALS:	67	8109	8176

b.

	<-0.75 A9(B) Deletion
Case	0.49%
Combined Controls	0.92%
P-value	0.036
Odds Ratio	1.90

The one-tailed test was also a suitable choice because we were looking for change in the frequency of A9(B) deletions in a particular direction (i.e. an increase in control samples compared to cases).

Results show that, as was the case in the Swedish cohort, there is a significantly higher frequency of A9(B) deletions in the UK control group compared to RA cases ($p=0.036$). The odds ratio for this is 1.90 (95% CI 0.98-3.82). Comparing the frequency of these variants in the two populations reveals that, despite the fact that A9(B) deletions occur at a lower frequency in UK samples, the frequency of A9(B) deletions in each control group is approximately half of that seen for RA patients in the same population (Table 5.7). This data shows it has been possible to replicate the association seen previously,

thereby validating our previous discovery that copy number variation at 12p13.31, specifically A9(B) deletions, is associated with susceptibility to RA.

Table 5.7: Summary of RA results from Swedish and UK studies

	A9(B) Deletion	Any Change
Swedish RA	1.18%	5.42%
Swedish Controls	2.64%	6.87%
UK RA	0.49%	5.36%
UK Combined Controls	0.92%	4.96%

5.3.3 Investigating Other Forms of Variation

Having replicated the A9(B) deletion association, we went on to investigate the frequency of other forms of variation in the UK RA sample group (Table 5.8). A9(B) deletions were excluded from this analysis since these have already been examined. For A9(B) deletions and A9(B) duplications, a two-tailed chi-squared test was performed to investigate the significance of any differences in variant frequency between case and control groups. However, in the case of A9(A) deletions and A9(A) duplications, a two-tailed Fisher's exact test was employed due to the low number of samples in these categories. Results show that the frequency of A9(A) duplications in UK RA patients is significantly higher than in matched control samples. This result is extremely statistically significant, however, it is somewhat surprising that an association of this strength was not also seen in the Swedish RA cohort.

Table 5.8: UK RA Frequency Data for all Variants

a.	<-0.75 A9(B) Deletion	-0.45- -0.75 A9(A) Duplication	-0.45-0.45 Normal	0.45-0.8 A9(B) Duplication	>0.8 A9(A) Deletion	TOTALS
Case	9	18	1747	69	3	1846
1958 Controls	34	10	3608	140	5	3797
UK Blood Controls	24	3	2408	94	4	2533
TOTALS:	67	31	7763	303	12	8176

b.	<-0.75 A9(B) Deletion	-0.45- -0.75 A9(A) Duplication	0.45-0.8 A9(B) Duplication	>0.8 A9(A) Deletion	Any Change
Case	0.49%	0.98%	3.74%	0.16%	5.36 %
Combined Controls	0.92%	0.21%	3.70%	0.14%	4.96 %
P-value (Fisher)	-	<0.0001	-	0.74	-
P-value (Chi squared)	-	-	0.93	-	0.49
Odds Ratio	-	4.78	-	-	-

Due to the fact that the frequency of A9(A) variants in the combined control group is so low, it may be that this result is due to a sampling error. Replication of this study using a larger UK RA cohort would be necessary to validate this result.

5.3.4 The Effect of Adjusting Category Boundaries

When determining the UK category boundaries, we found it difficult to position the lower boundary, which is used to classify A9(B) deletions. Ultimately, after examination of the density plots, it was decided that this should remain at -0.75, the same position which had been used in the Swedish investigations. However, since this

marks the category of interest for this study, we decided to investigate the effect of setting the A9(B) deletion boundary at three different values (-0.8, -0.75 and -0.7). For each value, the number of variants in the A9(B) deletion category was calculated in both case and control groups, and the significance of the differences in frequency were determined using a one-tailed chi-squared test (Table 5.9). The results show that a small adjustment of the boundary either way could have a considerable impact on the p-value, and therefore affect whether the results are considered statistically significant. The effect of accurate positioning is therefore something that needs to be carefully considered in the future when defining the boundaries for each category of variation.

Table 5.9: Effect of Changing Category Boundaries for UK Samples

	No. Samples	<-0.7	<-0.75	<-0.8
Case	1846	11	9	8
Combined Controls	6330	60	58	53
P-value		0.076	0.036	0.038
Odds Ratio		1.60	1.90	1.94

5.4 Swedish Psoriasis Cohort

Having detected a region of copy number variation which is associated with RA susceptibility, we went on to investigate whether this copy number variant is also associated with psoriasis. A Swedish psoriasis cohort was provided by collaborators at the Karolinska Institutet, and the same Swedish control panel genotyped for the RA study was used for comparison. Samples were genotyped using the A9 assay and analysis carried out as for the RA cohort (Table 5.10). A two-tailed chi-squared test was used to determine the statistical significance of differences in the frequency of each type of variant between case and control groups.

Table 5.10: Swedish Psoriasis Data

a.

	<-0.75 A9(B) Deletion	-0.5- -0.75 A9(A) Duplication	-0.45- -0.5	-0.45-0.45 Normal	0.45-0.5	0.5-0.7 A9(B) Duplication	>0.7 A9(A) Deletion	TOTALS
Case	14	11	8	1076	3	22	12	1146
Control	33	8	7	1166	10	22	23	1269
TOTALS:	47	19	15	2242	13	44	35	2415

b.

	<-0.75 A9(B) Deletion	-0.5- -0.75 A9(A) Duplication	0.5-0.7 A9(B) Duplication	>0.7 A9(A) Deletion	Any Change
Case	1.23%	0.97%	1.94%	1.06%	5.20%
Control	2.64%	0.64%	1.76%	1.84%	6.87%
P- value	0.013	0.365	0.740	0.113	0.088
Odds Ratio	2.16				

The results of this study show that a deletion of the region encompassing A9(B) occurs at a significantly higher frequency in control samples compared to psoriasis cases ($p=0.013$, odds ratio 2.16: 95% CI 1.2-4.1). This suggests that variation at this locus is associated with psoriasis, as well as RA. Viewing the results graphically (Figure 5.8) shows that the pattern of frequencies for all four types of variation is similar between case and control cohorts, except for A9(B) deletions, where a statistically significant difference has been identified between the two groups. There is also appears to be a change in frequency of A9(A) deletions, however this difference is not statistically significant.

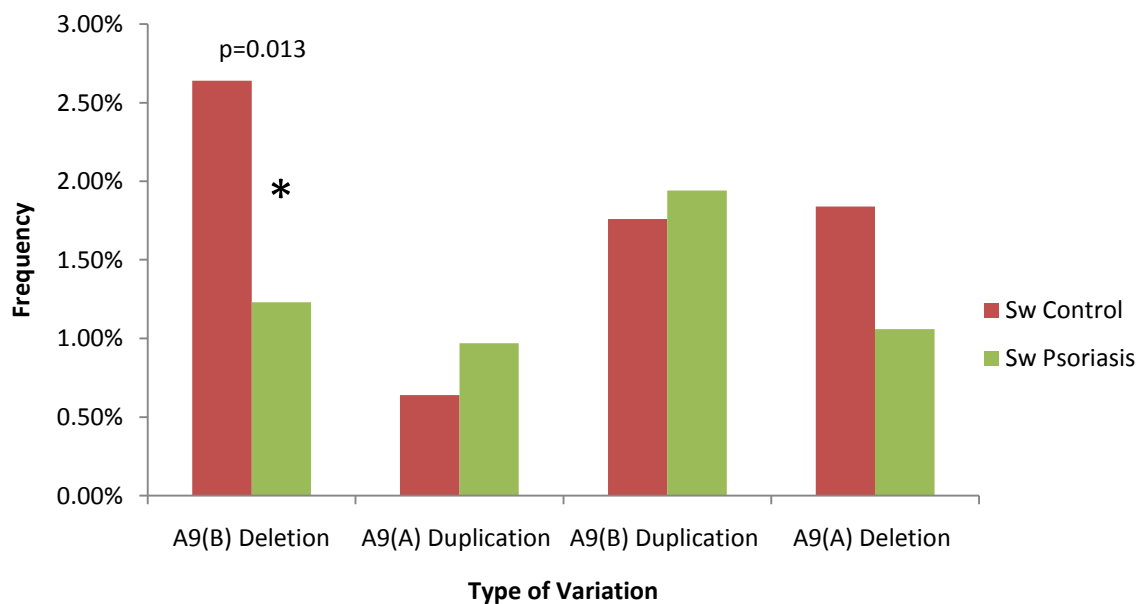


Figure 5.8: Swedish Psoriasis Cohort Data

The frequencies of each of four types of structural variation in a Swedish psoriasis case-control cohort were plotted on a graph. The two groups (case and control) are represented by differently coloured lines, as indicated by the key. An asterisk is used to mark a statistically significant difference in the frequency of A9(B) deletions.

5.5 Cardiovascular Disease

To test whether structural variation within this region is also involved in susceptibility to cardiovascular disease (CVD), a UK CVD cohort of DNA samples from 2339 patients provided by collaborators in the Department of Cardiovascular Sciences, University of Leicester, was obtained and genotyped using the A9 assay. The two UK control sample groups previously employed for the UK RA study were used again here for comparison. The frequency of each type of variation was determined by placing the genotyped samples in categories of variation, using the same cut-off boundaries as for the UK RA study described in Section 5.6. (Table 5.11).

Table 5.11: UK Cardiovascular Disease Data

a.

	<-0.75 A9(B) Deletion	-0.45- -0.75 A9(A) Duplication	-0.45-0.45 Normal	0.45-0.8 A9(B) Duplication	>0.8 A9(A) Deletion	TOTALS
Case	14	11	2222	84	8	2339
1958 Controls	34	10	3608	140	5	3797
UK Blood Controls	24	3	2408	94	4	2533
TOTALS:	72	24	8238	318	17	8669

b.

	<-0.75 A9(B) Deletion	-0.45- -0.75 A9(A) Duplication	0.45-0.8 A9(B) Duplication	>0.8 A9(A) Deletion	Any Change
Case	0.60%	0.47%	3.59%	0.34%	5.00%
Combined Controls	0.92%	0.21%	3.70%	0.14%	4.96%
P-value	0.15	0.037	0.82	0.062	0.92
Odds Ratio		2.3			

Since on this occasion we did not want to restrict our investigations to a single variant or direction of effect, a two-tailed chi-squared test was used to determine the significance of the data. Results suggest that there is a significant difference in the frequency of A9(A) duplications between the case and control groups, with the variant occurring at a higher frequency in CVD case samples ($p=0.037$, OR 2.3; 95% CI 1.03-5.10). (Figure 5.9). However, as for the UK RA data, the low frequency of A9(A) variants in the control group must be taken into account when considering this result. In order to confirm whether this is a true association or a sampling error, it would be necessary to validate this association using a larger sample group.

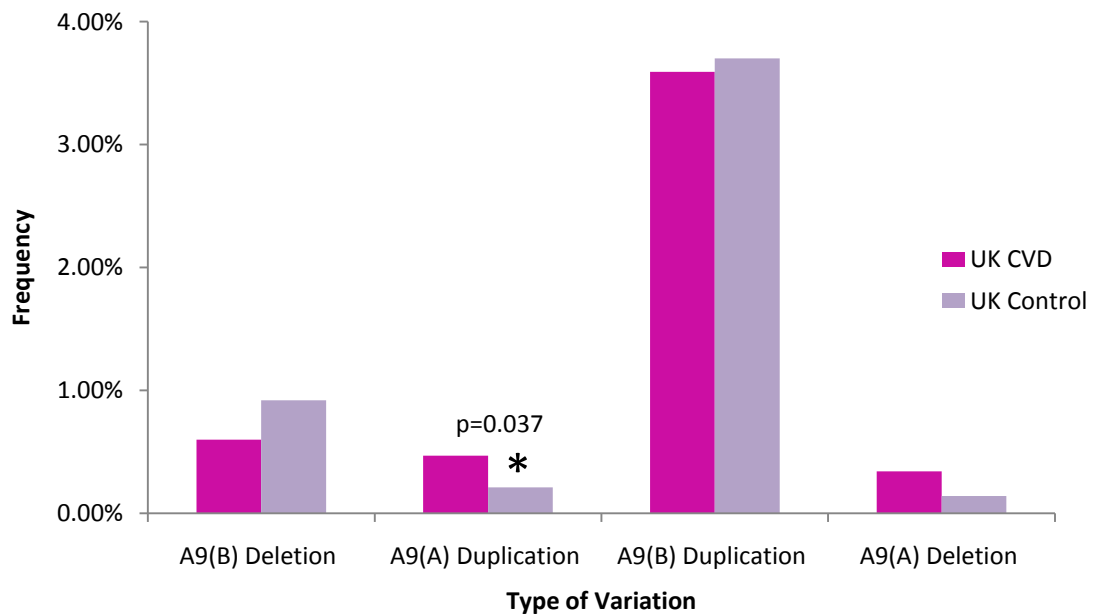


Figure 5.9: Frequency of Variation in UK CVD Case-Control Cohort

The frequencies of each of four types of structural variation in a UK CVD case-control cohort were plotted on a graph. The two groups (case and control) are represented by differently coloured lines, as indicated by the key. An asterisk is used to mark a statistically significant difference in the frequency of A9(A) duplications.

5.6 Summary and Discussion

Work described in this chapter has concentrated on investigating the association of copy number variation within a novel tandem duplication on chromosome 12p13.31 with complex disease. We have shown that the frequency of deletions within the region containing the A9(B) assay occurs at a significantly higher frequency in Swedish control samples compared to RA ($p = 0.001$) and psoriasis ($p = 0.013$) case samples from the same population. This apparent protective effect has an odds ratio of 2.3 (95% CI 1.4-3.9) in RA and 2.16 (95% CI 1.2-4.1) in psoriasis; this has also been replicated in a UK RA cohort ($p = 0.036$, OR 1.90; 95% CI 0.93-3.82). The size of this effect is comparable to that reported for the 1858T allele of *PTPN22*, the second most significant RA risk loci after the HLA genes, which has an odds ratio of 1.8 (95% CI 1.2-2.8) (Michou *et al*, 2007).

Risk loci for autoimmune diseases are known to often cluster together in the genome. We have already shown that a deletion within 12p13.31 is also associated with psoriasis ($p = 0.013$, OR 2.16 (95% CI 1.2-4.1)), another common autoimmune disorder, although this result has not yet been replicated. It would be interesting to investigate whether this region is also involved in susceptibility to other autoimmune disorders, for example Type I diabetes and SLE, which have previously been shown to share other risk loci with RA (Jawaheer *et al*, 2001). It is interesting to note that *SLC2A3* has previously been suggested as one of a number of genes for which leukocyte expression levels can be used to distinguish children with active juvenile rheumatoid arthritis (JRA) (Jarvis *et al*, 2004). Studies are underway within our research group to investigate whether copy number variation at 12p13.31 is associated with JRA.

Due to the presence of QTLs across this tandem duplication associated with factors such as blood pressure as well as insulin, glucose and cholesterol levels, we consider it is possible that genes within this region may also be associated with other complex disorders. Endophenotypes such as blood pressure and cholesterol levels are indicators of susceptibility to cardiovascular disease, which often shows co-morbidity with RA. For this reason we also investigated whether variation in this region is also associated with cardiovascular disease in a UK CVD cohort. Results of this study revealed a significant association between duplications within A9(A) and CVD disease ($p < 0.037$, OR 2.3; 95% CI 1.03-5.10), suggesting a possible association of variation in the A unit of the tandem duplication with CVD. There was also a significantly higher frequency of A9(A) duplications in UK RA cases compared to controls ($p < 0.0001$). However, this effect was not seen in the Swedish RA cohort, which suggests that it is either a difference between the two populations, or a statistical anomaly. Due to the low frequency of A9(A) duplications in both case and control groups, in particular the UK Blood Service control group, it seems likely that this association may be a statistical anomaly; to investigate this further, it would be necessary to repeat this study using a much larger UK cohort.

For this thesis, a subset of the 1958 British Birth Cohort has been genotyped and used as a UK control group. Detailed endophenotypic data have been collected from the individuals in this study over a number of decades. There are plans to complete genotyping of the remainder of these samples (approximately 3900 samples) shortly. Once this has been completed, it will be possible to investigate whether there are any correlations between CNV in the 12p13.31 region and phenotypes such as blood pressure, glucose and cholesterol levels, characteristics for which there are QTLs within this locus.

Comparing the frequencies of each of the four types of variation identified within the 12p13.31 tandem duplication in the two populations studied (Swedish and UK) revealed differences between the two groups. These differences occur in both the case and control samples. Overall, we detected an increased frequency of any type of variant in the Swedish samples compared to the UK cohort. In regard to A9(B) deletions, the variant of interest from the Swedish studies, it can be seen that the ratio of samples containing this variant compared to non-variant samples between case and control groups is very similar in both populations. Since the frequency of A9(B) deletions is lower in the UK samples compared to the Swedish cohorts, the power of the study to detect an association is reduced. This means that, ideally, a greater number of UK samples need to be genotyped to detect an effect of the same size as was seen in the Swedish population.

Due to differences between the two populations, the values assigned as the boundaries for each category of variation in the UK study were reviewed. To determine how the subsequent alteration of these boundaries influenced the results, we investigated the effect of using different cut-off points to categorise A9(B) deletions. This was shown to affect the statistical significance of the results, highlighting the importance of accurately assigning the categories of variation.

Up to this point, all genotyping of structural variation within the tandem duplication on chromosome 12p13.31 has been carried out using a single assay. This is only able to provide information about variation at a single point in the sequence, so it has not yet been possible to draw any conclusions as to whether variation at the 12p13.31 locus involves the entirety of each unit of the tandem duplication, smaller sections within them or a region which encompasses segments from both units. There may also be other variations within this region, which do not involve sequence at the position of the A9

assay and therefore would not be detected using this method. In order to investigate copy number variation within this tandem duplication further and also to identify the location of historical recombination breakpoints, additional assays are therefore required. The development of such assays forms the basis of the next chapter.

Chapter 6

Investigating Historical Recombination Events at 12p13.31

6.1 Introduction

Previous chapters described the characterisation of a copy number variable tandem duplication located on chromosome 12p13.31. Genotyping with a single assay, the A9 assay, enabled the detection of four variant genotype classes for this region. These are visualised as changes in the relative amount of product produced from equivalent single sites in each unit of the tandem duplication. We hypothesise that these four variant classes reflect deletion or duplication of a section of either unit. Since all genotyping so far has been carried out using the same assay, we have only been able to investigate copy number of the DNA sequence at one particular location within each unit. So for example, in the case of a putative deletion, we have evidence to suggest that the sequence amplified by the A9 assay is absent, but this result cannot provide any information about the size or boundaries of the deleted sequence.

It is likely that the two units of the tandem duplication were once identical (after the presumed initial duplication event) but have since diverged. The most obvious indicator of this divergence is the size difference between the two regions; unit A is 100 kb in size, whereas unit B is closer to 140 kb. The similarity between the two units is structured as blocks of near identical sequence, which are interrupted by stretches of sequence (mostly Alu repeat elements) present in one unit but not the other. The regions

of shared sequence show an average similarity of 94%, whereas the overall sequence identity between the two units is considerably lower, at around 50-60%.

Due to the high degree of sequence similarity which exists between the two units, the simplest mechanism by which CNV at this locus may have originated is non-allelic homologous recombination (NAHR) during meiotic division (Figure 6.1). This process involves the misalignment of low copy repeat regions, in this case the two units of the tandem duplication. NAHR between the two units of the tandem duplication would result in the gain or loss of a segment of DNA from both units.

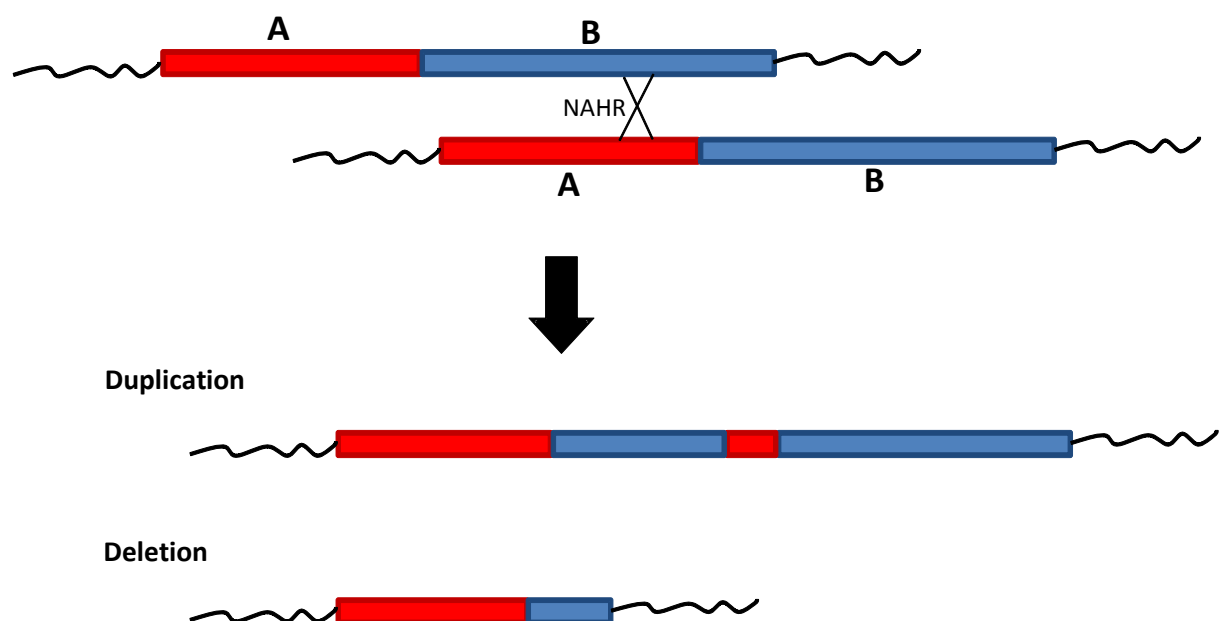


Figure 6.1: NAHR within the Tandem Duplication at 12p13.31

NAHR between the two units of the tandem duplication would result in the formation of two reciprocal products, containing either a duplication or deletion of a segment from each unit. The two units of the tandem duplication are represented as coloured bars; a red bar is used for unit A and a blue bar represents unit B. A black cross is used to indicate a hypothetical recombination event.

Nimblegen arrayCGH data from within this region supports this theory, since data produced from the comparison of one pair of DNAs revealed a structural variant that spans a section from each of the two units (discussed in Chapter 4). It is not clear how many different structural variants are segregating in today's populations. Genotyping data produced using the A9 assay has suggested four distinct classes of variation; however, as discussed in Chapter 5.2, this assay is not able to resolve all possible variants.

To learn more about how the sequence identity between the two units of the tandem duplication is structured, we carried out detailed examination of regions of sequence shared between the units. Following this, a new series of assays was designed in order to investigate copy number changes across the tandem duplication, and potentially reveal possible sites of recombination.

6.2 Characterisation of Sequence Identity

Identifying the largest stretches of sequence shared between the two units of the tandem duplication may help to predict the location of historical recombination events, and aid the development of further assays. To learn more about the structure of sequence identity between the two units of the tandem duplication, the sequences were compared using two complementary *in silico* methods; dot plots and BLAT alignments.

6.2.1 Dot Plots

Dot plots allow the direct comparison of two sequences (for example protein or DNA) and provide a visual representation of their similarity. The sequences are screened for exact sequence matches using ‘windows’ of a specified length, e.g. for DNA this could be 50 bp. The larger the window, the more stringent the process is, since there is less likelihood of a match occurring by chance. The resulting data are displayed on a graph, with two axes, one representing each sequence. Where there is a match of the specified length and identify, a dot is drawn in the corresponding position on the graph. So, for example, two identical sequences would produce a diagonal line, whereas a region unique to one of the sequences would appear as a break in this line.

Dot plots were created using the online software Gepard (Krumšek *et al*, 2007) (Figure 6.2). Two different window sizes were used. It can be seen that setting the search window to 100 bp reveals many regions of sequence identity, which tend to localise in a number of clusters. If the stringency is increased to detect exact matches of greater than 400 bp, only three regions remain; these represent the largest regions of identical sequence shared between the two units of the tandem duplication (Figure 6.3).

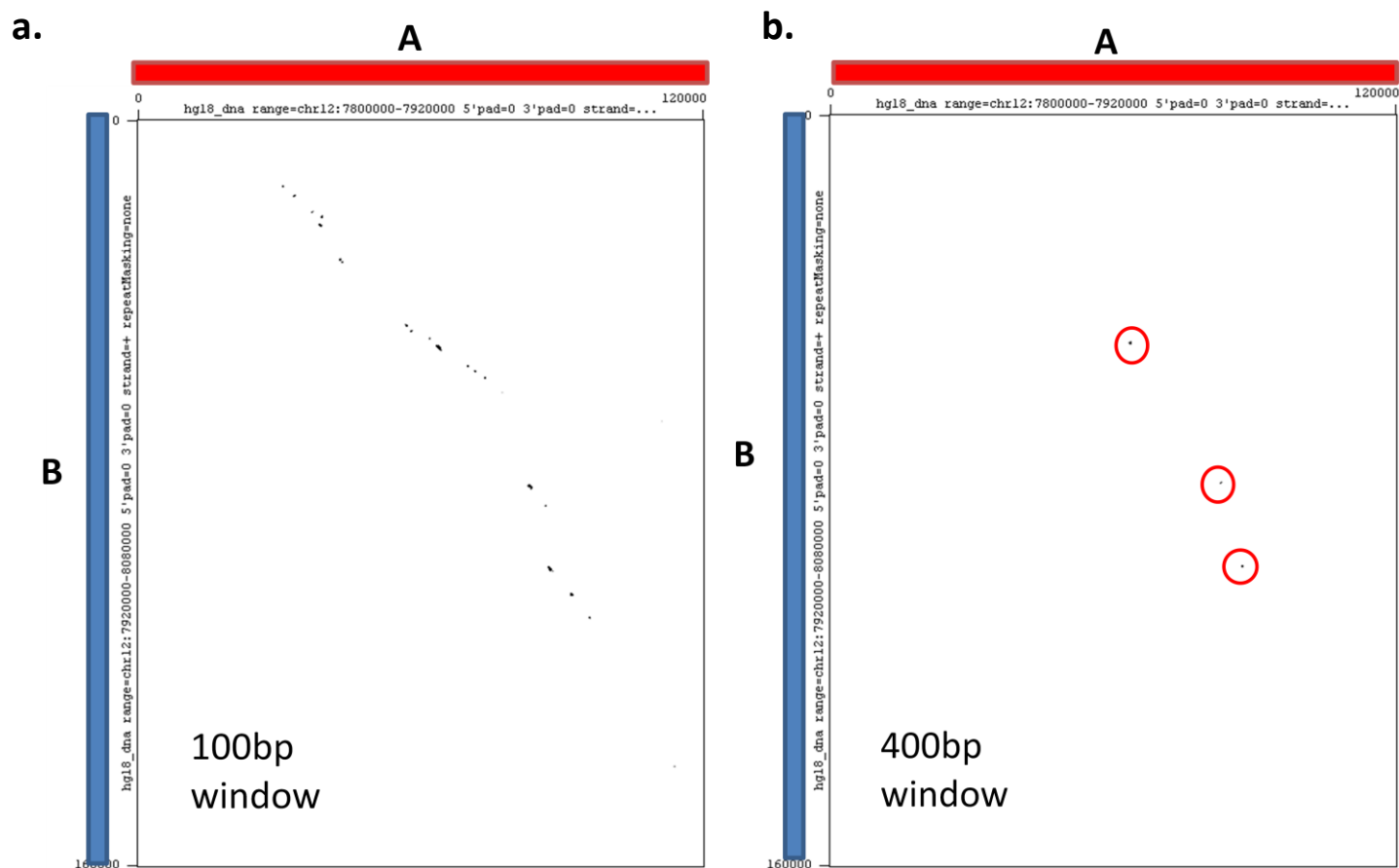


Figure 6.2: Dot Plots to show Regions of Sequence Identity

Dot plots were created to compare regions of sequence identity in the two units of the tandem duplication. A red bar indicates the axis representing unit A, whereas a blue bar shows the axis for unit B. Each dot on the graph represents a match of the required window size between the two units. Two different sized windows were used; a.) 100 bp, and b.) 400 bp. In b.) only three regions of sequence identity are displayed, these are circled in red.

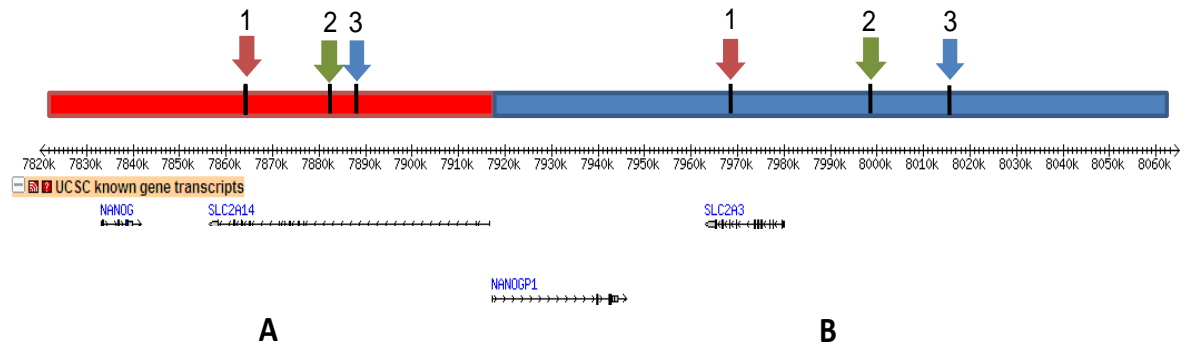


Figure 6.3: The Location of the Longest Regions of Identical Sequence

The location of the longest stretches of identical sequence shared between the two units was identified using a dot plot (Figure 6.2b). The coloured bars represent the two units of the tandem duplication (Red represents A, and blue B) and the genes are shown in UCSC genome browser. The locations of the longest regions of shared sequence are marked on the diagram by black lines. Each of the three regions are numbered and marked with a different coloured arrow, indicating the position in each unit.

6.2.2 BLAT Alignments

BLAT (Blast-like Alignment Tool) alignments were used to further investigate the sequence similarity between the two units of the tandem duplication (Kent, 2002). BLAT works in several stages. First, an index is created which contains the location and sequence of a set of non-overlapping unique DNA segments that represent the whole genome (the default setting is for 11 basepair sequences). Sequences within highly repetitive regions, i.e. satellite DNA, which occur above a certain frequency threshold, are discarded. The contents of this index are aligned to the query sequence and adjacent matches from the same genomic location are joined together to produce longer fragments. BLAST (Basic Local Alignment Search Tool) works the opposite way around, scanning through the genome for matches to the query sequence, rather than scanning the query sequence for regions of sequence which have a match in the genome index. This approach means that BLAST keeps the entire genome sequence in memory,

whereas BLAT requires only enough 11mers to represent the entire genome. Consequently the memory demands of BLAT are lower, meaning that this method is faster than alignment tools such as BLAST.

The DNA sequence of each of the two units of the tandem duplication was split into fragments in a tiling path across each unit using a Perl script (Owen Lancaster). Six separate sequence libraries were created, containing the entire sequence of the tandem duplication separated into different fragment sizes (100 bp, 500 bp or 1 kb) both with and without repeat masking prior to fragmenting. In cases where repeat-masking was carried out, any fragments which contained repeat elements (represented by runs of N residues in the sequence) were discarded. The sequences were then aligned to the current genome build (build NCBI36/hg18, March 2006). For this purpose the stand-alone BLAT software (Kent, 2002) was downloaded and used to run the search locally, since high volume BLAT searches are not permitted on the UCSC BLAT server.

The results of the alignments were filtered to remove hits which did not correspond to the tandem duplication under investigation, as well as matches to the position of the fragments themselves and hits with a sequence identity of less than 80%. This left the longest sequence matches in each category (for example, those equal to 100 bp in the 100 bp category and those >400 bp in the 500 bp category). Separate alignments were performed for each of the six sets of sequence fragments. In the case of the 100 bp fragments which had not been repeat masked, the short length meant that some fragments contained only repetitive element sequence and therefore produced a vast number of uninformative matches. Therefore, for this fragment length, only the results from the repeat masked data were used for analysis. However, since the majority of repetitive elements are less than 400 bp in size, this was not as much of a problem for the 500 bp fragments. These were likely to contain enough non-repeat sequence to

produce unique matches, and so in this case, the non-repeat masked sequences were used for analysis. The results of the BLAT alignments are displayed graphically in Figure 6.4. Since there were few regions of sequence identity greater than 500 bp, the 1 kb fragment alignments did not provide a great deal more information than the 500 bp fragments, so these data are not shown.

As was shown by the dot plots, 100 bp alignments produce large clusters of similar sequence, whereas increasing the stringency, in this case to 500 bp, produces considerably fewer matches of high identity. The data tends to form clusters representing regions of high identity, and it can be seen that these clusters form similar patterns in both units of the tandem duplication. In unit B the clusters are spread over a larger distance, due to the fact that the regions of sequence identity have been broken up by the insertion or deletion of sequence, largely repetitive elements.

Data produced from both the dot plots and BLAT alignments illustrate the numerous regions of sequence identity which exist between the two units of the tandem duplication, many of which are 400-500 bp in length and share over 90% sequence identity. These shared blocks of sequence are dispersed across the two units. Considering these results alongside our hypothesis that NAHR may occur (or has previously occurred) between the two units of the tandem duplication, this means that there are many possible locations at which historical recombination events might be expected to occur.

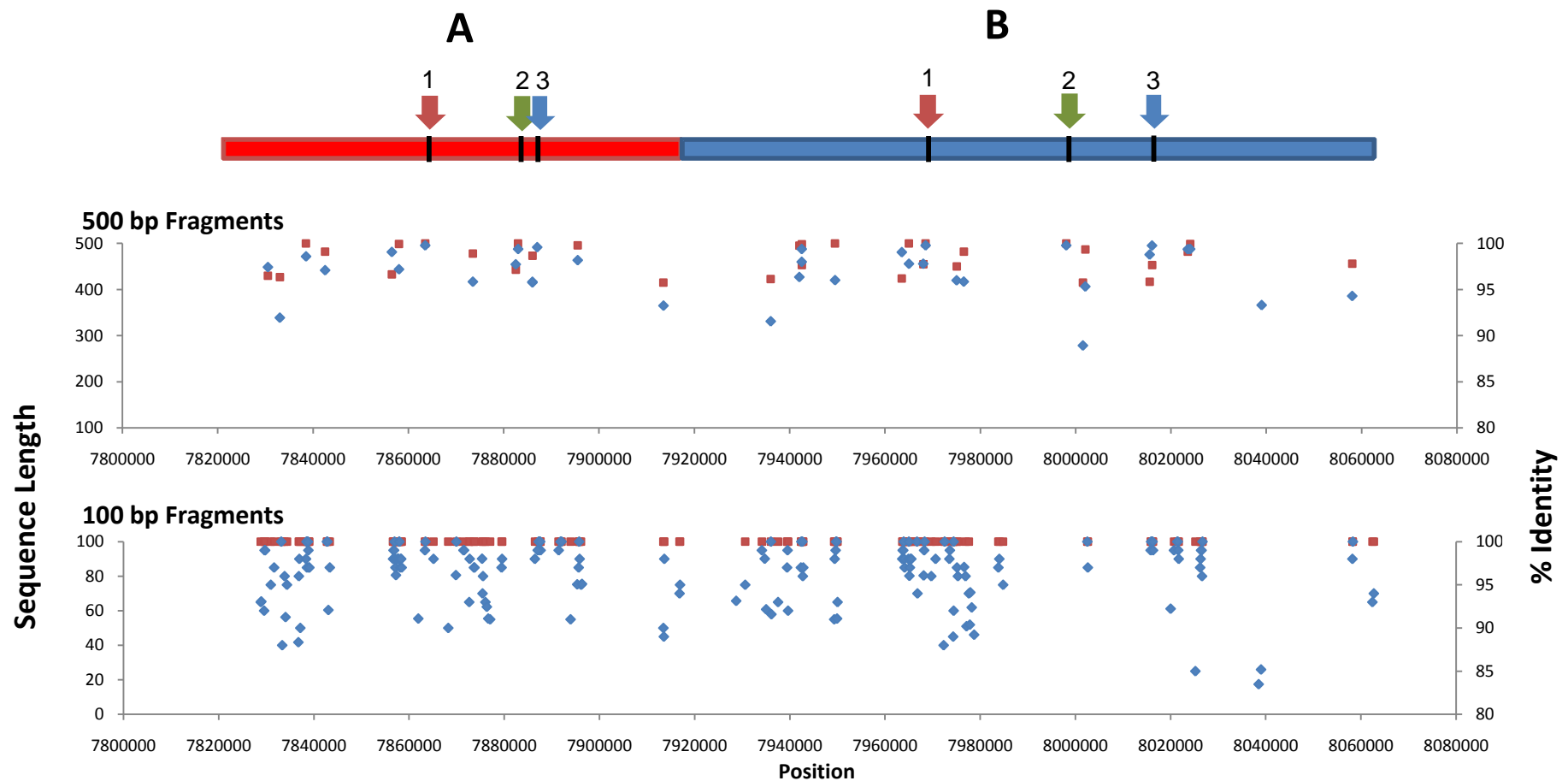


Figure 6.4: Graphical View of Sequence Similarity Between the Two Units of the Tandem Duplication

BLAT alignments were carried out using sequence from the two units of the tandem duplication broken up into different sized fragments. The resulting data is shown graphically for different fragment sizes. The 100 bp fragments have been repeat masked whereas the 500bp fragments have not (see text for details). Red dots indicate the sequence length whereas blue dots represent the % sequence identity. The position of the two units is shown above the graphs, the red bar represents unit A and the blue bar unit B. The position of the three largest regions of identical sequence identified from the dot plots are also shown marked on the tandem duplication for reference.

6.3 Development of a New Series of Assays for CNV

We hypothesise that copy number variation at the 12p13.31 locus occurs as a result of NAHR between blocks of highly similar sequence shared between the two units of the tandem duplication. This would result in the simultaneous deletion and duplication of a segment of DNA which is a combination of sequence from both units of the tandem duplication (See Figure 6.5). As this figure shows, if we imagine a series of theoretical quantitative assays designed at various points across the tandem duplication, each of these acts as a sequence marker. Every unit (be it A, B or the combined unit) will contain one copy of each marker. As the assays move across the region in a variant sample, the relative amount of product amplified from sequence originating from each of the two units will change.

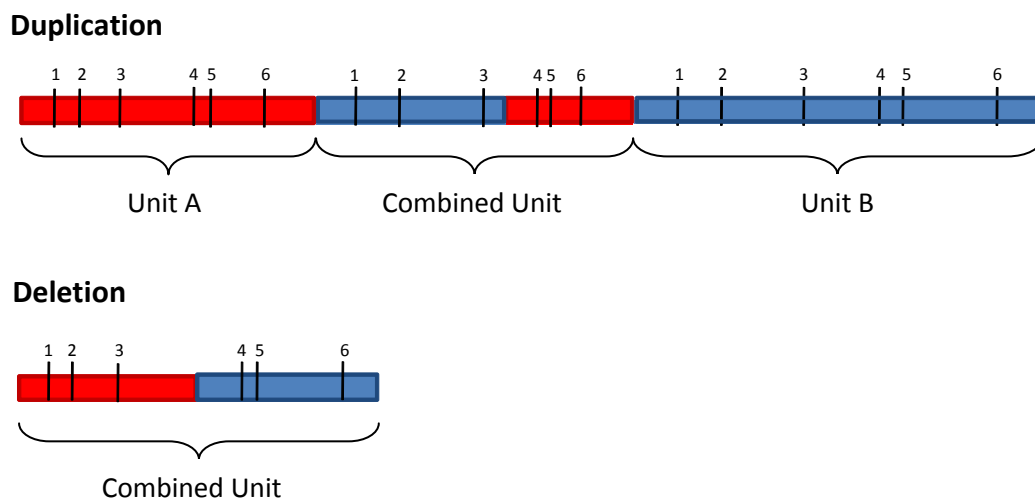


Figure 6.5: Investigating the Location of Recombination Events

NAHR between the two units of the tandem duplication results in deletion or duplication of a region which is a combination of sequence from units A and B, as shown above. Coloured bars represent sequence from units A (red bar) and B (blue bar). A hypothetical series of quantitative assays, shown by numbered lines, act as sequence markers.

For example, in the case of the duplication shown in Figure 6.5, the ratio will change from suggesting an increase in product from unit B (in the case of assays 1-3) to an increase in product from unit A (assays 4-6). The position at which this change occurs can be used to reveal the location of historical recombination events.

It was not possible to test this hypothesis with the A assays, since other than assay A9 these tended not to be reproducible (Described in Chapter 4). Therefore in order to narrow down the location of putative historical recombination points further, a new series of assays was developed to allow the comparison of copy number at different points across the two units.

6.3.1 Assay Design

As described in section 4.4, the A series of assays were designed around repetitive elements which were differentially present or absent in the two units. To complement this approach, a new series of assays was designed. Each of these new assays (termed 'B' assays), uses three primers to simultaneously amplify a region from each of the two units of the tandem duplication (Figure 6.6). For each assay, the single forward primer is located within a block of sequence shared between both units of the duplication, whereas the two reverse primers are positioned within sequences specific to either A or B, and therefore each is unique to one of the two units. As for the previous assays, two products of different sizes are amplified, which can be separated on an agarose gel. Changes in the relative amount of product from each unit of the tandem duplication are used to reveal copy number variation.



Figure 6.6: Design of B Assays

The coloured bars represent sequence from each unit. Primers are indicated by arrows, showing the 5' to 3' direction of amplification. For each assay the common forward primer (shown by a green arrow and labelled 'F') is located within sequence shared between units A and B, whereas the two reverse primers are unique to each unit (Labelled RA for Reverse primer unique to unit A and RB for Reverse primer unique to unit B). Black bars show sequence shared between the two units, whereas gaps between these bars indicate unique sequence.

In total, twelve B assays were designed at a range of intervals across the tandem duplication on chromosome 12p13.31 (Figure 6.7). We have termed these 'B' assays. The exact positioning of each assay was limited to regions where it was possible to design both a unique primer and also a shared primer in close enough proximity in both units for successful amplification to take place. It was also necessary for the two products to be similar in size, but with a large enough size difference to allow separation on an agarose gel. Despite these restrictions, it was possible to design assays spread out across the two units.

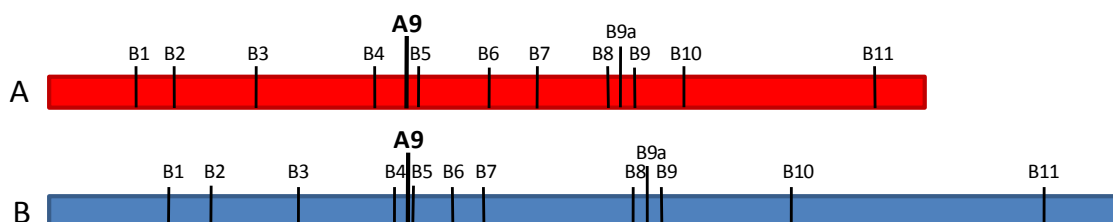


Figure 6.7: Location of B Assays

Twelve sets of B primers were designed across the two units of the tandem duplication. The red bar represents unit A whereas the blue bar shows unit B. The location of each assay is represented by a numbered line in each unit. The location of the A9 assay used for previous genotyping is also highlighted.

6.3.2 Optimisation

The B assays were optimised using a set of 23 DNA samples, some of which had previously been shown to contain variation within the tandem duplication as a result of genotyping with the A9 assay. A series of amplifications were carried out using temperature gradients and a range of different reaction conditions, in order to determine the optimal conditions for each assay (refer to Chapter 2 for further details). These optimisations revealed that assays located within the middle section of each unit (B5 to B10) tended to be the most robust (data not shown).

One of the problems associated with using a set of three primers to simultaneously amplify two different regions is that the efficiency of the two pairs of primers may be different. In a number of cases, for example assay B7, sequence from one unit of the tandem duplication was amplifying at a higher efficiency than the other (Figure 6.8a). This resulted in a considerable difference between the quantities of each product, and made it difficult to reliably compare the signal intensities. In an attempt to overcome this problem, the concentration of the unique primer amplifying with the slowest efficiency was increased. Although after this change the product intensity of the two bands appeared to be more equal on the gel, when the ratio was calculated the results were still variable from experiment to experiment (Figure 6.8 b & c).

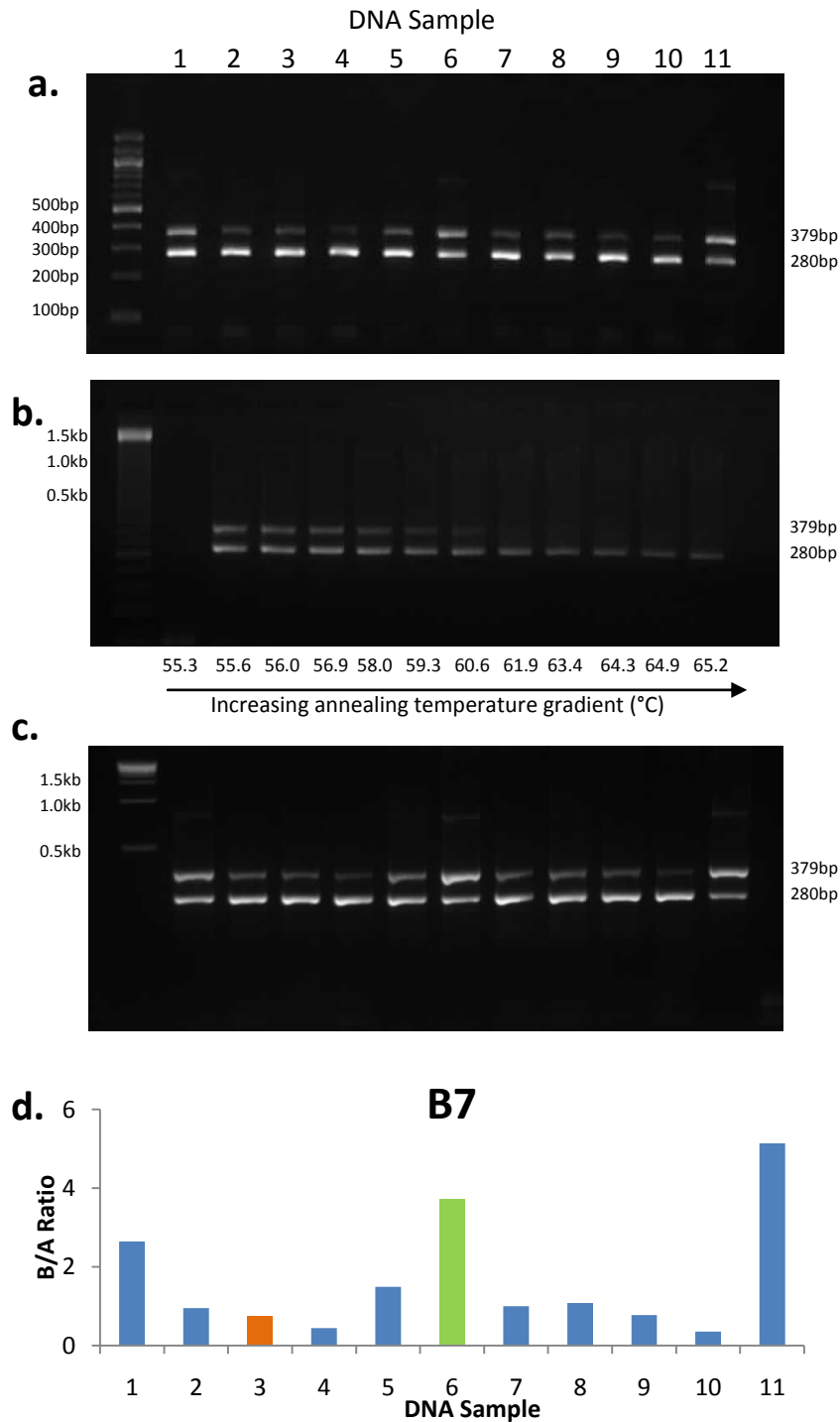


Figure 6.8: Optimisation of an Unequal B Assay

Results of B7 assay optimisation, using the same set of 11 DNA samples in each figure. a.) As this agarose gel shows, products from the two units of the tandem duplication appeared to be amplifying at different rates. b.) Increasing the concentration of the unique reverse primer amplifying at the lowest efficiency resulted in a more equal amplification at lower temperatures. Gel here shows an annealing temperature gradient (55 - 65°C) c.) Subsequent genotyping using the adjusted conditions was not reproducible, shown here an agarose gel and d.) B/A product ratio data from gel c.) shown on a bar chart. Blue bars indicate normal samples, a red bar shows a deletion and green a duplication within this region.

6.3.3 Detection of Putative Recombination Intervals

The most reproducible of the B assays (B5, B6, B8 and B10) were used to search for intervals within which historical recombination events may have taken place. Previous genotyping with the A9 assay had led us to conclude that A9(B) variations are considerably more frequent than those involving A9(A). Therefore our search for sites of recombination initially focussed on samples which had been shown to contain A9(B) duplications or deletions. A set of 33 samples, containing three samples with A9(B) duplications and two with A9(B) deletions, was genotyped with assays A9, B5, B6, B8 and B10 (Figure 6.9).

The results from this genotyping, shown in Figure 6.9, reveal that the product ratio of both A9(B) deletions and A9(B) duplications appears to change between assays B8 and B10. This suggests that there may be a point of recombination located within this interval. Since the distance between assays B8 and B10 in unit B is several kilobase pairs in size, the same set of samples were also genotyped with assay B9 to enable the location of this recombination site to be estimated more precisely (Figure 6.10). Although assay B9 is less reproducible than the others, results suggest that the ratio changes of the variant samples are in the same direction as for assay B8. This locates the point of recombination proximal to this assay, between assays B9 and B10.

In order to investigate these results further, a set of 17 samples, 5 of which contain A9(B) duplications, was genotyped using assays A9, B5, B6, B8, B9 and B10 (Figure 6.11). This series of samples contained a number of individuals from the HapMap collection with European or Yoruba ancestry.

Figure 6.9: Investigating of Sites of Recombination Using B Assays

The four most robust B assays (B5, B6, B8, and B10), as well as assay A9, were used to genotype 33 samples of European ancestry, some of which were known to contain CNV in this region. The B/A product ratio for this set of samples is displayed graphically for each assay. Blue bars represent 'normal' samples, green bars represent A9(B) duplications and red bars represent A9(B) deletions. In all cases, samples A1, B1 and C1 are negative controls. Any other instance where there is no data for a sample, for example B10 in assay A9, is due to failure of the PCR. Black arrows indicate the direction of the product ratio changes in variant samples. This figure continues on the following page.

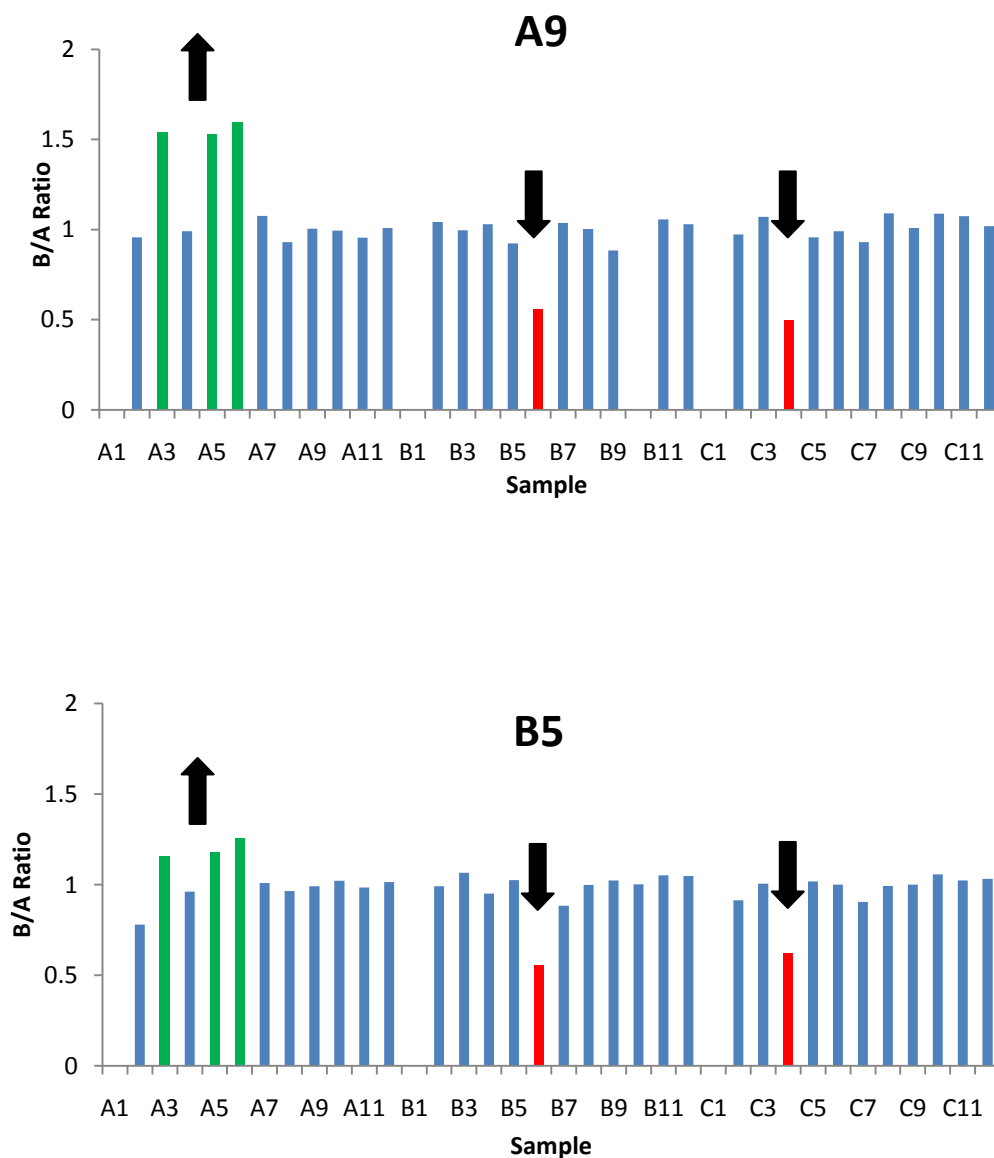
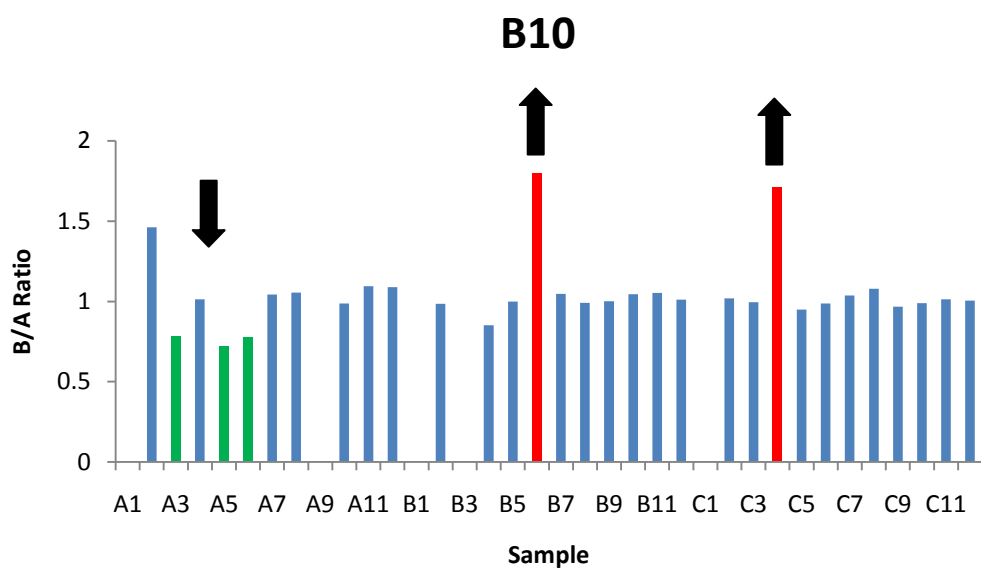
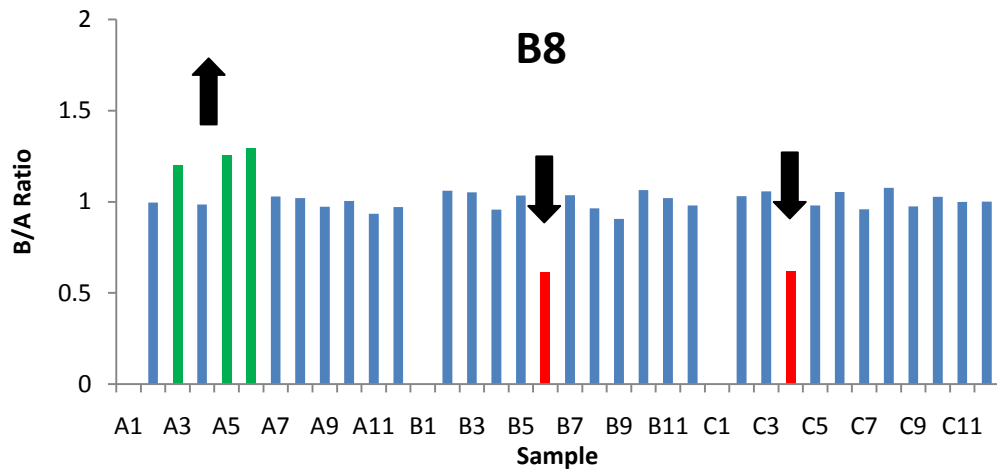
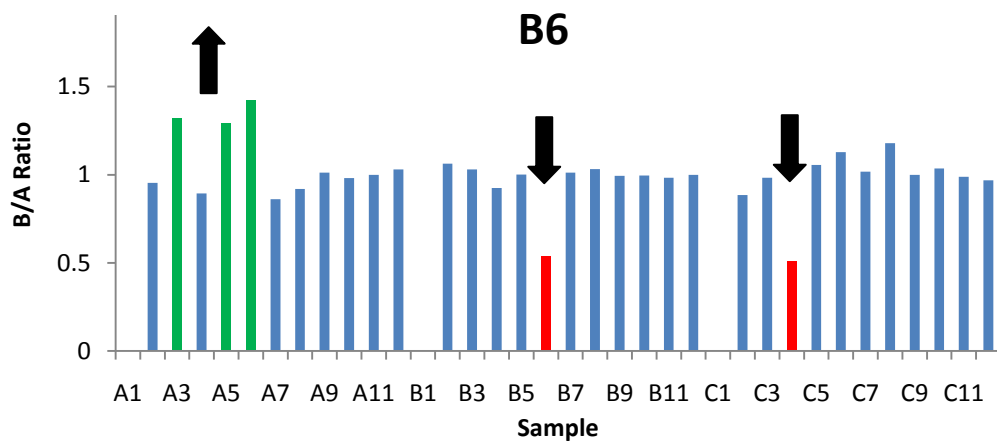


Figure 6.9 continued



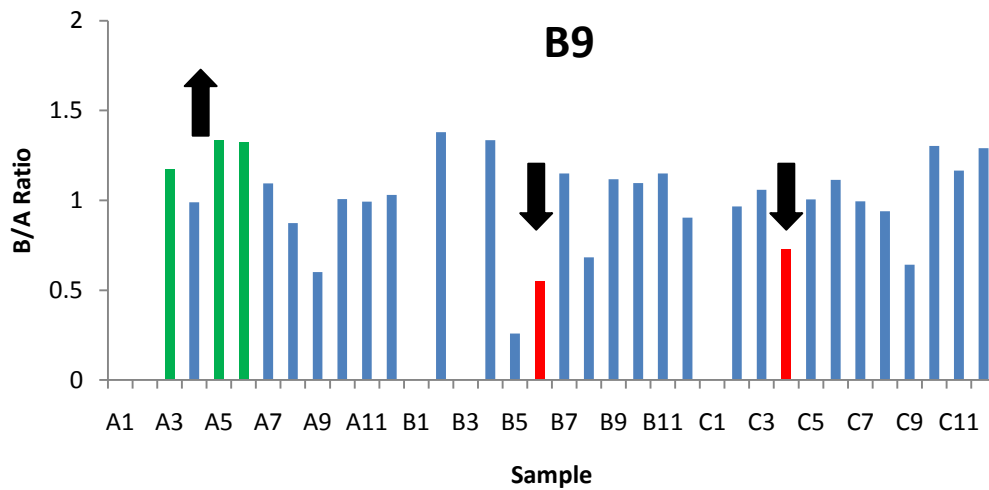


Figure 6.10: Locating Site of Recombination using Assay B9

The set of DNA samples shown in Figure 6.9 were genotyped with assay B9. The B/A product ratio for this set of samples is displayed graphically. Blue bars represent 'normal' samples, green bars represent A9(B) duplications and red bars represent A9(B) deletions. Black arrows indicate the direction of ratios changes in variant samples.

Figure 6.11: Identification of Two Distinct Points of Recombination

17 samples, including 5 A9(B) duplications, were genotyped with assays B5, B6, B8, B9 and B10. Data from assay A9 is shown for comparison. Samples D1-D5 and H1-H7 are CEPH samples, D6 has African ancestry and H8-H11 are from Yoruba individuals (For more details see Appendix C). The B/A product ratio is shown for each assay. Blue bars represent 'normal' samples, whereas green bars represent those with known A9(B) duplications. On each diagram, the black arrows indicate the direction of the product ratio change for variant samples.

Figure 6.11 continued

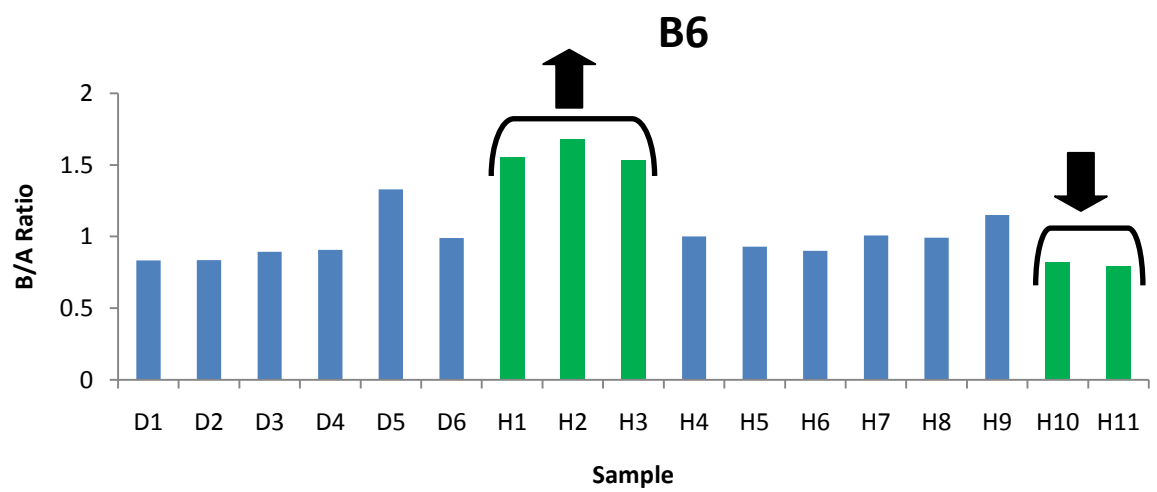
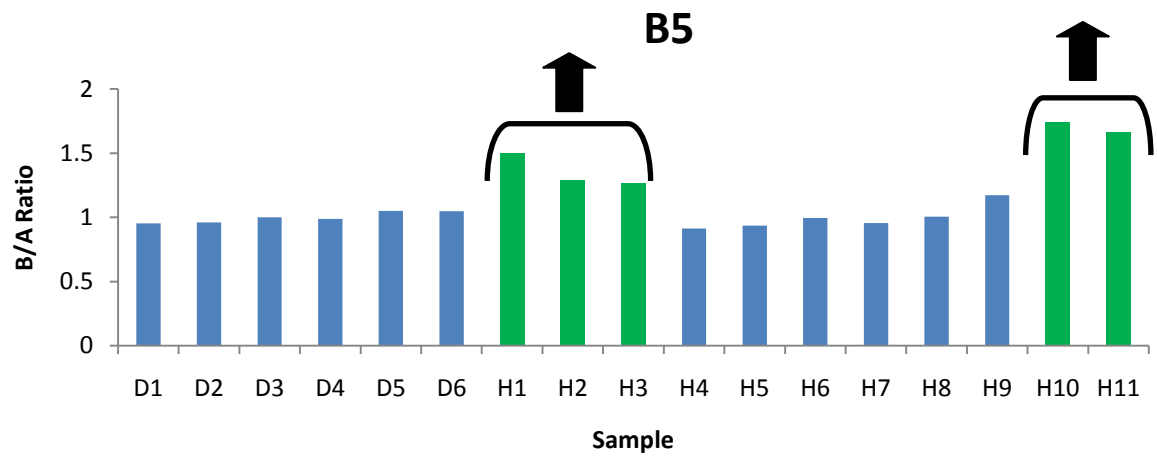
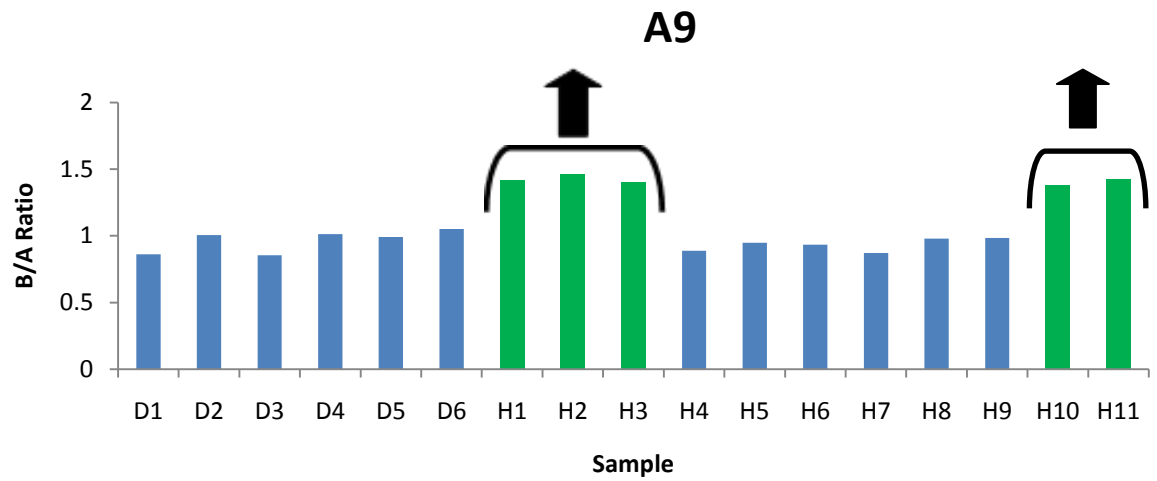


Figure 6.11 continued

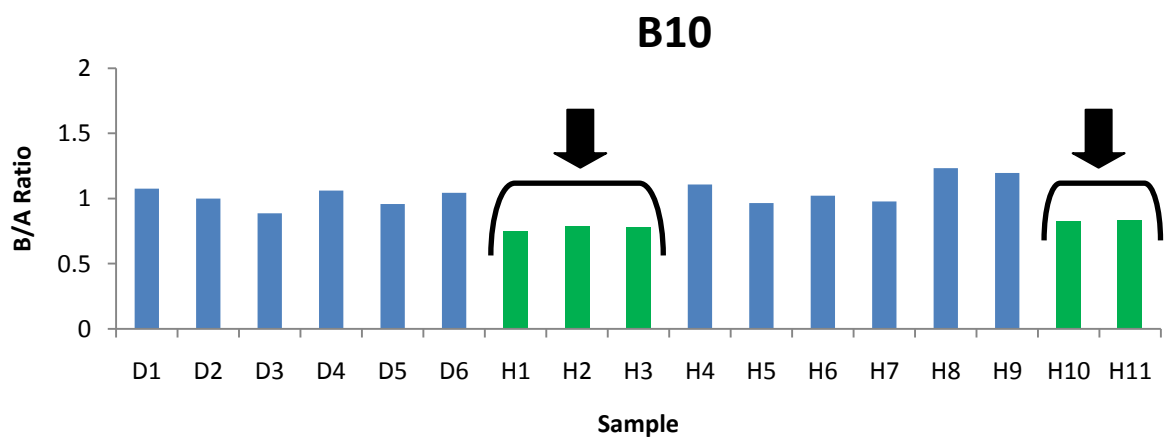
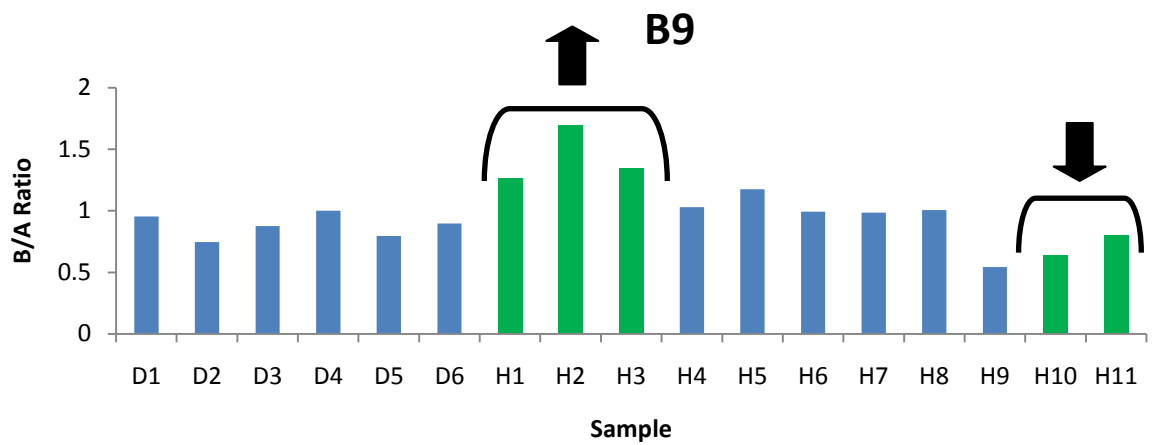
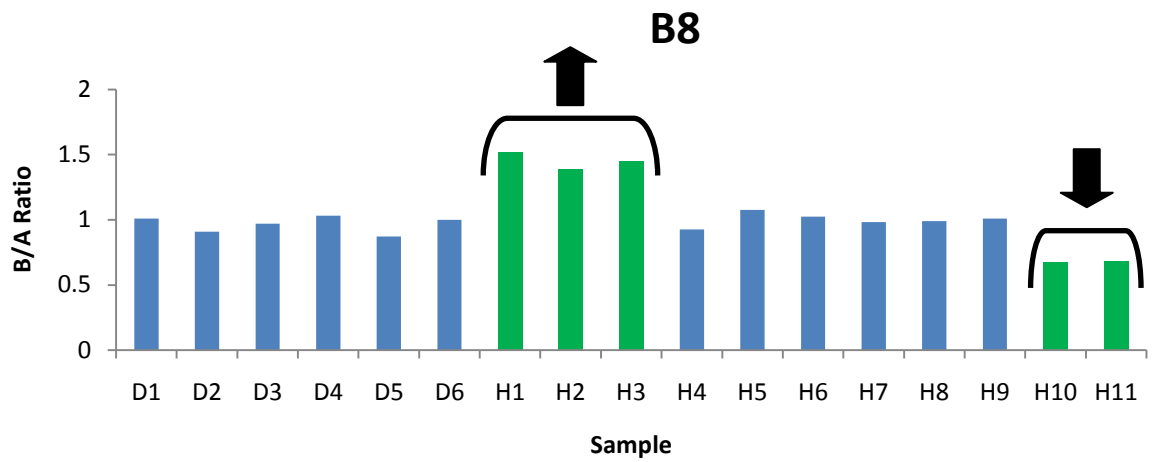


Figure 6.11 shows that there appears to be two different points of change in product ratio; samples H10 and H11 change from an increase in product from unit B to an increase in product from A between assays B5 and B6, whereas H1, H2 and H3 change between assays B9 and B10. This effect would occur as a result of adjacent assays crossing a region within which one or more historical recombination events had taken place, therefore these results suggest the presence of at least two distinct points of historical recombination (Figure 6.12). It should be noted that samples H2 and H3, as well as H10 and H11 are related individuals (parent and child). The position of these putative breakpoints was compared with the results of the *in silico* studies described in section 6.2, which investigated regions of sequence shared between the two units of the tandem duplication (Figure 6.13). Two of the longest stretches of identical sequence are located within these intervals, one between assays B5 and B6 and the other between B9 and B10. These are potentially good candidates for sites of recombination.

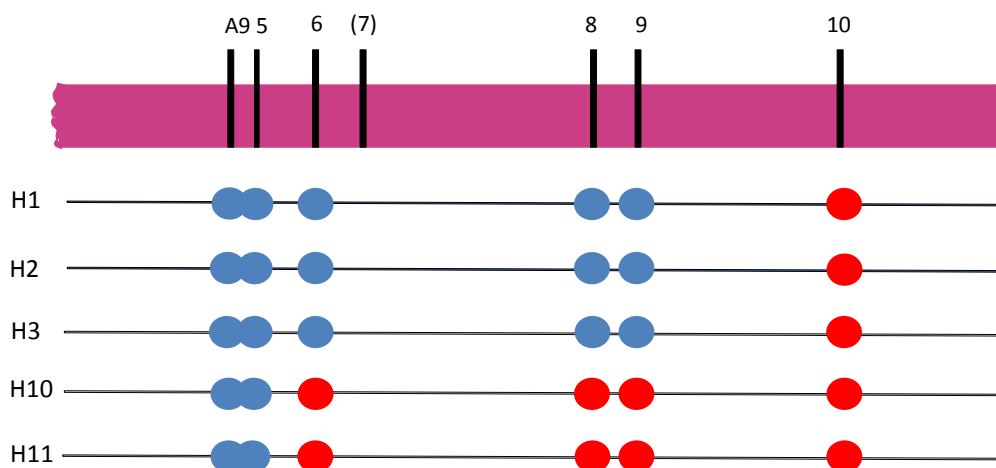


Figure 6.12: B Assays Reveal Two Distinct Sites of Recombination

The purple bar represents a section of a consensus unit from the tandem duplication, showing the relative position of the B assays. The assay results from the five variant samples shown in Figure 6.11 are displayed below; coloured dots indicate whether the B:A product ratio for each assay suggests there is a relative increase in product from unit B (shown as a blue dot) or unit A (red dot). Samples H1, H2 and H3 change from an increase in B to an increase in A between assays B9 and B10, whereas samples H10 and H11 change between assays B5 and B6.

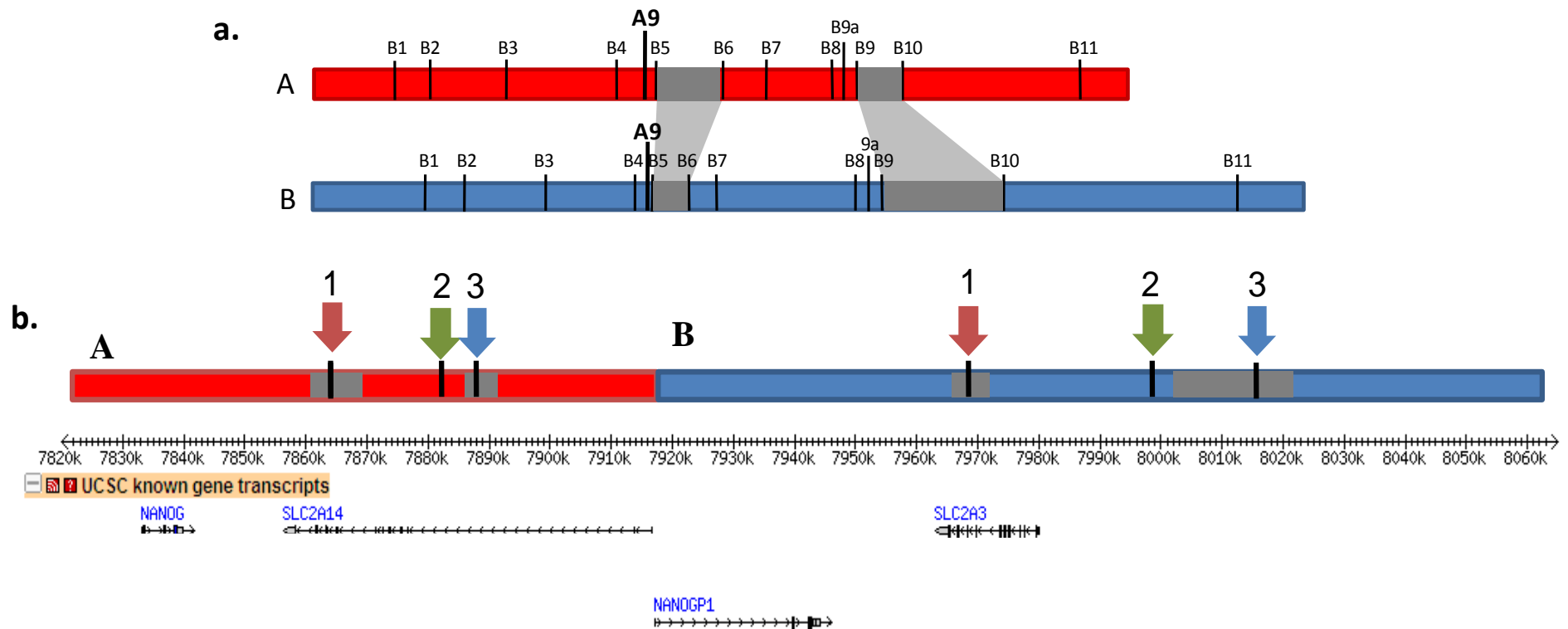


Figure 6.13: Putative Sites of Historical Recombination

B assays enabled the identification of two distinct historical recombination events. a.) The location of these events was narrowed down to two intervals, flanked by assays which show differences in the relative product ratio for the two units of the tandem duplication. Coloured bars represent the two units of the tandem duplication. The numbered lines indicate the position of the B assays in each unit. Grey shading shows two intervals within which there is evidence for historical recombination events. b.) These were compared with the longest stretches of sequence shared between the two units, identified using dot plots. As before coloured bars represent each unit. Coloured arrows show the position of the three largest blocks of sequence shared between the two units. Genes in UCSC genome browser are shown for reference.

To investigate the frequency of historical recombination events within each interval, 352 samples from the 1958 UK control cohort were genotyped with the assays B5, B6, B8, B9 and B10 (Technical Support, Jaya Brakenbury). A number of these samples had previously been shown to contain CNV in this region using the A9 assay. First degree relatives of the CEPH and Yoruba samples were excluded from this analysis. The results of this genotyping reveal that every variant sample within this cohort shows evidence of recombination within the B9/B10 interval. Combining all the B Assay data shows that evidence for recombination within the B5/B6 interval has so far only been detected in two related Yoruba individuals, whereas currently all A9(B) deletions or duplications identified in Caucasian individuals have been shown to contain historical recombination events at B9/B10 (Table 6.1).

Table 6.1: Location of Historical Recombination Events

Type of CNV	Recombination Interval	UK	CEPH	Yoruba
None	-	309	7	2
A9(B) Duplication	B5/B6	0	0	1
	B6/B8	0	0	0
	B8/B9	0	0	0
	B9/B10	28	2	0
A9(B) Deletion	B5/B6	0	0	0
	B6/B8	0	0	0
	B8/B9	0	0	0
	B9/B10	15	0	0
TOTAL:		352	9	3

6.4 Identifying Points of Recombination

Genotyping data from the B assays (described in section 6.3) revealed that at least two distinct recombination events have taken place within the tandem duplication on chromosome 12p13.31. These events have been localised to two stretches of sequence, each several kilobase pairs in size. The first of these intervals is flanked by assays B5 and B6 (the B5/B6 interval) and the second by assays B9 and B10 (the B9/B10 interval). Both intervals contain stretches of sequence previously identified as being amongst the longest regions of sequence identity shared between the two units of the tandem duplication (described in section 6.2). Although these are good candidates for sites of recombination, there are also many stretches of sequence 300-500 bp in size shared between the two units, within which NAHR may also have taken place. Therefore we decided to investigate points of recombination all across the B5/B6 and B9/B10 intervals, rather than just concentrating on the two longest stretches of shared sequence.

6.4.1 Primer Design

Primers were designed within the B5/B6 and B9/B10 intervals by Colin Veal, to amplify across potential recombination breakpoints. A combined approach to primer design was employed, depending on the properties of the target sequence (Figure 6.14). Some pairs of primers were designed so that one primer was specific to A and the other specific to B, so that a product would only be produced if recombination had taken place. Other primer pairs were designed to generate an additional product of a different size from a recombinant DNA sequence compared to a normal sequence.

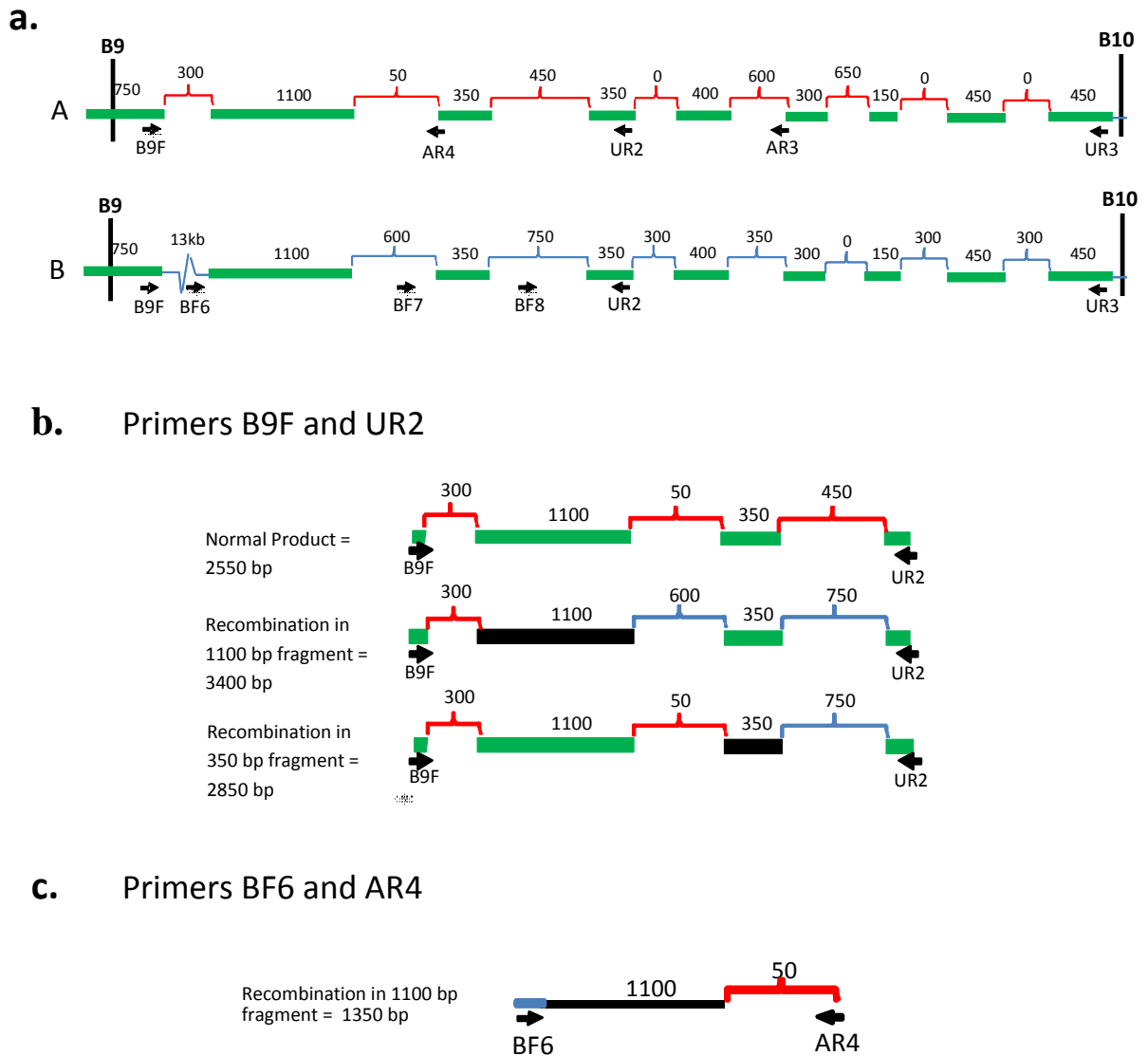


Figure 6.14: Design of Primers to Reveal Recombination Events

Primers were designed to locate recombination breakpoints within the two intervals identified using the B assays. a.) The B9/B10 interval is shown as an example; a similar approach was used to design primers within the B5/B6 interval. Black lines indicate the position of the B assays. Green bars represent regions of shared sequence which are differentially broken up in the two units of the tandem duplication by regions of sequence found only in one unit. These are shown as coloured gaps, red in unit A and blue for unit B. These are not shown to scale, but the size of the region is indicated to the nearest 50 bp. Primers are shown as black arrows. Two complementary approaches were used for primer design, which are as follows; b.) One primer was unique to one unit, and the other common to both units. In a diploid sample within which recombination has occurred, two products are produced. Due to differences in size between the unique regions, the recombinant product is a different length to the normal product. The size of the product indicates the block of shared sequence within which recombination has occurred – on the diagram this is indicated by a black rather than green bar. c.) Primers were designed to be unique to either unit of the duplication, so a product is only produced if recombination has occurred.

In this scenario, the exact size of the recombinant product can be used to reveal the block of sequence within which recombination has occurred.

6.4.2 Primer Optimisation

Primer optimisation was carried out using pairs of primers designed to produce a product in the absence of recombination in this region (Technical support, Kelly Rooke) (Tables 6.2 & 6.3). For more details on reaction conditions see Chapter 2.

Table 6.2: Control Primers for B5/B6 Interval

Forward Primers	Reverse Primers			
	AR1	AR2	UR	BR
	AF1	TD		TD
	AF2	TD(2)		
	BF2		FS	
	BF3		FS	
	BF4		10x	
	BF5		10x	

Table 6.3: Control Primers for B9/B10 Interval

Forward Primers	Reverse Primers			
	AR3	AR3B	UR2	UR3
	BF6		FS	
	BF7		FS	FS
	BF8		10x	FS
	B9F	10x	11x	

Grey shaded boxes indicate primers that were not designed to work together as controls. Green shaded boxes show primer pairs which were successfully optimised. The buffer or amplification system which, when used, resulted in the amplification of a single clean band is also shown; 10x, 10x Buffer; 11x, 11.1x Buffer; FS, FastStart System (Roche); TD, Touchdown PCR. For primer pair AF2/AR2, the (2) indicates that two products were amplified under the given conditions, since these primers unexpectedly annealed to both units of the tandem duplication it was not possible to amplify one band only.

6.4.3 Investigating Recombination within the B9/B10 Interval

Once we had established that the control primers were amplifying the correct sequence successfully, we moved on to investigate putative historical recombination events within the B9/B10 interval using samples previously shown to contain A9(B) duplications or deletions. Pairs of primers designed to work together (as described previously) were used to amplify across blocks of sequence which are shared between the two units of the tandem duplication, and are therefore potential sites of recombination.

6.4.3.1 *A9(B) Duplication*

Primers BF6 and UR2 were used to investigate recombination events with the B9/B10 interval in samples which had previously been shown to contain a duplication of the A9(B) region (Figure 6.15). A normal sample will produce a single 3000 bp product. All samples used in these studies are diploid, and we have made the assumption that each sample contains a maximum of one variant. Therefore, A9(B) duplications should produce two products, one of 3 kb amplified from the normal tandem duplication, and a second, shorter product from the duplicated sequence.

Figure 6.15 shows that, as expected, two products are amplified from A9(B) duplications. The second PCR product, which occurs only in variant samples, is 2.15 kb in size. This suggests that a historical recombination event has taken place within a 1100 bp block of sequence, which has a sequence identity of 99% between the two units of the tandem duplication. In order to confirm this, two additional primers (UR4 and UR5) were designed at the distal end of the 350 bp block of shared sequence.

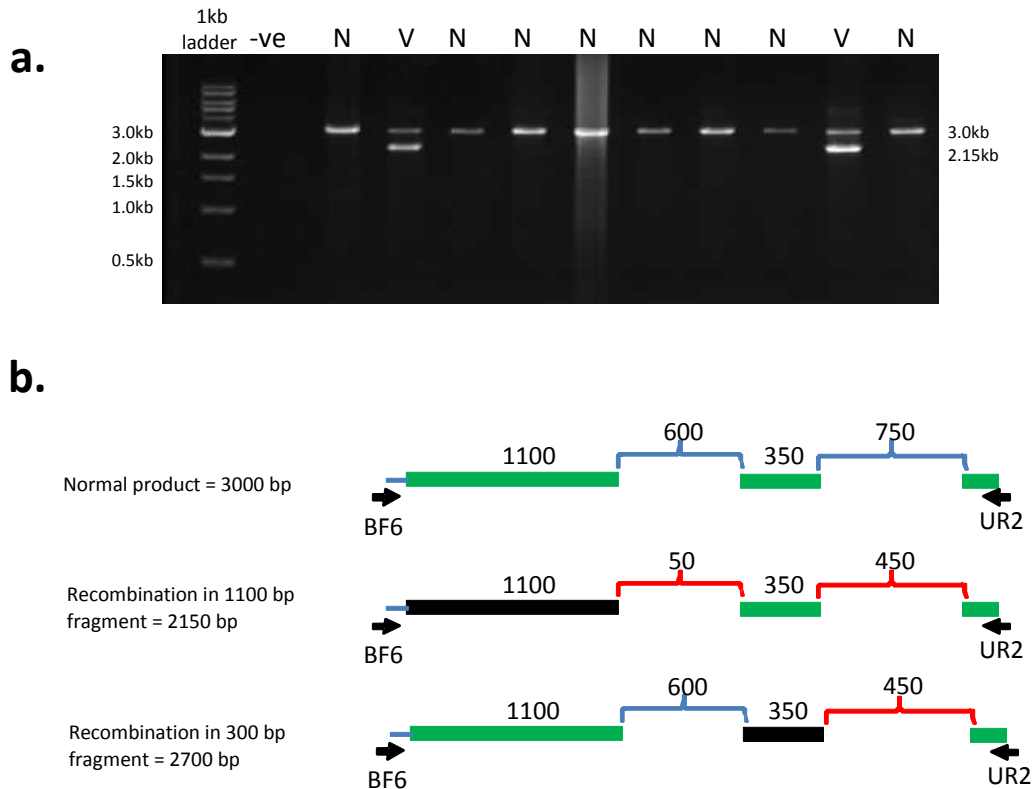


Figure 6.15: Recombination Within the B9/B10 Interval Resulting in a Duplication

a.) The primer pair BF6/UR2 was used to amplify sequence from 10 independent samples. In most cases a single band is produced, but in two variant samples two products are amplified, shown here on an agarose gel. N marks a normal sample whereas V shows a sample containing a duplication. b.) Possible products from a BF6/UR2 amplification are shown. Green bars represent sequence shared between the two units of the tandem duplication, gaps of different sizes between these shared regions are marked in red for unit A and blue for unit B, with numbers indicating the length of each region. Black bars mark the block of sequence within which recombination has occurred in each scenario. Primers used to detect these events are shown as arrows pointing in the direction of amplification.

Any recombination events detected by amplifications carried out using BF6 and either UR4 or UR5 can only have occurred within the 1100 bp region, since this is the only block of shared sequence covered by this amplification. UR4 and UR5 were each individually paired with BF6 and used to amplify sequence in normal and recombinant samples (Figures 6.16 and 6.17).

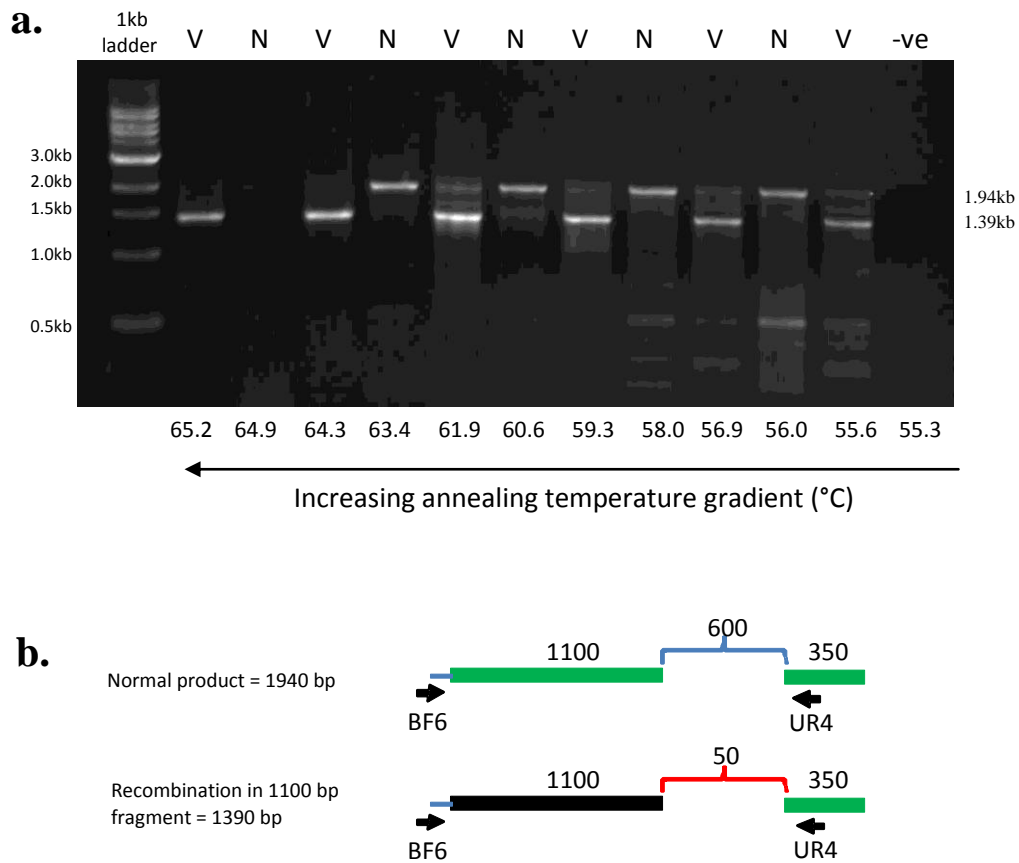


Figure 6.16: Identification of Recombination Within a 1.1 kb Region (1)

The primer pair BF6/UR4 was used to confirm that a historical recombination event had taken place within the 1100bp block of sequence shared between the two units of the tandem duplication. a.) The agarose gel shows amplification of two alternating samples, one normal and one known to contain an A9(B) duplication (labelled 'V' for 'variant'). An annealing temperature gradient is used in this example. The normal DNA sample produces a single product, whereas the variant produces an additional band. b.) Illustration of the recombination event which has occurred to result in a product of approximately 1390 bp in size. As before, green bars represent sequence shared between the two units, gaps between these shared regions are marked in red for unit A and blue for unit B and numbers represent the size of each region. Black bars mark the putative location of recombination events in each scenario. The primers used to detect these events are shown as arrows pointing in the direction of amplification.

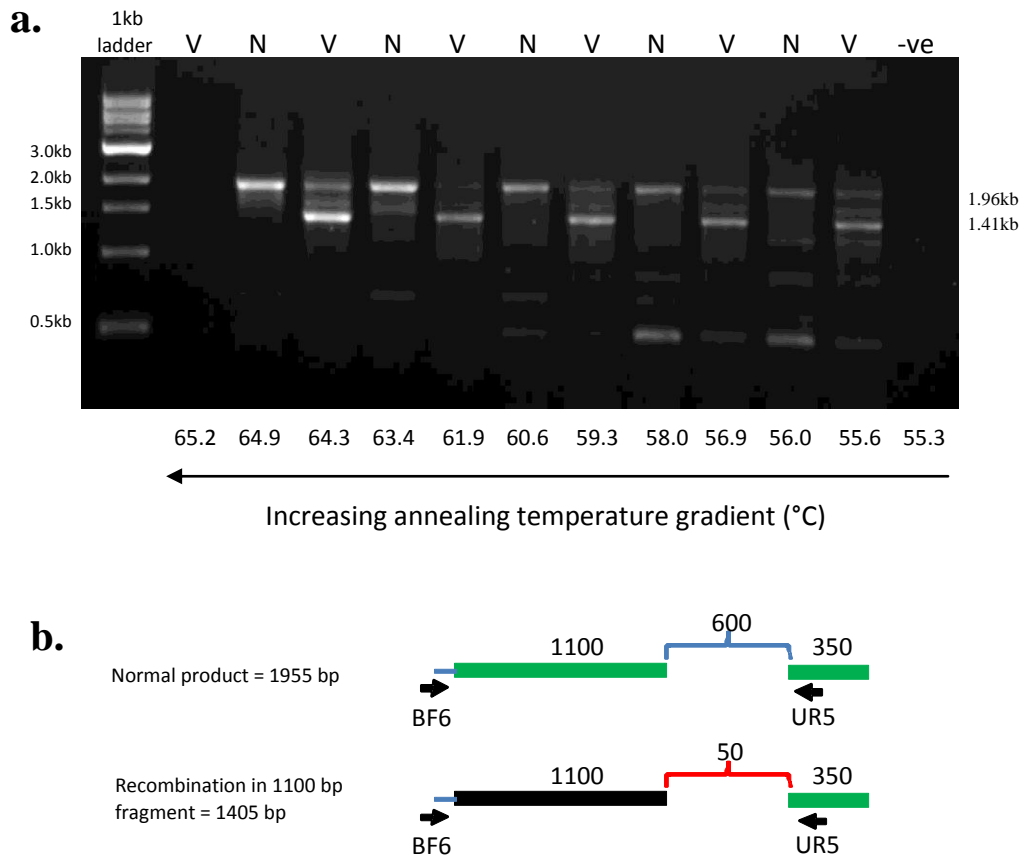


Figure 6.17: Identification of Recombination Within a 1.1 kb Region (2)

The primer pair BF6/UR5 was used to confirm that a historical recombination event had taken place within the 1.1 kb block of sequence shared between the two units of the tandem duplication. a.) The agarose gel shows amplification of two alternating samples, one normal and one known to contain an A9(B) duplication (labelled 'V' for 'variant'). An annealing temperature gradient is used in this example. Normal samples produce a single band, whereas recombinant samples produce an additional band. b.) Illustration of the recombination event which has occurred to result in a product of approximately 1405 bp in size. Green bars represent sequence shared between the two units, gaps of different sizes between these shared regions are marked in red for unit A and blue for unit B and numbers represent the size of each region. Black bars mark the putative location of recombination events in each scenario. The primers used to detect these events are shown as arrows pointing in the direction of amplification.

In both cases, two bands were produced from recombinant samples, one corresponding to the normal product and one recombinant. The band representing the normal product is faint in the variant samples; this is likely to be a result of preferential amplification of the recombinant product.

A second approach used to design primers for breakpoint mapping was to locate one primer of a pair within unit A and the other within unit B. In this design a product is only produced if recombination has occurred, bringing the two primers in close proximity to each other. This is the more direct method of the two, as the presence of a product must indicate a recombinant and is unlikely to be a false positive. Using primers BF6/AR4 it was possible to confirm the presence of a recombination event within the 1.1 kb block of shared sequence which shows a 99% sequence identity between the two units of the tandem duplication (Figure 6.18).

6.4.3.2 *A9(B) Deletion*

Primers B9F and UR2 were used to investigate recombination events within two Caucasian samples which had previously been shown to contain A9(B) deletions (Figure 6.19). In normal samples, as expected, a single product 2.55 kb in size was produced. However, in all of the suspected deletion samples tested, a second band of 3.4 kb was also present. The presence of a product of this size indicates that recombination has taken place within the 1100 bp stretch of sequence shared between the two units of the tandem duplication. We have previously shown that the same block of sequence is also the location of recombination events in a number of samples containing a duplication of sequence within this region.

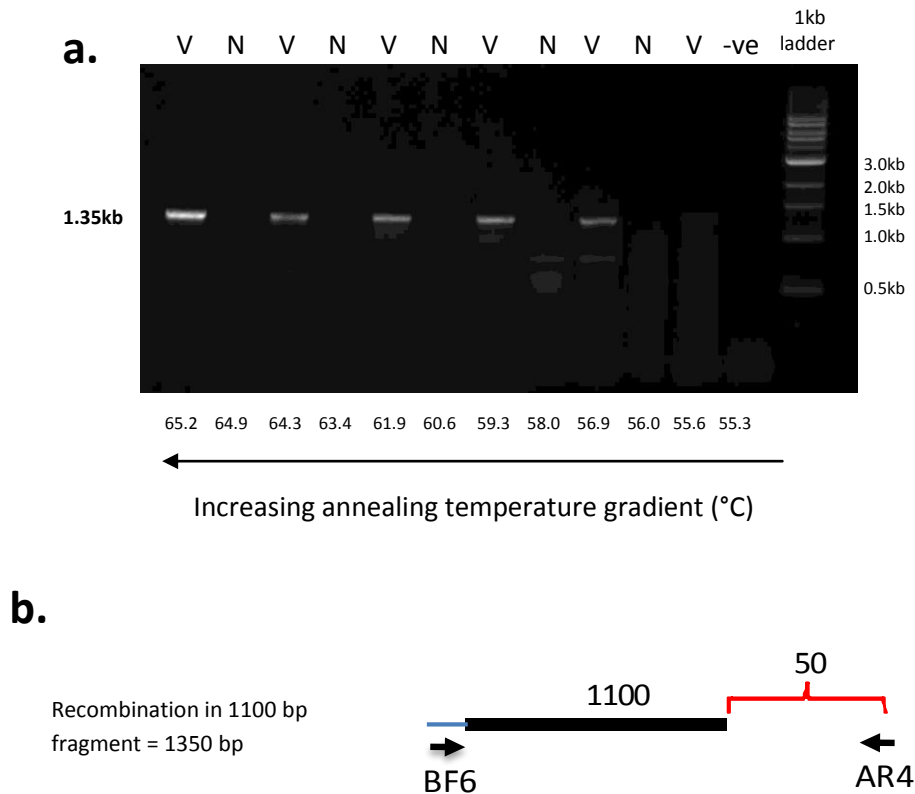


Figure 6.18: Direct PCR to Identify an A9(B) Duplication Event

A PCR was designed to product a product only in samples containing an A9(B) duplication. a.) The gel shows alternating normal and variant samples. A product is only produced by variant samples. b.) As one of the primers is located in the B unit of the tandem duplication, whereas the other is located within A, a historical recombination event must have occurred in the sample to bring these two primers close enough together for amplification to occur. A black bar represents the region of sequence shared between the two units, within which recombination has occurred. The red region shows a region present in unit A only. Numbers represent the sizes of each region. The primers used to detect this event are shown as arrows pointing in the direction of amplification.

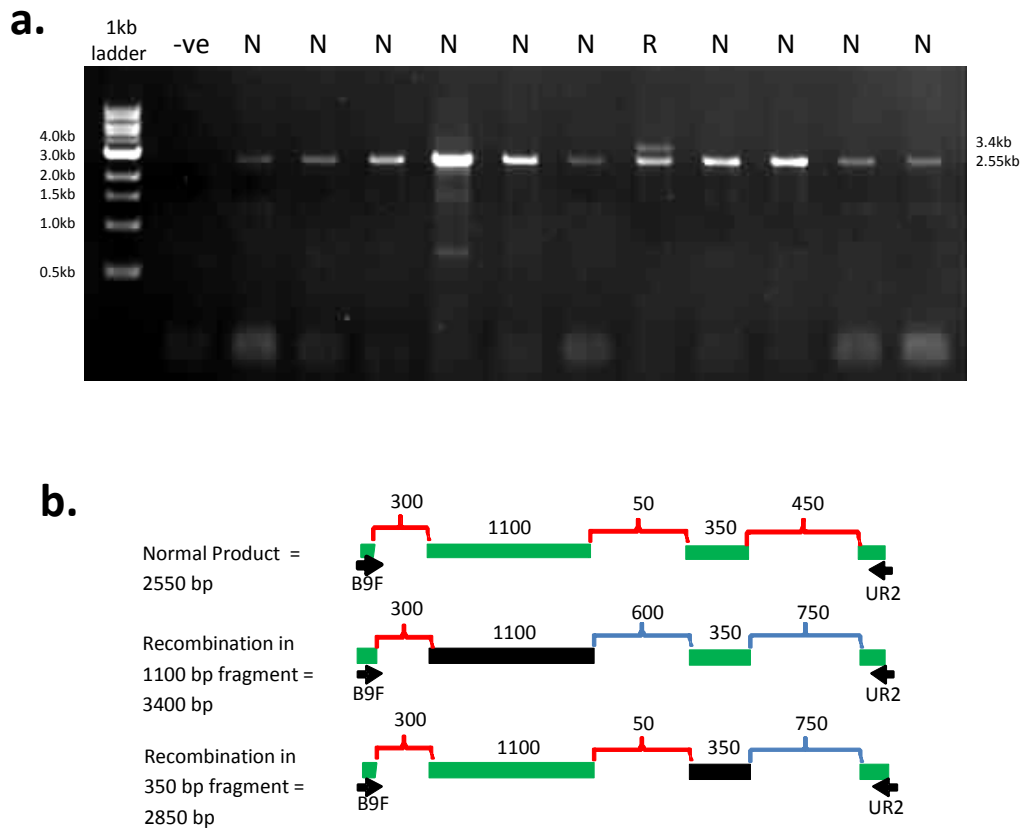


Figure 6.19: Recombination at B9/B10 resulting in a Deletion

Primers B9F/UR2 were used to detect a recombination event within B9/B10 interval resulting in a deletion of sequence. a.) As shown here on an agarose gel, amplification of normal samples results in just one product. However, in the single recombinant sample shown here, two products are produced. N indicates normal samples whereas R marks the recombinant sample. b.) Potential products of a B9F/UR2 amplification. Green bars represent sequence shared between the two units, whereas gaps of different sizes between these shared regions are marked in red for unit A and blue for unit B. Numbers represent the sizes of each region. Black bars mark the putative location of recombination events in each scenario. The primers used to detect these events are shown as arrows pointing in the direction of amplification.

6.4.4 Investigating Recombination within the B5/B6 Interval

The same approach as described for the B9/B10 interval was used to investigate recombination events seen only in Yoruba samples within a region flanked by the B5/B6 assays (Figure 6.20). Amplification across the smaller regions of sequence

shared between the two units of the tandem duplication, for example using primer pairs BF3/UR, BF4/UR and BF5/UR, failed to detect the presence of any recombination events. However, we have not yet been able to achieve successful amplification across the 1500 bp block of sequence, which is the largest region of shared sequence within this interval. There are also a number of smaller blocks of sequence identity which have not yet been studied. Further investigations are therefore required in order to locate recombination events within this sequence interval.

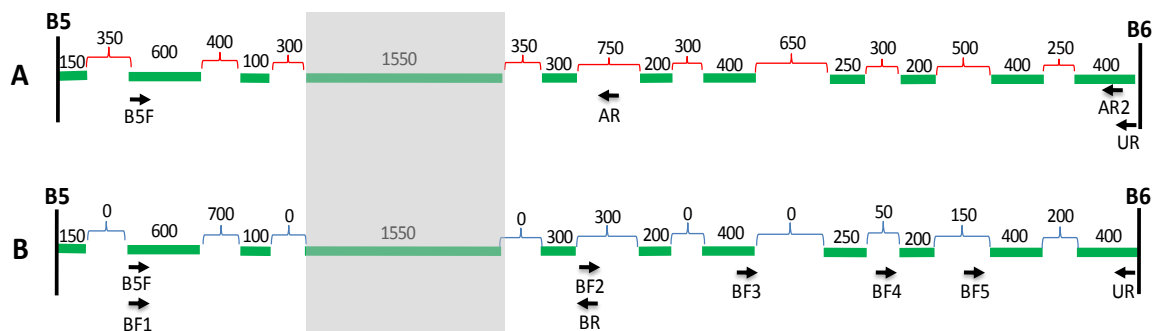


Figure 6.20: Structure of Sequence Similarity in the B5/B6 Interval

Sequence similarity within the B5/B6 interval is shown. The position of the B assays are marked by black lines at the two ends of the region. Green bars represent regions of shared sequence which are differentially broken up in the two units of the tandem duplication by regions of sequence shown as red gaps (in unit A) or blue gaps (in unit B). Numbers indicate the size the region to the nearest 50bp. Black lines indicate the position of the B assays. Primers are shown as black arrows.

6.4.5 Validation of Genotyping Results

To determine the frequency of historical recombination events within the 1100 bp block of shared sequence in the B9/B10 interval, and also to validate the accuracy of our previous genotyping, a subset of samples from the 1958 UK control cohort which had been used in the disease association studies (described in Chapter 5) were investigated

using a number of the recombination PCRs described above. Amplifications with primer pairs BF6/UR2 and BF6/AR4 were used to investigate duplications, whereas B9F/UR2 was used to investigate deletions. The results were compared with those from the A9 assay (Table 6.4). It can be seen that all the A9(B) deletion and duplications identified previously were confirmed using the three recombination PCRs. Every recombination event we have detected within this set of samples has taken place within the 1100 bp stretch of sequence shared between the two units of the tandem duplication.

Table 6.4: Confirmation of Variants in UK Samples using Recombination PCRs

	A9 Assay	Recombination within 1100bp region
Normal	309	0
A9(B) Duplication	29	29
A9(B) Deletion	14	14
TOTAL:	352	43

So far the B9/B10 interval has been the point of recombination for all Caucasian samples with either a duplication or deletion of A9(B). In contrast, one A9(B) duplication event detected in two related Yoruba individuals shows evidence for a historical recombination event within the region flanked by assays B5 and B6. In order to investigate the frequency of historical recombination events within the two intervals within another population, four Chinese/Japanese samples identified from the HapMap plates as containing A9(B) duplications were also genotyped using the BF6/UR2 PCR. Results showed that in all four samples, recombination has taken place within the 1100

bp block of shared sequence, as for the Caucasian samples (B9/B10). Since no A9(B) deletions were identified in the small sample set available for this population, it was not possible to investigate whether the same is true for these variants. Due to limited sample availability it was not possible to test a larger set of samples; however this result suggests that the historical recombination event we have identified within the B9/B10 interval is not restricted to individuals of Northern European origin. Table 6.5 shows the frequency of each event in the four populations studied. At present, the Yoruba A9(B) duplication remains the only one to have shown evidence of recombination within the B5/B6 interval.

Table 6.5: Population Differences in the Location of Recombination Events

Variant	UK	CEPH	Yoruba	Japanese/Chinese
None	309	7	2	0
B5/B6 Duplication	0	0	1	0
B5/B6 Deletion	0	0	0	0
B9/B10 Duplication	29	2	0	4
B9/B10 Deletion	14	0	0	0
TOTAL:	352	9	3	4

6.5 Summary and Discussion

Work described in this chapter involved the investigation of extant allele structures for copy number variations within the tandem duplication at chromosome 12p13.31.

Two intervals were identified within which distinct historical recombination events have taken place (termed B5/B6 and B9/B10). Within the B9/B10 interval a site of recombination was detected within a stretch of sequence 1.1 kb in size, which is shared between the two units of the tandem duplication. Both deletion and duplication events within this sequence were detected. Results for variation within this interval appear to be consistent with a simple model of NAHR. However, it has not yet been possible to identify breakpoints in the B5/B6 interval using the same approach and therefore at present we cannot firmly establish whether variation within the B5/B6 interval is also consistent with the simple NAHR model, or if it is a result of more complex structural rearrangements.

In this chapter, the design of PCRs which are able to provide direct confirmation of recombination events (i.e. a variant is either present or absent) have been described. These are a more reliable method of detecting copy number variation than the previous assays, which relied on the comparison of product ratios and were therefore subject to inaccuracy. The advantage of the PCRs which produce a product from both normal and variant units is that these assays contain an inbuilt positive control; since one band should be produced even in a non-variant sample, this can be used to check whether the PCR has failed. A limitation of the second approach, in which a product is only produced as a result of recombination, is that if no band is detected this could either be due to the fact that there is no recombination, or that the PCR has failed. There is no way of differentiating between the two scenarios.

We have detected evidence of historical recombination within the B9/B10 interval in a number of populations, including Caucasian and Chinese/Japanese. So far every recombination event at this position appears to have taken place within the same region of sequence, a 1.1 kb stretch which has a 99% sequence identity between the two units of the tandem duplication. It is not possible to conclude whether this represents a single ancestral recombination event, or a number of distinct events which have spread across the population. Future work could involve sequencing across the 1.1 kb region in recombinant samples, in order to identify the precise location of the recombination breakpoint(s). However, due to the high level of sequence identity which exists between the two units of the tandem duplication, it may be difficult to identify the exact points of recombination.

The B9/B10 interval has so far been identified as the point of recombination in the majority of variant samples studied. However, we have also identified evidence for ancestral recombination within the B5/B6 interval, which has so far only been detected in two related Yoruba individuals. The presence of another point of historical recombination in this population is perhaps not surprising, as studies have shown that there is a more diverse range of CNVs in samples of African origin (Armengol *et al*, 2009). It is possible that this recombination event did not escape Africa, and is therefore restricted to individuals of this population. Further studies involving a much larger number of samples could be used to investigate this further.

Given the relatively small sample size of all the populations genotyped for this research, it is possible that recombination events may also have occurred at locations other than the two identified here. In the future it would be interesting to look at a larger number of individuals from a wider variety of populations, to identify whether other points of

recombination do exist, and also to determine to what extent the frequency of recombination at different loci varies between populations.

Although variations involving copy number changes at locus A9(B) are the most common form of CNV at 12p13.31, the A9 assay was also able to detect copy number changes of A9(A). Since we have been unable to design reproducible assays towards the distal end assay A9, it has yet not been possible to identify recombination events which would lead to variation of A9(A). Genotyping data for A9(A) deletions and duplications tends to be more inconsistent than that for A9(B) recombinants, and results from association studies described in Chapter 5 suggest that A9(A) changes occur at a much lower frequency. We do not currently have any conclusive evidence that variation of A9(A) fits with the simple model of NAHR which we have proposed for A9(B) variants; it may be that A9(A) variants are a result of more complex structural rearrangements. In the future, the design of additional assays and further optimisation of primers may allow us to validate the results of the A9 assay for A9(A) variants, and also investigate recombination events which result in copy number variation of A9(A).

It is possible to consider the effect of recombination events within the B5/B6 and B9/B10 intervals (Figure 6.21). Table 6.6 shows how each of the four genes within the tandem duplication would be affected by each of these events. Recombination within the B5/B6 interval results in the fusion of sequence from the two glucose transporter genes, *SLC2A3* and *SLC2A14*, which may lead to the production of hybrid mRNA transcripts. Since these genes share a high degree of sequence similarity, it would be interesting to study whether such fusion products would be functional. This could be investigated using both the mRNA, to determine whether transcription would take place, as well as the protein.

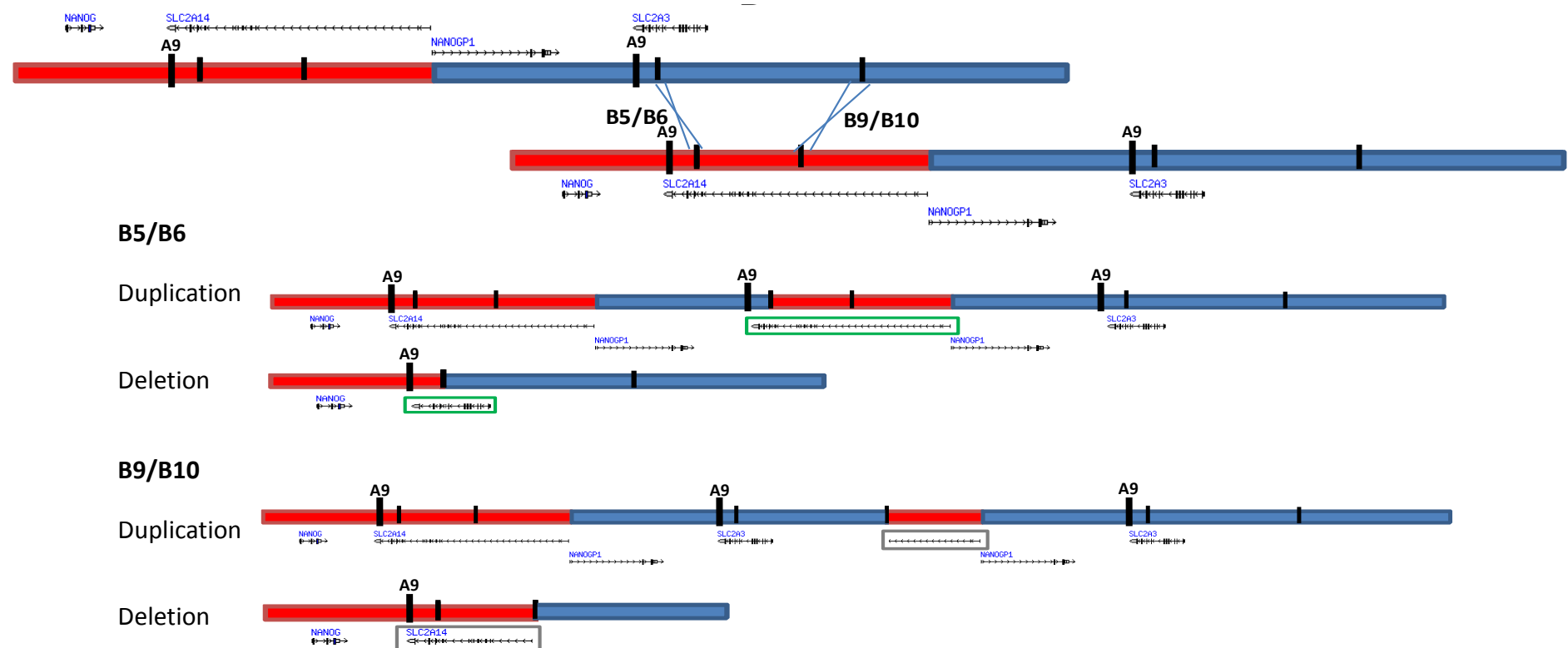


Figure 6.21: NAHR in the B Unit of the Tandem Duplication

Representation of the likely consequences of recombination events at B5/B6 and B9/B10. The coloured bars represent the two units of the tandem duplication, with the genes shown below. Black lines represent the position of the two intervals within which historical recombination events have been identified. The position of the A9 assay is also marked with labelled black lines. The products produced from recombination at the two loci are shown. Green boxes represent possible gene fusion products which may be produced as a result of fragments of the two glucose transporter genes, *SLC2A3* and *SLC2A14*, being brought together. Grey boxes indicate that part of the gene has been lost.

Table 6.6: Effect of Recombination Events within the B5/B6 and B9/B10 intervals on Genes within the Tandem Duplication (in the recombinant chromosome only)

	<i>NANOG</i>	<i>SLC2A14</i>	<i>NANOGP1</i>	<i>SLC2A3</i>
B5/B6 Duplication	1	1 + fusion with <i>SLC2A3</i>	2	1 + fusion with <i>SLC2A14</i>
B5/B6 Deletion	1	Fusion with <i>SLC2A3</i>	0	Fusion with <i>SLC2A3</i>
B9/B10 Duplication	1	1 + 3' fragment	2	2
B9/B10 Deletion	1	5' fragment	0	0

B9/B10 duplications and deletions result in either gain or loss of *SLC2A3*, as well as loss of a segment of *SLC2A14*. The potential physiological effects of a deletion of *SLC2A3* will be discussed later (section 7.3.2). The effect of the fragmentation of *SLC2A14* is unknown, but may warrant further investigation.

While the recombination events appear to mostly affect the glucose transporter genes, it should be noted that in all cases there is either a loss or gain of *NANOGP1*. Although this is a pseudogene it is known to be transcribed, and therefore the changes in copy number may have consequences. Although here we have concentrated on the genes within the tandem duplication, it is possible that these structural changes may also confer effects on other genes in the surrounding regions.

This chapter has described studies of recombination events which have led to the formation of copy number variation within the 12p13.31 tandem duplication. Two distinct loci within which there is evidence for historical recombination events have been detected, and within one of these it has been possible to pin the location of this

event down, to a short region of sequence. Future work should focus on the identification of other recombination events, for example within the B5/B6 interval. Having identified the location of recombination events it is possible to hypothesise how these might affect expression of the genes within this region. Understanding the effect that CNV within the tandem duplication may have on gene and protein levels may provide important insights into mechanisms by which variation at this locus is involved in complex disease, and therefore such studies should form a major part of any follow-on research from this project.

Chapter 7

Discussion

Research described in this thesis has focused on the characterisation of a novel tandem duplication on chromosome 12p13.31, within which structural variation has been detected. Specifically, the aims of my project were a) to develop an optimised protocol for oligo-arrayCGH on an in-house microarray platform, b) to use this technique, as well as other methods, to investigate a region of putative copy number variation on chromosome 12p13.31, and c) to study association of this structural variant with rheumatoid arthritis and other complex disorders.

7.1 Development of oaCGH on the Geniom Platform

As knowledge of the extent to which copy number variation is present in the genome increases, there is a growing need for techniques to study this form of variation. One method currently employed for detecting CNVs is oligo-arrayCGH (oaCGH) (Carvalho *et al*, 2004). The paucity of reliable in-house platforms for this technique is a limiting factor for investigating structural variation on a large scale. The focus of my research was to develop an optimised protocol for oaCGH on an in-house platform, specifically the Geniom microarray platform. This system has been developed for a number of applications, including expression studies and microRNA profiling, but had not previously been used for oaCGH.

By modifying various reaction conditions, we were able to achieve a reasonably optimised protocol for oaCGH on the Geniom platform, with which it was possible to detect regions of CNV. During the optimisation process, a great deal was learnt about the complexities of this technique, and the delicate balance of conditions required to differentiate between false positive results and true regions of variation.

One of the unique features of the Geniom biochips is that the oligonucleotide probes are located within microchannels rather than on the array surface. This presents a number of challenges, for example it is difficult to ensure that the target DNA solution can easily access all regions of the array (and therefore all of the probes) during hybridisation. Additionally, we have shown that this format is susceptible to the formation of networks of DNA molecules within the microchannels. These networks can lead to the creation of a ‘smearing’ effect and mask hybridisations. Adjustments to the protocol, for example fragmenting the target DNA with a double restriction enzyme digest prior to biotin labelling, appeared to go some way to overcoming this limitation, however further investigations are necessary in order to completely overcome this problem.

OaCGH is a powerful method frequently employed for the study and detection of structural variations. A recent WTCCC CNV association study, for example, employed high resolution Agilent microarrays for high-throughput detection of copy number variants (WTCCC, 2010). OaCGH also has potential as a diagnostic tool, for example it can be used to detect and identify aberrations responsible for a range of developmental disorders (Siggberg *et al*, 2010) as well as to distinguish between different subtypes of brain tumour, each of which has a distinct CNV profile, to better inform treatment (Kool *et al*, 2008). However, despite the popularity of this technique, oaCGH does have a number of limitations. We believe that many of the challenges we have encountered

are not specific to the Geniom system, but are likely to be encountered on all microarray platforms. Additionally, oaCGH may not be suitable for the study of loci which show a high degree of sequence similarity to other regions of the genome, since at the design stage oligonucleotide probes must be filtered to leave only those which are unique. This may restrict the coverage of complex regions of the genome, thereby leaving many CNVs undetected. For example, the WTCCC study mentioned above excluded multiallelic and complex CNVs from their investigations as it was not possible to accurately genotype such regions on their chosen platform (WTCCC, 2010). Reproducibility of results both within and across platforms can also be a problem (Draghici *et al*, 2006).

It is becoming increasingly clear that a considerable portion of the genome varies in copy number, which to some extent undermines the notion of a single reference genome. This makes the selection of reference regions for copy number investigations difficult, and emphasises the gaps in our current knowledge of CNV. In the three years since this work begun, new technologies have already begun to emerge for the study of CNVs. For example, it is now possible to employ specially designed SNP arrays to simultaneously genotype SNPs as well as CNVs (McCarroll *et al*, 2008). Another emerging technique, which may in future rival oaCGH as the method of choice for the study of CNVs, is sequencing. Next-generation sequencing platforms such as the 454 (Margulies *et al*, 2005) and ABI SOLiD (Valouev *et al*, 2008) have advantages over microarrays in that they are rapid, lowering in cost, and the sequencing trace data generated can be reused for multiple investigations. Methods are in development for the extraction of CNV information from sequencing data, for example CNV-Seq, which is a statistical package designed to identify regions of CNV from high-throughput sequencing traces (Xie & Tammi, 2009). However, there are currently major limitations

of this technique, in particular concerning the costs associated with both the experimental technique as well as subsequent storage of the data. It is not yet possible to predict whether, as the cost of sequencing becomes increasingly lower, sequencing-based technologies will replace or work alongside oaCGH as the methods of choice for studying genome-wide CNV.

7.2 Characterisation of CNV at 12p13.31

As well as the development of a protocol for oaCGH, my research was concerned with the detailed characterisation of a putative region of structural variation on chromosome 12p13.31, using PCR-based methods. Sequence analysis of this locus revealed the presence of a novel tandem duplication, which we showed to be copy number variable. Several different types of assays were developed to investigate this variation in greater detail.

7.2.1 Assay Development

CNVs are often located within areas of complex genomic structure, for example regions of segmental duplication or with a high density of repetitive elements, which can make studying this form of variation difficult. In our case, there were particular challenges associated with studying the tandem duplication on chromosome 12p13.31 due to the increased frequency of repetitive elements found at this locus, and also the high degree of sequence similarity between the two units of the duplication. In order to overcome these challenges, we adapted the traditional method of PRT to better suit our region of study. Since it proved difficult to identify sequence unique to one unit which also had a single copy paralogue located elsewhere in the genome, we instead amplified sequences in each of the two units simultaneously and looked for relative changes in copy number between the two products.

Initially assays were designed to exploit differences in repetitive elements between the two units of the tandem duplication to amplify products of different sizes, which could be distinguished on an agarose gel. Differences between the ratio of the two products revealed copy number variation within this region. Using the most robust and

reproducible of these, assay A9, we were able to distinguish four classes of variation at this locus. We hypothesised that these four variants could be a result of deletion or duplication of a section of sequence within either of the two units of the duplication.

Assay A9 was only able to determine copy number of the region within which the assay primers were located, therefore to study CNV at different points across the tandem duplication another series of assays, referred to as the B assays, were designed. Combining genotyping data from the two sets of assays enabled us to obtain a clearer picture of how the copy number changed across the region and led to the identification of two distinct intervals within which there is evidence for historical recombination events. One of these, referred to as the B9/B10 breakpoint, has been confirmed as a site of recombination using a specific PCR assay, and appears to be the main location of rearrangements seen in the Caucasian population as well as in four Asian samples. The rearrangements which we have detected appear to fit with a simple model of NAHR between the two units of the tandem duplication at 12p13.31.

Agarose gel electrophoresis was used to separate the assay products and determine the relative amount of each product. Although useful, gel electrophoresis is an imperfect method as it is susceptible to variability, for example between samples at different positions on the gel. Other methods of separating DNA products include capillary-based systems, which have the advantage of being more suitable for high-throughput experiments. The development of PCR-based assays which, through the presence or absence of specific bands, directly reveal the presence of variation, overcome some of the difficulties associated with the use of agarose gels. Since these do not rely on a comparison of the signal intensities of two products, each of which could potentially be altered independently as a result of various gel effects, they are more reliable detectors of CNV.

7.2.2 Studies of Inheritance

It is currently unclear what proportion of CNVs are inherited, and how many are generated as a result of *de novo* events. It has been suggested that new CNVs are constantly being formed (Egan *et al*, 2007); however, other studies are in disagreement, suggesting that over 99% of CNVs are inherited (McCarroll *et al*, 2008). To investigate this in regards to our region of interest, a number of family trios were genotyped from within the HapMap populations (CEPH and Yoruba), along with four larger three-generation CEPH families. The results of this study provided no evidence of a high rate of instability within this region. However so far, due to sample availability, it has only been possible to study the inheritance of A9(B) duplications.

Our studies of family trios have not detected any *de novo* recombination events within this region. A possible explanation for this may be that the number of samples we have genotyped so far is insufficient to detect any *de novo* recombinations. To investigate the frequency of *de novo* recombination events within the 12p13.31 region in more detail, single-molecule amplification of DNA from gametes could be used.

7.2.3 Population Differences

Populations are known to differ in terms of both the number of copy number variable regions identified as well as the diversity of individual CNVs (Redon *et al*, 2006). We detected such differences within our region of interest on chromosome 12p13.31. For example, the frequency of structural variations at this locus is increased in Swedish compared to UK control samples. Additionally, while every Caucasian or Asian recombinant sample genotyped to date shows evidence of historical recombination within an interval flanked by the B9/B10 assays, in two Yoruba variant samples we

have detected the result of a presumptive recombination event distal to this locus, between assays B5 and B6. Further studies of variation at 12p13.31 in individuals of African descent, as well as other populations, will enable us to conclude whether historical recombination events at the B5/B6 position are unique to individuals of African descent, and if rearrangements at the B9/B10 locus also occur within this ancestral group, as well as potentially revealing other points of recombination.

We have detected differences in the frequency of CNV within 12p13.31 between populations, and have discovered two distinct points of historical recombination which vary in individuals of different ethnic backgrounds. This evidence, along with the absence of *de novo* recombinants, suggests that alleles at this locus may share common ancestry. The fact that the frequency of variation at this locus differs between populations might also suggest that these variants may have at some point been selected for. To investigate this further, it would be necessary to genotype a large number of samples from a wide range of populations. Studies of the frequency of *de novo* recombination in gametes could enable us to confirm whether recombination at this locus is still active in the population.

7.3 Association of CNV at 12p13.31 with Complex Disease

A major field of research is concerned with the association of genetic variations, for example SNPs or CNVs, with susceptibility to complex disease. The majority of studies to date, in particular GWAS, have focussed on SNPs; however the number of CNV-focussed investigations has increased rapidly over the last few years, and is likely to continue increasing. A role for CNV has already been identified in susceptibility to many common complex disorders including schizophrenia (International Schizophrenia Consortium, 2008), autism (Sebat *et al*, 2007) and HIV (Gonzalez *et al*, 2005).

Somewhat disappointingly, the recent WTCCC CNV GWAS, which investigated the role of common CNVs in susceptibility to eight complex disorders (including RA), identified only three loci where CNV was associated with disease (WTCCC, 2010). This led them to conclude that common CNVs are unlikely to explain any significant proportion of the remaining ‘missing heritability’ for complex disorders. However, a number of the authors of this paper, including Charles Lee and Stephen Scherer, have since emphasised the limitations of the study (Petrone, 2010). Specifically, rare CNVs and those which were difficult to genotype on the array platform used, such as multiallelic and complex variants, were excluded from the study. It is important not to dismiss the contribution of these variants to complex disease, since many of the CNVs which have currently been associated with disease phenotypes are located within complex genomic regions, for example CNV of the β -defensins has been associated with psoriasis and Crohn’s disease (Hollox *et al*, 2008; Fellermann *et al*, 2006). Additionally, rare structural rearrangements have been implicated in susceptibility to disorders including schizophrenia and autism (International Schizophrenia Consortium, 2008; Pinto *et al*, 2010). Future CNV association studies are therefore likely to

concentrate on these classes of variant, which will necessitate the development of methods better suited to studying rarer and more complex CNVs.

7.3.1 RA Association

Rheumatoid arthritis is a common complex autoimmune disorder with a significant genetic component (heritability is currently estimated at around 60% (Macgregor *et al*, 2000)). A number of loci have been implicated in RA susceptibility; however with the exception of alleles at the HLA locus, these associations tend to have small effect sizes. We have shown that a deletion within 12p13.31 is associated with protection against RA in a Swedish cohort ($p = 0.001$, odds ratio 2.3; 95% CI 1.4-3.9) as well as a UK cohort ($p = 0.036$, OR 1.90; 95% CI 0.93-3.82). This is only the second CNV to be associated with RA, the other being *CCL3L1* (McKinney *et al*, 2008). A similar association was also seen for psoriasis, another autoimmune disease, in a Swedish cohort ($p = 0.013$, odds ratio 2.16; 95% CI 1.2-4.1). Given this result, as well as the observation that autoimmune diseases often tend to share common susceptibility loci (Jawaheer *et al*, 2001), we consider it likely that variation on chromosome 12p13.31 may also be involved in susceptibility to other autoimmune disorders. This is especially suggested due to the fact that one of the genes within this region is an extremely high-affinity glucose transporter, and glucose is an essential requirement of the immune response.

The recent WTCCC CNV investigation did detect a variant within the 12p13.31 tandem duplication, however the consortium did not report an association with any of the eight complex disorders examined, which included RA (WTCCC, 2010). This may be for a number of reasons, in particular that this study had little power to detect associations with rare variants (defined as those with a $MAF < 5\%$), and our investigations suggest

that A9(B) deletions within 12p13.31 occur at a frequency of <1% in the UK population. In addition, this is a complex CNV, which may pose further challenges for accurate detection. Also of concern is that the cited allele frequency for this variant is 48%, which is inconsistent with our data. Interestingly, this study also failed to replicate the previously described association of *CCL3L1* with RA (McKinney *et al*, 2008; WTCCC, 2010).

7.3.2 Physiological Effects of a Deletion within 12p13.31

It is interesting to consider the possible physiological effects of a deletion within the tandem duplication on 12p13.31. *SLC2A3*, one of the two genes located within this region, codes for the membrane glucose transporter GLUT3. Although often referred to as a neuronal glucose transporter, GLUT3 is also expressed in many other tissues and is notable due to its particularly high affinity for glucose. Increased expression of glucose transporters, in particular GLUT1 and GLUT3, is a characteristic feature of cancer cells (Yamamoto *et al*, 1990; Macheda *et al*, 2005). There is evidence to suggest that in some cancers, the over expression of these glucose transporters may be indicative of a poor prognosis, as it signifies an enhanced glycolytic metabolism which is characteristic of malignant cells (Baer *et al*, 2002; Ayala *et al*, 2010).

The deletion of either *SLC2A3* or *NANOG* (another of the genes located within this region) has been shown to be lethal in mice (Mitsui *et al*, 2003; Ganguly *et al*, 2007), however it is not yet known whether this genotype has the same effect in humans. Although individuals heterozygous for an A9(B) deletion have been detected at a frequency of 2.64% in a Swedish population and 0.92% in a UK cohort, no homozygous deletions have been identified in the 15,000 samples genotyped to date. To investigate whether we would have expected to detect any homozygous deletions in a

sample set of this size, the expected allele frequency was calculated for each population (Swedish and UK) according to Hardy-Weinberg equilibrium. The allele frequency of A9(B) deletions in the two populations are 1.32% and 0.46% respectively. Using these values, we calculated that the number of homozygous A9(B) deletions which would be expected to occur in the cohorts studied is less than one for both populations (0.22 for the UK samples and 0.81 in the case of the Swedish samples). These values are consistent with the fact that no homozygous deletions have been detected to date, and therefore a greater number of samples must be genotyped before any conclusions can be drawn as to whether this genotype occurs in humans.

It has been shown that cells of the immune system such as monocytes, B and T cells increase their expression of GLUT proteins, including GLUT3, upon activation, resulting in an increased glucose uptake to fuel the immune response (Maratou *et al*, 2007; Fu *et al*, 2004). It is possible that individuals with a low copy number of *SLC2A3* may have decreased expression of GLUT3 and therefore be less efficient at transporting glucose. Such individuals may not be able to raise and maintain an immune response at the level required for the development of an autoimmune disease, putting them at a reduced risk of such disorders. It would be interesting to investigate whether individuals with a deletion within this region also show a lower than average immune response to general infections.

7.3.3 Chondrocytes in RA Pathogenesis

There is a well established role for chondrocytes, the cells of the cartilage, in articular joint destruction and therefore in RA pathogenesis. One of the early examples of this came from an investigation involving mice in which TNF- α was over expressed,

resulting in high levels of chondrocyte apoptosis (Butler *et al*, 1997). In these mice a relative preservation of cartilage was observed, even in mice with advanced arthritis, which was attributed to the lack of cartilage degradation by chondrocytes.

Facilitated diffusion of glucose into chondrocytes occurs through glucose transporter proteins such as GLUT3 (Reviewed in Mobasheri *et al*, 2008). We have detected an association between a decreased copy number (loss of *SLC2A3*, the gene which codes for GLUT3) and protection against RA. It is therefore possible that a reduction in GLUT3 expression in individuals with an *SLC2A3* deletion may lead to lower levels of cartilage degradation.

So far our investigations have examined the frequency of CNV within 12p13.31 in one specific type of somatic tissue (blood lymphocytes). To gain further insight into potential mechanisms by which CNV in this region may influence RA susceptibility, future work could investigate the effect of altered gene expression levels in chondrocytes, the cells of the cartilage. If a deletion of *SLC2A3* is found to correlate with expression levels, this would offer scope for the development of drugs to treat RA, either by reducing expression of *SLC2A3* (for example through targeted gene knockout or RNAi), or by inhibiting the glucose transporting action of GLUT3. Drugs could be specifically introduced to the joints and the limited vascular supply to the chondrocytes would concentrate the effects locally, preventing a potentially harmful systemic reduction in GLUT3.

7.3.4 Future Perspectives

12p13.31 is only the second region of CNV to be associated with rheumatoid arthritis, although we expect more to be discovered in the coming years. Further work is required

in order to understand more completely the dynamics of the 12p13.31 locus and to determine how the copy number variation described in this thesis may alter the expression of genes within this region. The mechanism by which this variation influences susceptibility to autoimmune disorders also remains to be elucidated. Given the universal importance of genes located within this tandem duplication, it seems likely that genetic variation at this locus may contribute towards susceptibility for many forms of complex disease.

Future genome-wide CNV disease studies will provide a wealth of data which will add to our growing insight of the role in which such variations play in disease. As methods for studying CNV improve, it is likely that many more complex regions of the genome, such as the tandem duplication described in this thesis, will be revealed to contain CNVs. It is probable that there will also be an increasing emphasis on the ways in which different forms of variation, for example SNPs, CNVs and epigenetic interactions, work together to influence disease susceptibility and progression.

Appendix A: Geniom Workflows

SYNTHESIS

Open Argon cylinder : 9.0 bar !
Turn on : geniom device, then application PC
Start : geniom client

DNAPro : _____
Date : _____
Operator : _____

Select module group INSTRUMENT CONTROL
Start module MAINTENANCE

Select your geniom device and press NEXT

Make sure that a wash-DNA-processor is inserted

☐

Set geniom STANDBY, then open front door
Remove old H2O bottle, renew filters, load new H2O bottle
Close front door, then set geniom OPERABLE

☐
☐
☐

Select plc-program SENSORS_CLEANING and press START

☐

(duration : ~ 48 min)
→ Dilute Amidites (separate instructions)

Set geniom STANDBY, then open front door
Remove old chemical bottles, renew filters then load new chemicals in order:
ACNx2, Deblock L, Oxidizer, Deblock F, Cappingx2 (mix first), Activator, Amidites
Close front door, then set geniom OPERABLE

☐
☐
☐

Check your DNA-processor to be synthesized for clean surface

☐

Eject wash-DNA-processor and insert your new DNA-processor

☐

Stop module MAINTENANCE
Start module SYNTHESIS

Select your geniom device and press NEXT

Create your DNAProcessor number with DNAProCreator software and import into geniom software

☐

Select your DNA-processor from database and press NEXT

☐

Select array template for each array (drag&drop) and press NEXT

☐

array template : _____

Make sure that all requirements are fulfilled, then press NEXT

☐

Pre-Synthesis is running :
purge/ check tightness /optical calibration /write data to database

(duration : ~ 40 min)

GENIOM SYNTHESIS RUNNING

☐

After synthesis is completed successfully : Stop module SYNTHESIS

Start module MAINTENANCE

Select your geniom device and press NEXT

If you don't want to continue with hybridization:
Eject your synthesized DNA-processor and insert a wash-DNA-processor and proceed with HYB_DENATURING and HYBPATHWAY_DRYING.

proceed with : EXTERNAL INCUBATION

EXTERNAL INCUBATION

from *SYNTHESIS*

Unless you just finished the SYNTHESIS:

DNAPro : _____
Date : _____
Operator : _____

Select module group INSTRUMENT CONTROL
Start module MAINTENANCE
Select your geniom device and press NEXT
Eject wash-DNA-processor and insert your synthesized DNA-processor

Select plc-program HYB_DENATURING and press START

☐ (duration : ~ 20 min)

Eject your synthesized DNA-processor and insert a wash DNA-processor

☐
→ Insert DNA-processor in
Hybridization holder*

Select plc-program HYBPATHWAY_DRYING and press START

☐ (duration : ~ 30 min)

Stop module MAINTENANCE

☐

* EXTERNAL HYBRIDISATION HOLDER:

Make sure that your external hybridisation holder is clean

☐

Make sure that sponges are wet

☐

Insert your synthesized DNA-processor

☐

Fill in each well 15 µl PREHYBRIDISATION BUFFER

☐

Move PREHYBRIDISATION BUFFER into arrays by use of syringe

☐

Close lid & incubate at RT for 15 min

☐

Denature target at 95°C for 3min, place on ice, spin down, place on ice again

☐

Remove PREHYBRIDISATION BUFFER from arrays by use of syringe

☐

Fill in each well 10-15 µl of your TARGET-solution

☐

Move TARGET-solution into arrays by use of syringe

☐

Close lid & incubate at incubation temperature (oven) for 4-16 hrs

☐

After removal of hybridization holder from oven let cool down for 20 min

☐
→ Prepare Geniom with
buffer solutions**

Make sure that GENIOM device is ready for washing routines

☐

Remove TARGET-solution from arrays by use of syringe & pipette

☐

If multiple samples used each array can be washed with 15ul 6xSSPE to avoid
contamination between arrays

Remove slowly your hybridized DNA-processor from holder

☐

Check the surface of your hybridized DNA-processor for residues. In case of residues
clean the surface carefully.

Quickly insert the DNA processor in the Geniom
proceed with : WASHING & MARKER

→ ***

WASHING & MARKER

		DNAPro : _____ Date : _____ Operator : _____	
<i>from EXTERNAL INCUBATION</i>			
**	Select module group INSTRUMENT CONTROL Start module MAINTENANCE		
	Select your geniom device and press NEXT		
	Make sure that a wash-DNA-processor is inserted	<input type="checkbox"/>	
	Prepare fresh SAPE-solution (44µl SAPE + 9ml 6x SSPE) in a 15ml Falcon tube	<input type="checkbox"/>	
	Set geniom STANDBY, then open front door Remove tube & replace with fresh SAPE-solution in PRE-BUFFER Place fresh de-ionized water in H2O Renew filters & place fresh buffer-solutions in BUFFER-1, -2, -3 Close front door, then set geniom OPERABLE	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	NO filters for Pre-Buffer ! BUFFER-1 = stringent BUFFER-2 = non-stringent BUFFER-3 = non-stringent
	Select plc-program HYB_CLEANARRAY_AR and press START	<input type="checkbox"/>	(duration : ~ 3 min)
***	Replace wash-DNA-processor with your hybridized DNA-processor	<input type="checkbox"/>	
	Stop module MAINTENANCE		
	Start module HYBRIDIZATION		
	Select your geniom device and press NEXT		
	Select your DNA-processor from database and press NEXT	<input type="checkbox"/>	
	Select HYBRIDIZATION PROFILE (drag & drop) Check all arrays to be processed and press NEXT	<input type="checkbox"/> <input type="checkbox"/>	
	Make sure waste bottle is empty Make sure that all requirements are fulfilled and press NEXT	<input type="checkbox"/> <input type="checkbox"/>	
	Press YES in dialog window to start the washing & marker routines	<input type="checkbox"/>	
	GENIOM WASHING & MARKER ROUTINES RUNNING	<input type="checkbox"/>	
	Stop module HYBRIDIZATION		

proceed with DETECTION & ANALYSIS

DETECTION & ANALYSIS

from WASHING & MARKER

DNAPro : _____
Date : _____
Operator : _____

Start module DETECTION
Select your geniom device and press NEXT
Select dna-processor from list and press NEXT
Check acquisition configuration
Press START DETECTION
optional: save image to TIF-file
Press NEXT with image analysis
Select arrays/hybridizations for analysis and press NEXT
Enter experiment description
Save analysis data to database (WRITE TO DB)
optional: save data to Excel-file
Stop module DETECTION
Start module MAINTENANCE
Select your geniom device and press NEXT
Select plc-program LAMP_OFF and press START
Select plc-program HYB_WASH_80C and press START
Eject your DNA-processor and insert a wash-DNA-processor
Stop module MAINTENANCE

standard parameters :

autoexposure : yes

mean over 16 : no

hardware gain : 1

filter : Cy3

background subtraction:
yes

☐
☐
☐
☐
☐

check data is written!

check data is written!


☐
☐

(duration : ~ 30 min)

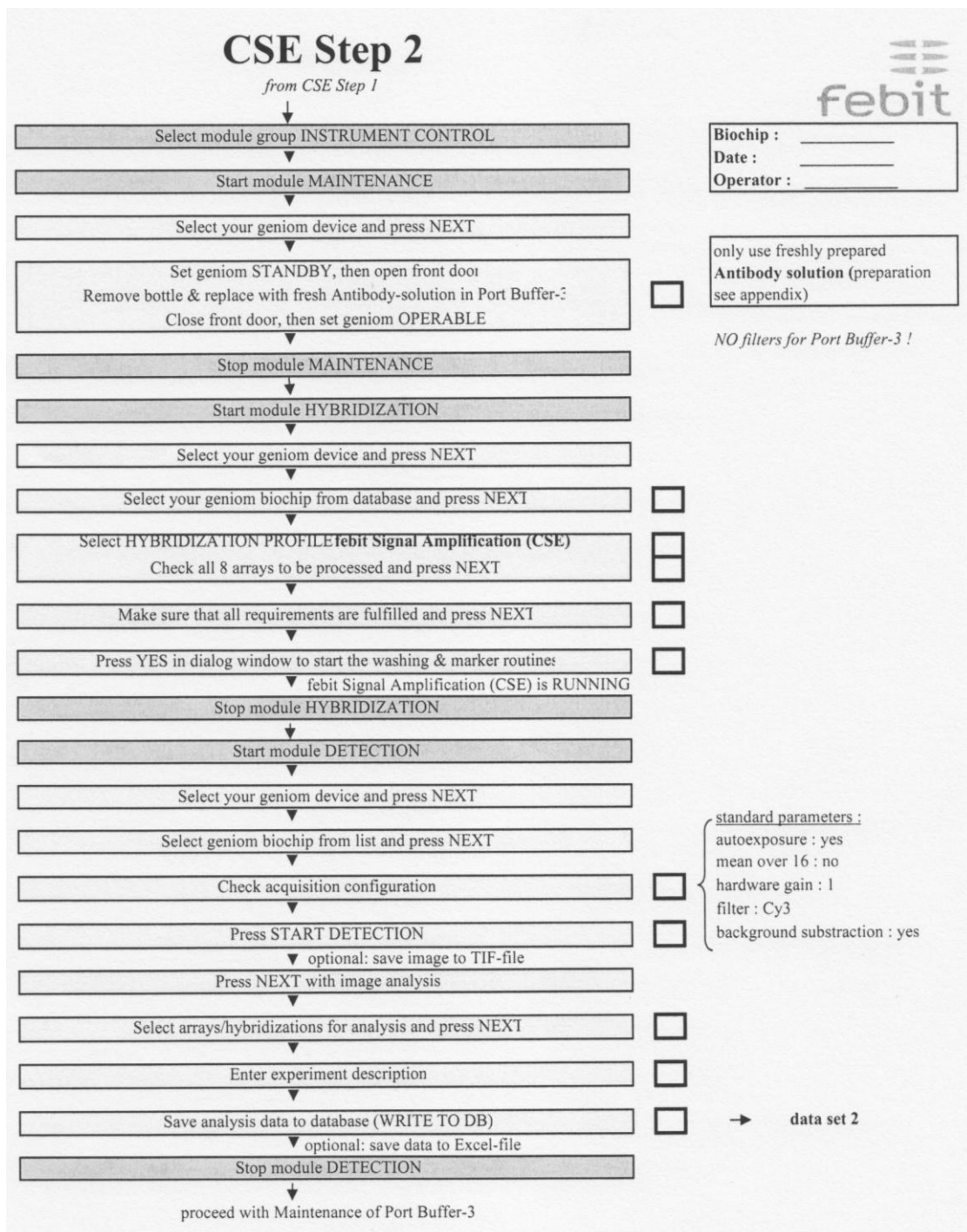
☐

proceed with new Synthesis, Hybridization or SHUT DOWN geniom device

Wash Profile

Geniom Hybridization Profile						
Leicester Expression Standard 40°C(external incubation)						
Date	2007-11-15	Author	BeH	Version	1.0	

Step	Description	Port (default)	Temp [°C]	Flow [µl/min]	Time [s]	Volume [µl]
A Non-stringent washing at 25°C						
2	Set temperature		25			
2	Flow buffer	variable (Buffer-2)	25	500	15	125
3	Pause		25		30	
4	Flow buffer	variable (Buffer-2)	25	500	5	42
5	Pause		25		30	
6	Flow buffer	variable (Buffer-2)	25	500	5	42
7	Pause		25		30	
8	Flow buffer	variable (Buffer-2)	25	500	5	42
9	Pause		25		30	
10	Flow buffer	variable (Buffer-2)	25	500	5	42
B Stringent washing at 40°C						
1	Fill with buffer	Buffer-2 (6x SSPE)				
2	Set temperature		40			
2	Flow buffer	variable (Buffer-1)	40	500	15	125
3	Pause		40		30	
4	Flow buffer	variable (Buffer-1)	40	500	5	42
5	Pause		40		30	
6	Flow buffer	variable (Buffer-1)	40	500	5	42
7	Pause		40		30	
8	Flow buffer	variable (Buffer-1)	40	500	5	42
9	Pause		40		30	
10	Flow buffer	variable (Buffer-1)	40	500	5	42
C Staining with SAPE-solution						
11	Fill array with buffer	Buffer-2	25	500	15	125
12	Flow buffer	Pre-Buffer	25	500	20	167
13	Flow buffer	Pre-Buffer	25	500	20	167
14	Pause		25		20	
15	Grab Image (Cy3, 500ms)		25			
D Incubation with SAPE-solution						
16	Incubation - Pause		25		900	
E Non-stringent washing at 25°C						
17	Set temperature		25			
18	Flow buffer	variable (Buffer-2)	25	500	15	125
19	Pause		25		30	
20	Flow buffer	variable (Buffer-2)	25	500	5	42
21	Pause		25		30	
22	Flow buffer	variable (Buffer-2)	25	500	5	42
23	Pause		25		30	
24	Flow buffer	variable (Buffer-2)	25	500	5	42
25	Pause		25		30	
26	Flow buffer	variable (Buffer-2)	25	500	5	42



Appendix B: Oligonucleotides

PCR Primers

Table 1: Boundary Primers (Chapter 4)

Primer	Sequence
1F	CCCTAACTGCTTCTACATCCC
1R	ATCGTAAGAACTCCCATTTCACC
2F	AGCACTTGAACTCCTGGTAGG
2R	TTAGCAAGAGTACAAATATACACCC
3F	TGCAAGTACAATGACAGTTACTCT
3R	CAGGATTTAGTGCTTGCCACC
UNREP1F	TGGGTAACAGCTCAAGATTCC
UNREP1R	AGTGGAATTCTTGGATCGTGG
UNREP2F	AAGCAATTCTCTTGACACCAGC
UNREP2R	TGGGTAGATCTCTTGACGTTAGG
3FREP	AAGGCTGTGATTTAGACCTCTGG
3RREP	AGTGCAATGTTGCATCTCGG
2FB	ACAGCACGAAGGTGTTTTGG
3FB	GTCCATCTCTCTGCTATCCTAGG
3RB	AGGATTTAGTGCTTGCCACC

Table 2: Gaps Primer Sequences (Chapter 4)

Gap Number	Gap Location	Gap Size	Forward Primer	Sequence	Reverse Primer	Sequence
1	7838572-7838970	398bp	12	TGAGACATCATAATGACATTAGC	6	GTAGATGCACATGTACAATGC
			12a	ATAGCAATGGTGTGACGCAGG	6a	TGACTGGATGGGCATCATGG
2	7858914-7859030	116bp	6	CCGTCCTTCCCAGATCTATC	29	GACTACTGTAATTACTGACGC
3	7886905-7887766	861bp	28	CTCTGTGTAGCTATCAGTGG	9a	CTATAGACTGGAACCAGCAC
4	7892157-7893387	1230bp	22a	AACAGGCGCAGACATAGGTG	22	CAGCATTATCTCTTCTAGACTC
5	7927103-7927787	684bp	4	CTTGAGACTTGTTCACTTGG	42	GAGTTCAAATTATAGCACATAGTG
6	7968270-7968880	610bp	7a	GCTACAAAGACCAAGATAGCC	10a	GATGTAGTCTTCACTCTTCAAGC
7	7997990-7998704	714bp	17a	GTAAGTCTGAGATGTTGTTAGG	51	CCTGCCTGTCCTCTGAAGTC
8	8001389-8001555	166bp	39a	CTCAGAACATGCCATAGTCTG	17	ATAGATTGCCATGTGCTTCTG
9	8004263-8004312	49bp	32a	CATTTACGCTTCCACACAG	35aR	CACGTCTGCACTTGCTTACG
10	8015595-8016453	858bp	35aF	AGCTTGCCACCATGCACAGC	33	GACCTCTTCACTCACGAGGC
11	8021345-8021773	428bp	31	AACAATAGGCACAGACATAGG	25	GAGCTAGCACCAGTGGTATTCTG
			33a	GTGTCTGGCCAACCTCCATC	13a	TAACAGTCTTGCTCTCTTGCC
12	8043200-8043431	231bp	46	GCTAAAGAAGTGAGGCAGGG	1aR	CAGTGTGATAGCTGGTGGAAGC
13	8050277-8050530	253bp	1aF		1	GGTGCTGCTTCACCTTGCAC

Table 3: Recombination Assay Primers for B5/B6 Interval (Chapter 6)

Primer	Sequence
AF1	CCTGTTGAGTAATAGAACACCTAGG
AF2	ATTCAGCTGACTCAGGAATCCC
AR1	AAAAGGACTTCTTACCTCCTGGG
AR2	CCACTTCAGAGTGGTATGTGC
BF1	CTGGTTTTAAGATAAGGTGAGTTTAC
BF2	AAGTCCCAGTGGGTAAGAAGTCC
BF3	CTAAAGTTGCTATAACTACCATAG
BF4	TGCTCTAGGGCTAATTAACACTG
BF5	ATGCAACCTCTTACCCATCTGC
BR	AAGGACTTCTTACCCACTGGGAC
UF1	CATATAGTCACTTTGTACTGAGACC
UR	CTGAGGCAACTCCCTAGCTG
UR6	TGCTACAGATGTTCAAAGATGG

Table 4: Recombination Assay Primers for B9/B10 Interval (Chapter 6)

Primer	Sequence
AR3	TCAGTTTTCCCTCAAACCTTAGC
AR3B	CCACCGTTTCCTTGACCTC
AR4	TATAGACTGGAACCAGCACTAGG
BF6	CAACTCAATGTAAAAGGGAAGGC
BF7	TTGTGCCGGTTCAAGTCTACAG
BF8	CTAAACAGAAACAAGTGGGCTAC
B9F	GCTTGCATGAAGTGAAGACAG
UR2	TTTGGGAATGGAGGAGCATG
UR3	GGAGGTTCTTTCTGTGCAC
UR4	GTA CT CGTAGACACGATGCTCAGTG
UR5	CCAACACCACCACATAGTACTC

PRT-based Assay Primers

Table 5: A Assay Primers

Assay	Forward Primer	Reverse Primer
A1	CTGTATATCACATTGATGTGCAG	ACTACTCTTCTAGAGAGAACCAGG
A2	TCATTTTCTCTCATGCCTTTACCC	GAGCCTCCGGAGTGAAAGACC
A3	AGAGTGCAGAGGAGAATGAGTC	ATTTTCCTTYCTATTCCCAAACC
A4	GTTGAAGGACACAGAATTCGG	CAGAGAATGGAGAAGAGACAGCC
A5	TTAGATCTACTTATCTATAGCCAGAGAC	GTATGGTTGGAGCCTAATCAG
A6	TGTTCTGGTTTCCATGATGCC	TTCTTGAGGTTAGATTCTAAACCC
A7	GGAGGTAAATACCTTTMAGAGTAC	GGYAGTTAACCTCAGAGTATAGC
A8	GGAAATATTCACAATCTTCTCAGTGAG	TTCTTGGTGCTTACACATCTCAG
A9	TATTGCACCTTAACCTCTCCAGC	CCTCACTTCATACAGCTCTACG
A10	TACRTAGGCTGGTTTTAAGATAAG	ACTTCCTAGTCGGATTGCTC
A12	CATACTRTATAGTCCCTTCTATAC	GATGTTCAAAGATGGATTGC
A13	GCCTAGGCCCTTTTGAAGTG	ATGTTCCCTTGTTAACTTGAGG
A14	GATATTTGCATGGAATACAAGGG	GTGTGATACTTTAGGGTCAAACG
A16	CGGAAAATATGGCCACAGAC	GTCTATTTATTCCCTTGATTCTCTAGG
A17	GATCTACTCACTGCATCCTGGC	GAATTATTCCAGGCAAGGTGAAG
A18	ACGCATTGAGGTAAAGACAGG	GCAAAGTTTTAGAGCAGTTACTAGC
A19	AGAGAAGCAGTGACAGAAGCAG	TCTAGAAGTAAATGTGTTGATGGC
A20	GCCTGTCATCTTCTGACGTG	GGAYGACTAGGTAATAGAGACTGG
A21A	TCCAAAGGCTCATATGTTGC	TAGAGCAGACTATAGAGTGTAAGTG
A21B	CAGTTACACTCTATAGTCTGCTC	TTTATGATCAGCATTGTAATTGC

Table 6: Trio Assay Primers

Assay	Forward	Reverse A	Reverse B	Size Product A	Size Product B
B1	TGCCTGGTTCTCTCTAGAAGAG	TGAATGGATCATTACCAGC	TTGTTAGAGATGCTAACTGATGG	228bp	180bp
B2	TGTCTCTTCTCCATTCTCTGG	GTTACCAATGGCATTCTTGG	AGAAGCGTAAGTACTAGCACAGG	343bp	210bp
B3	GGATCTGCAGAACAGATGTACCTG	TCCTATGTTGTTGAGTGCTGTG	GCTAAGTATGATTGCAGGCCTG	815bp	765bp
B4	TTCGGCATAGGACCACAGTG	CCTGAGGTCTGTACATTACCTC	TGTAGTTAATGATGGTTGATGGG	616bp	438bp
B5	AGCAATCCGACTAGGAAGTTGG	CTTGCCATTAGTTGCCTCTGG	GTGAGCTATTGTAGTGCCACTGC	270bp	407bp
B6	CAGCTAGGGAGTTGCCTCAG	GGAAGAGAGTACTCCTGTGCC	CAGAATCTATGAGGAGCCAAGTG	701bp	439bp
B7	GGTAGTAGAGCTCCAAGCAAGG	GAGGTGCTGCTCACGAATCTC	CTGTAACAGTCTGTTCTTCCACC	280bp	379bp
B8	CTAGGAGTCTGCTCATGCAGG	GAGTGATGTGTCCTCACATGG	CCTCTGCACTCAAGCAATCC	749bp	1030bp
B9	GCTTGCATGAAGTGAAGACAG	GGTAGCCTGGTGATTGTGTCC	GTGGGCATCTGTAATCCAGG	119bp	352bp
B9A	GGAGTCAAGAATGATCCTAAAGG	CATTACCCTAGCATCATCGC	TTATCTGTATATTCCCAGCATCTG	150bp	206bp
B10	AGGATTCTGCAGAAAGCAGC	AGAGATGGCAGGAGAACAGTC	CTAACACGATGAAGCCTTGTCTC	289bp	173bp
B11	TTAGGAAGATGATAGGTACTGTG	CAACATAACGTATCTCAGCAGAC	TCGTAACAGCTACAGCATCG	750bp	796bp

Appendix C: DNA Samples

CEPH Trios (The origin of each trio (maternal or paternal) to the third generation is shown)

Trio Number	Family Number	Maternal/Paternal
1	1334	Paternal
2	1334	Maternal
3	1340	Paternal
4	1340	Maternal
5	1341	Paternal
6	1341	Maternal
7	1344	Paternal
8	1345	Maternal
9	1346	Paternal
10	1347	Maternal
11	1349	Maternal
12	1350	Paternal
13	1350	Maternal
14	1358	Paternal
15	1362	Paternal
16	1362	Maternal
17	1375	Paternal
18	1408	Paternal
19	1408	Maternal
20	1416	Paternal
21	1420	Paternal
22	1420	Maternal
23	1444	Maternal
24	1447	Paternal
25	1447	Maternal
26	1454	Paternal
27	1454	Maternal
28	1459	Paternal
29	1459	Maternal
30	1463	Maternal

Yoruba Trios

Trio	Family Number
1	4
2	5
3	9
4	12
5	13
6	16
7	17
8	18
9	23
10	24
11	28
12	40
13	42
14	43
15	45
16	47
17	48
18	50
19	51
20	56
21	58
22	60
23	71
24	72
25	74
26	77
27	101
28	105
29	112
30	117

Coriell Samples (Used in Chapter 6 for B assays)

Sample ID	Coriell Reference Number	Population
D1	NA07048	CEPH
D2	NA10847	CEPH
D3	NA10861	CEPH
D4	NA11920	CEPH
D5	NA12043	CEPH
D6	NA14700	Black
H1	NA07034	CEPH
H2	NA10854	CEPH
H3	NA11840	CEPH
H4	NA12236	CEPH
H5	NA12753	CEPH
H6	NA12762	CEPH
H7	NA12763	CEPH
H8	NA18502	Yoruba
H9	NA18507	Yoruba
H10	NA19092	Yoruba
H11	NA19094	Yoruba

References

- 1000 Genomes. Available at: www.1000genomes.org [Accessed August 2010].
- Abramoff, M.D., Magelhaes, P.J. & Ram, S.J. (2004) Image Processing with ImageJ. *Biophotonics International*, 11(7); 36.
- Aldred, P.M., Hollox, E.J. & Armour, J.A. (2005) Copy number polymorphism and expression level variation of the human alpha-defensin genes DEFA1 and DEFA3. *Human Molecular Genetics*, 14(14); 2045-2052.
- Armengol, L., Villatoro, S., Gonzalez, J.R., Pantano, L., Garcia-Aragones, M., Rabionet, R., Caceres, M. & Estivill, X. (2009) Identification of copy number variants defining genomic differences among major human groups. *Public Library of Science One*, 4(9); e7230.
- Armour, J.A., Palla, R., Zeeuwen, P.L., den Heijer, M., Schalkwijk, J. & Hollox, E.J. (2007) Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats. *Nucleic Acids Research*, 35(3); e19.
- Armour, J.A., Sismani, C., Patsalis, P.C. & Cross, G. (2000) Measurement of locus copy number by hybridisation with amplifiable probes. *Nucleic Acids Research*, 28(2); 605-609.
- Arnett, F.C., Edworthy, S.M., Bloch, D.A., McShane, D.J., Fries, J.F., Cooper, N.S., Healey, L.A., Kaplan, S.R., Liang, M.H. & Luthra, H.S. (1988) The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis and Rheumatism*, 31(3); 315-324.
- Ayala, F.R., Rocha, R.M., Carvalho, K.C., Carvalho, A.L., da Cunha, I.W., Lourenco, S.V. & Soares, F.A. (2010) GLUT1 and GLUT3 as potential prognostic markers for Oral Squamous Cell Carcinoma. *Molecules*, 15(4); 2374-2387.
- Baer, S., Casaubon, L., Schwartz, M.R., Marcogliese, A. & Younes, M. (2002) Glut3 expression in biopsy specimens of laryngeal carcinoma is associated with poor survival. *The Laryngoscope*, 112(2); 393-396.
- Barker, D.L., Hansen, M.S., Faruqi, A.F., Giannola, D., Irsula, O.R., Lasken, R.S., Latterich, M., Makarov, V., Oliphant, A., Pinter, J.H., Shen, R., Sleptsova, I., Ziehler, W. & Lai, E. (2004) Two methods of whole-genome amplification enable accurate genotyping across a 2320-SNP linkage panel. *Genome Research*, 14(5); 901-907.
- Barton, A., Thomson, W., Ke, X., Eyre, S., Hinks, A., Bowes, J., Plant, D., Gibbons, L.J., Wellcome Trust Case Control Consortium, YEAR Consortium, BIRAC Consortium, Wilson, A.G., Bax, D.E., Morgan, A.W., Emery, P., Steer, S., Hocking, L., Reid, D.M., Wordsworth, P., Harrison, P. & Worthington, J. (2008) Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. *Nature Genetics*, 40(10); 1156-1159.

- Baum, M., Bielau, S., Rittner, N., Schmid, K., Eggelbusch, K., Dahms, M., Schlauersbach, A., Tahedl, H., Beier, M., Guimil, R., Scheffler, M., Hermann, C., Funk, J.M., Wixmerten, A., Rebscher, H., Honig, M., Andreae, C., Buchner, D., Moschel, E., Glathe, A., Jager, E., Thom, M., Greil, A., Bestvater, F., Obermeier, F., Burgmaier, J., Thome, K., Weichert, S., Hein, S., Binnewies, T., Foitzik, V., Muller, M., Stahler, C.F. & Stahler, P.F. (2003) Validation of a novel, fully integrated and flexible microarray benchtop facility for gene expression profiling. *Nucleic Acids Research*, 31(23); e151.
- Begovich, A.B., Carlton, V.E., Honigberg, L.A., Schrodi, S.J., Chokkalingam, A.P., Alexander, H.C., Ardlie, K.G., Huang, Q., Smith, A.M., Spoerke, J.M., Conn, M.T., Chang, M., Chang, S.Y., Saiki, R.K., Catanese, J.J., Leong, D.U., Garcia, V.E., McAllister, L.B., Jeffery, D.A., Lee, A.T., Batliwalla, F., Remmers, E., Criswell, L.A., Seldin, M.F., Kastner, D.L., Amos, C.I., Sninsky, J.J. & Gregersen, P.K. (2004) A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *American Journal of Human Genetics*, 75(2); 330-337.
- Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Keira Cheetham, R., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S.V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H., Walter, K., Wu, X., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Chiara E Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fuentes Fajardo, K.V., Scott Furey, W., George, D., Gietzen, K.J., Goddard, C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., Hoshler, K., Hurwitz, S., Ivanov, D.V., Johnson, M.Q., James, T., Huw Jones, T.A., Kang, G.D., Kerelska, T.H., Kersey, A.D., Khrebtukova, I., Kindwall, A.P., Kingsbury, Z., Kokko-Gonzales, P.I., Kumar, A., Laurent, M.A., Lawley, C.T., Lee, S.E., Lee, X., Liao, A.K., Loch, J.A., Lok, M., Luo, S., Mammen, R.M., Martin, J.W., McCauley, P.G., McNitt, P., Mehta, P., Moon, K.W., Mullens, J.W., Newington, T., Ning, Z., Ling Ng, B., Novo, S.M., O'Neill, M.J., Osborne, M.A., Osnowski, A., Ostadan, O., Paraschos, L.L., Pickering, L., Pike, A.C., Pike, A.C., Chris Pinkard, D., Pliskin, D.P., Podhasky, J., Quijano, V.J., Racz, C., Rae, V.H., Rawlings, S.R., Chiva Rodriguez, A., Roe, P.M., Rogers, J., Rogert Bacigalupo, M.C., Romanov, N., Romieu, A., Roth, R.K., Rourke, N.J., Ruediger, S.T., Rusman, E., Sanches-Kuiper, R.M., Schenker, M.R., Seoane, J.M., Shaw, R.J., Shiver, M.K., Short, S.W., Sizto, N.L., Sluis, J.P., Smith, M.A., Ernest Sohna Sohna, J., Spence, E.J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C.L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S.M., Wakelin, S., Walcott, G.C., Wang, J., Worsley, G.J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J.C., Hurles, M.E., McCooke, N.J., West, J.S., Oaks, F.L., Lundberg, P.L., Klenerman, D., Durbin, R. & Smith, A.J. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218); 53-59.
- Bentley, D.R. (2006) Whole-genome re-sequencing. *Current Opinion in Genetics & Development*, 16(6); 545-552.

- Botstein, D., White, R.L., Skolnick, M. & Davis, R.W. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32(3); 314-331.
- Bottini, N., Vang, T., Cucca, F. & Mustelin, T. (2006) Role of PTPN22 in type 1 diabetes and other autoimmune diseases. *Seminars in Immunology*, 18(4); 207-213.
- Bottini, N., Musumeci, L., Alonso, A., Rahmouni, S., Nika, K., Rostamkhani, M., MacMurray, J., Meloni, G.F., Lucarelli, P., Pellecchia, M., Eisenbarth, G.S., Comings, D. & Mustelin, T. (2004) A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. *Nature Genetics*, 36(4); 337-338.
- Britten, R.J. (1994) Evidence that most human Alu sequences were inserted in a process that ceased about 30 million years ago. *Proceedings of the National Academy of Sciences of the United States of America*, 91(13); 6148-6150.
- Bruder, C.E., Piotrowski, A., Gijsbers, A.A., Andersson, R., Erickson, S., de Stahl, T.D., Menzel, U., Sandgren, J., von Tell, D., Poplawski, A., Crowley, M., Crasto, C., Partridge, E.C., Tiwari, H., Allison, D.B., Komorowski, J., van Ommen, G.J., Boomsma, D.I., Pedersen, N.L., den Dunnen, J.T., Wirdefeldt, K. & Dumanski, J.P. (2008) Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *American Journal of Human Genetics*, 82(3); 763-771.
- Brueck, C., Song, S. & Collins, J. (2007) Oligonucleotide Array CGH Analysis of a Robust Whole Genome Amplification Method. *BioTechniques*, 42(2); 230-233.
- Burchill, M.A., Yang, J., Vang, K.B. & Farrar, M.A. (2007) Interleukin-2 receptor signaling in regulatory T cell development and homeostasis. *Immunology Letters*, 114(1); 1-8.
- Butler, D.M., Malfait, A.M., Mason, L.J., Warden, P.J., Kollias, G., Maini, R.N., Feldmann, M. & Brennan, F.M. (1997) DBA/1 mice expressing the human TNF-alpha transgene develop a severe, erosive arthritis: characterization of the cytokine cascade and cellular composition. *Journal of Immunology*, 159(6); 2867-2876.
- Carvalho, B., Ouwerkerk, E., Meijer, G.A. & Ylstra, B. (2004) High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. *Journal of Clinical Pathology*, 57(6); 644-646.
- Chance, P.F., Alderson, M.K., Leppig, K.A., Lensch, M.W., Matsunami, N., Smith, B., Swanson, P.D., Odelberg, S.J., Disteche, C.M. & Bird, T.D. (1993) DNA deletion associated with hereditary neuropathy with liability to pressure palsies. *Cell*, 72(1); 143-151.
- Cheung, V.G., Burdick, J.T., Hirschmann, D. & Morley, M. (2007) Polymorphic variation in human meiotic recombination. *American Journal of Human Genetics*, 80(3); 526-530.
- Chevreur, B., Wetter, T. & Suhai, S. (1999) Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB)*, (99); 45-56.
- CombiMatrix. Available at: http://www.combimatrix.com/tech_microarrays.htm [Accessed July 2010].

- Cornelis, F., Faure, S., Martinez, M., Prud'homme, J.F., Fritz, P., Dib, C., Alves, H., Barrera, P., de Vries, N., Balsa, A., Pascual-Salcedo, D., Maenaut, K., Westhovens, R., Migliorini, P., Tran, T.H., Delaye, A., Prince, N., Lefevre, C., Thomas, G., Poirier, M., Soubigou, S., Alibert, O., Lasbleiz, S., Fouix, S., Bouchier, C., Liote, F., Loste, M.N., Lepage, V., Charron, D., Gyapay, G., Lopes-Vaz, A., Kuntz, D., Bardin, T., Weissenbach, J. & ECRAF (1998) New susceptibility locus for rheumatoid arthritis suggested by a genome-wide linkage study. *Proceedings of the National Academy of Sciences of the United States of America*, 95(18); 10746-10750.
- Cummings, P. (2009) The relative merits of risk ratios and odds ratios. *Archives of Pediatrics & Adolescent Medicine*, 163(5); 438-445.
- Daser, A., Thangavelu, M., Pannell, R., Forster, A., Sparrow, L., Chung, G., Dear, P.H. & Rabbitts, T.H. (2006) Interrogation of genomes by molecular copy-number counting (MCC). *Nature Methods*, 3(6); 447-453.
- De Keyser, F., Elewaut, D., Vermeersch, J., De Wever, N., Cuvelier, C. & Veys, E.M. (1995) The role of T cells in rheumatoid arthritis. *Clinical Rheumatology*, 14 Suppl 25-9.
- De Preter, K., Speleman, F., Combaret, V., Lunec, J., Laureys, G., Eussen, B.H., Francotte, N., Board, J., Pearson, A.D., De Paepe, A., Van Roy, N. & Vandesompele, J. (2002) Quantification of MYCN, DDX1, and NAG gene copy number in neuroblastoma using a real-time quantitative PCR assay. *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc*, 15(2); 159-166.
- Deighton, C.M., Walker, D.J., Griffiths, I.D. & Roberts, D.F. (1989) The contribution of HLA to rheumatoid arthritis. *Clinical Genetics*, 36(3); 178-182.
- Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. (2010) Rare variants create synthetic genome-wide associations. *Public Library of Science Biology*, 8(1); e1000294.
- Ding, B., Padyukov, L., Lundstrom, E., Seielstad, M., Plenge, R.M., Oksenberg, J.R., Gregersen, P.K., Alfredsson, L. & Klareskog, L. (2009) Different patterns of associations with anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis in the extended major histocompatibility complex region. *Arthritis and Rheumatism*, 60(1); 30-38.
- Diskin, S.J., Hou, C., Glessner, J.T., Attiyeh, E.F., Laudenslager, M., Bosse, K., Cole, K., Mosse, Y.P., Wood, A., Lynch, J.E., Pecor, K., Diamond, M., Winter, C., Wang, K., Kim, C., Geiger, E.A., McGrady, P.W., Blakemore, A.I., London, W.B., Shaikh, T.H., Bradfield, J., Grant, S.F., Li, H., Devoto, M., Rappaport, E.R., Hakonarson, H. & Maris, J.M. (2009) Copy number variation at 1q21.1 associated with neuroblastoma. *Nature*, 459(7249); 987-991.
- Dorner, T., Egerer, K., Feist, E. & Burmester, G.R. (2004) Rheumatoid factor revisited. *Current Opinion in Rheumatology*, 16(3); 246-253.
- Draghici, S., Khatri, P., Eklund, A.C. & Szallasi, Z. (2006) Reliability and reproducibility issues in DNA microarray measurements. *Trends in Genetics : TIG*, 22(2); 101-109.
- Egan, C.M., Sridhar, S., Wigler, M. & Hall, I.M. (2007) Recurrent DNA copy number variation in the laboratory mouse. *Nature Genetics*, 39(11); 1384-1389.

- Eichler, E.E. (2001) Segmental duplications: what's missing, misassigned, and misassembled--and should we care? *Genome Research*, 11(5); 653-656.
- Ettinger, R., Kuchen, S. & Lipsky, P.E. (2008) Interleukin 21 as a target of intervention in autoimmune disease. *Annals of the Rheumatic Diseases*, 67 Suppl 3iii83-6.
- Etzel, C.J., Chen, W.V., Shepard, N., Jawaheer, D., Cornelis, F., Seldin, M.F., Gregersen, P.K., Amos, C.I. & for the North American Rheumatoid Arthritis Consortium (2006) Genome-wide meta-analysis for rheumatoid arthritis, *Human Genetics*, 119(6); 634-641.
- Feldmann, M. (1996) What is the mechanism of action of anti-tumour necrosis factor-alpha antibody in rheumatoid arthritis? *International Archives of Allergy and Immunology*, 111(4); 362-365.
- Fellermann, K., Stange, D.E., Schaeffeler, E., Schmalzl, H., Wehkamp, J., Bevins, C.L., Reinisch, W., Teml, A., Schwab, M., Lichter, P., Radlwimmer, B. & Stange, E.F. (2006) A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *American Journal of Human Genetics*, 79(3); 439-448.
- Feuk, L., Carson, A.R. & Scherer, S.W. (2006) Structural variation in the human genome. *Nature Reviews Genetics*, 7(2); 85-97.
- Fisher, R., (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.
- Florijn, R.J., Bonden, L.A., Vrolijk, H., Wiegant, J., Vaandrager, J.W., Baas, F., den Dunnen, J.T., Tanke, H.J., van Ommen, G.J. & Raap, A.K. (1995) High-resolution DNA Fiber-FISH for genomic DNA mapping and colour bar-coding of large genes. *Human Molecular Genetics*, 4(5); 831-836.
- Flornes, L.M., Bryceson, Y.T., Spurkland, A., Lorentzen, J.C., Dissen, E. & Fossum, S. (2004) Identification of lectin-like receptors expressed by antigen presenting cells and neutrophils and their mapping to a novel gene complex. *Immunogenetics*, 56(7); 506-517.
- Fredman, D., White, S.J., Potter, S., Eichler, E.E., Den Dunnen, J.T. & Brookes, A.J. (2004) Complex SNP-related sequence variation in segmental genome duplications. *Nature Genetics*, 36(8); 861-866.
- Fu, Y., Maianu, L., Melbert, B.R. & Garvey, W.T. (2004) Facilitative glucose transporter gene expression in human lymphocytes, monocytes, and macrophages: a role for GLUT isoforms 1, 3, and 5 in the immune response and foam cell formation. *Blood Cells, Molecules & Diseases*, 32(1); 182-190.
- Gabriel, S.E. (2010) Heart disease and rheumatoid arthritis: understanding the risks. *Annals of the Rheumatic Diseases*, 69 Suppl 1i61-64.
- Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J. & Altshuler, D. (2002) The structure of haplotype blocks in the human genome. *Science*, 296(5576); 2225-2229.
- Ganguly, A., McKnight, R.A., Raychaudhuri, S., Shin, B.C., Ma, Z., Moley, K. & Devaskar, S.U. (2007) Glucose transporter isoform-3 mutations cause early pregnancy loss and fetal

- growth restriction. *American Journal of Physiology, Endocrinology and metabolism*, 292(5); E1241-55.
- Ganz, T. (1999) Defensins and host defense. *Science*, 286(5439); 420-421.
- Goemaere, S., Ackerman, C., Goethals, K., De Keyser, F., Van der Straeten, C., Verbruggen, G., Mielants, H. & Veys, E.M. (1990) Onset of symptoms of rheumatoid arthritis in relation to age, sex and menopausal transition. *The Journal of Rheumatology*, 17(12); 1620-1622.
- Goldring, M.B. & Marcu, K.B. (2009) Cartilage homeostasis in health and rheumatic diseases. *Arthritis research & therapy*, 11(3); 224.
- Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R.J., Freedman, B.I., Quinones, M.P., Bamshad, M.J., Murthy, K.K., Rovin, B.H., Bradley, W., Clark, R.A., Anderson, S.A., O'connell, R.J., Agan, B.K., Ahuja, S.S., Bologna, R., Sen, L., Dolan, M.J. & Ahuja, S.K. (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*, 307(5714); 1434-1440.
- Goronzy, J.J. & Weyand, C.M. (2009) Developments in the scientific understanding of rheumatoid arthritis. *Arthritis Research & Therapy*, 11(5); 249.
- Gregersen, P.K., Amos, C.I., Lee, A.T., Lu, Y., Remmers, E.F., Kastner, D.L., Seldin, M.F., Criswell, L.A., Plenge, R.M., Holers, V.M., Mikuls, T.R., Sokka, T., Moreland, L.W., Bridges, S.L., Jr, Xie, G., Begovich, A.B. & Siminovitch, K.A. (2009) REL, encoding a member of the NF-kappaB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. *Nature Genetics*, 41(7); 820-823.
- Gregersen, P.K. & Olsson, L.M. (2009) Recent advances in the genetics of autoimmune disease. *Annual Review of Immunology*, 27363-391.
- Gregersen, P.K., Lee, H.S., Batliwalla, F. & Begovich, A.B. (2006) PTPN22: setting thresholds for autoimmunity. *Seminars in Immunology*, 18(4); 214-223.
- Gregersen, P.K., Silver, J. & Winchester, R.J. (1987) The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis and Rheumatism*, 30(11); 1205-1213.
- Gu, W., Zhang, F. & Lupski, J.R. (2008) Mechanisms for human genomic rearrangements. *PathoGenetics*, 1(1); 4.
- Haiman, C.A., Patterson, N., Freedman, M.L., Myers, S.R., Pike, M.C., Waliszewska, A., Neubauer, J., Tandon, A., Schirmer, C., McDonald, G.J., Greenway, S.C., Stram, D.O., Le Marchand, L., Kolonel, L.N., Frasco, M., Wong, D., Pooler, L.C., Ardlie, K., Oakley-Girvan, I., Whittemore, A.S., Cooney, K.A., John, E.M., Ingles, S.A., Altshuler, D., Henderson, B.E. & Reich, D. (2007) Multiple regions within 8q24 independently affect risk for prostate cancer. *Nature Genetics*, 39(5); 638-644.
- Hastings, P.J., Ira, G. & Lupski, J.R. (2009) A microhomology-mediated break-induced replication model for the origin of human copy number variation. *Public Library of Science Genetics*, 5(1); e1000327.

- Heng, H.H., Squire, J. & Tsui, L.C. (1992) High-resolution mapping of mammalian genes by in situ hybridization to free chromatin. *Proceedings of the National Academy of Sciences of the United States of America*, 89(20); 9509-9513.
- DNA Baser Sequence Assembler v2.x (2010). Available from: <http://www.DnaBaser.com> [Accessed 2009].
- Hill, J.A., Southwood, S., Sette, A., Jevnikar, A.M., Bell, D.A. & Cairns, E. (2003) Cutting edge: the conversion of arginine to citrulline allows for a high-affinity peptide interaction with the rheumatoid arthritis-associated HLA-DRB1*0401 MHC class II molecule. *Journal of Immunology*, 171(2); 538-541.
- Hinks, A., Barton, A., John, S., Bruce, I., Hawkins, C., Griffiths, C.E., Donn, R., Thomson, W., Silman, A. & Worthington, J. (2005) Association between the PTPN22 gene and rheumatoid arthritis and juvenile idiopathic arthritis in a UK population: further support that PTPN22 is an autoimmunity gene. *Arthritis and Rheumatism*, 52(6); 1694-1699.
- Hollox, E.J., Huffmeier, U., Zeeuwen, P.L., Palla, R., Lascorz, J., Rodijk-Olthuis, D., van de Kerkhof, P.C., Traupe, H., de Jongh, G., den Heijer, M., Reis, A., Armour, J.A. & Schalkwijk, J. (2008) Psoriasis is associated with increased beta-defensin genomic copy number. *Nature Genetics*, 40(1); 23-25.
- Hollox, E.J., Armour, J.A. & Barber, J.C. (2003) Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster. *American Journal of Human Genetics*, 73(3); 591-600.
- Holmdahl, R., Lorentzen, J.C., Lu, S., Olofsson, P., Wester, L., Holmberg, J. & Pettersson, U. (2001) Arthritis induced in rats with nonimmunogenic adjuvants as models for rheumatoid arthritis. *Immunological Reviews*, 184; 184-202.
- Howell, W.M., Jobs, M., Gyllenstein, U. & Brookes, A.J. (1999) Dynamic allele-specific hybridization. A new method for scoring single nucleotide polymorphisms. *Nature Biotechnology*, 17(1); 87-88.
- Huizinga, T.W., Amos, C.I., van der Helm-van Mil, A.H., Chen, W., van Gaalen, F.A., Jawaheer, D., Schreuder, G.M., Wener, M., Breedveld, F.C., Ahmad, N., Lum, R.F., de Vries, R.R., Gregersen, P.K., Toes, R.E. & Criswell, L.A. (2005) Refining the complex rheumatoid arthritis phenotype based on specificity of the HLA-DRB1 shared epitope for antibodies to citrullinated proteins. *Arthritis and Rheumatism*, 52(11); 3433-3438.
- Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W. & Lee, C. (2004) Detection of large-scale variation in the human genome. *Nature Genetics*, 36(9); 949-951.
- Ikari, K., Momohara, S., Inoue, E., Tomatsu, T., Hara, M., Yamanaka, H. & Kamatani, N. (2006) Haplotype analysis revealed no association between the PTPN22 gene and RA in a Japanese population. *Rheumatology*, 45(11); 1345-1348.
- Illumina. Available from: www.illumina.com/support/faqs.ilmn [Accessed July 2010].
- International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, 437(7063); 1299-1320.

- International HapMap Consortium (2003) The International HapMap Project. *Nature*, 426(6968); 789-796.
- International Schizophrenia Consortium (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, 455(7210); 237-241.
- Jacobs, P.A., Matsuura, J.S., Mayer, M. & Newlands, I.M. (1978) A cytogenetic survey of an institution for the mentally retarded: I. Chromosome abnormalities. *Clinical Genetics*, 13(1); 37-60.
- Janecka, J.E., Miller, W., Pringle, T.H., Wiens, F., Zitzmann, A., Helgen, K.M., Springer, M.S. & Murphy, W.J. (2007) Molecular and genomic data identify the closest living relative of primates. *Science*, 318(5851); 792-794.
- Jansson, A.M., Jacobsson, L., Luthman, H. & Lorentzen, J.C. (1999) Susceptibility to oil-induced arthritis is linked to Oia2 on chromosome 4 in a DA(DA x PVG.1AV1) backcross. *Transplantation Proceedings*, 31(3); 1597-1599.
- Jarvis, J.N., Dozmorov, I., Jiang, K., Frank, M.B., Szodoray, P., Alex, P. & Centola, M. (2004) Novel approaches to gene expression analysis of active polyarticular juvenile rheumatoid arthritis. *Arthritis Research & Therapy*, 6(1); R15-R32.
- Jawaheer, D., Seldin, M.F., Amos, C.I., Chen, W.V., Shigeta, R., Etzel, C., Damle, A., Xiao, X., Chen, D., Lum, R.F., Monteiro, J., Kern, M., Criswell, L.A., Albani, S., Nelson, J.L., Clegg, D.O., Pope, R., Schroeder, H.W., Jr, Bridges, S.L., Jr, Pisetsky, D.S., Ward, R., Kastner, D.L., Wilder, R.L., Pincus, T., Callahan, L.F., Flemming, D., Wener, M.H., Gregersen, P.K. & North American Rheumatoid Arthritis Consortium (2003) Screening the genome for rheumatoid arthritis susceptibility genes: a replication study and combined analysis of 512 multicase families. *Arthritis and Rheumatism*, 48(4); 906-916.
- Jawaheer, D., Seldin, M.F., Amos, C.I., Chen, W.V., Shigeta, R., Monteiro, J., Kern, M., Criswell, L.A., Albani, S., Nelson, J.L., Clegg, D.O., Pope, R., Schroeder, H.W., Jr, Bridges, S.L., Jr, Pisetsky, D.S., Ward, R., Kastner, D.L., Wilder, R.L., Pincus, T., Callahan, L.F., Flemming, D., Wener, M.H. & Gregersen, P.K. (2001) A genomewide screen in multiplex rheumatoid arthritis families suggests genetic overlap with other autoimmune diseases. *American Journal of Human Genetics*, 68(4); 927-936.
- Jeffreys, A.J., Wilson, V. & Thein, S.L. (1985) Hypervariable 'minisatellite' regions in human DNA. *Nature*, 314(6006); 67-73.
- John, S., Amos, C., Shephard, N., Chen, W., Butterworth, A., Etzel, C., Jawaheer, D., Seldin, M., Silman, A., Gregersen, P. & Worthington, J. (2006) Linkage analysis of rheumatoid arthritis in US and UK families reveals interactions between HLA-DRB1 and loci on chromosomes 6q and 16p. *Arthritis and Rheumatism*, 54(5); 1482-1490.
- 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(5083); 818-821.
- Kan, Y.W., Golbus, M.S. & Dozy, A.M. (1976) Prenatal diagnosis of alpha-thalassemia. Clinical application of molecular hybridization. *The New England Journal of Medicine*, 295(21); 1165-1167.

- Kastbom, A., Strandberg, G., Lindroos, A. & Skogh, T. (2004) Anti-CCP antibody test predicts the disease course during 3 years in early rheumatoid arthritis (the Swedish TIRA project). *Annals of the Rheumatic Diseases*, 63(9); 1085-1089.
- Katz, D.A. & Bhathena, A. (2009) Overview of pharmacogenetics. *Current Protocols in Human Genetics*, Chapter 9; Unit 9.19.
- Kawasaki, E., Awata, T., Ikegami, H., Kobayashi, T., Maruyama, T., Nakanishi, K., Shimada, A., Uga, M., Kurihara, S., Kawabata, Y., Tanaka, S., Kanazawa, Y., Lee, I. & Eguchi, K. (2006) Systematic search for single nucleotide polymorphisms in a lymphoid tyrosine phosphatase gene (PTPN22): association between a promoter polymorphism and type 1 diabetes in Asian populations. *American Journal of Medical Genetics*, 140(6); 586-593.
- Kent, W.J. (2002) BLAT--the BLAST-like alignment tool. *Genome Research*, 12(4); 656-664.
- Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. & Haussler, D. (2002) The human genome browser at UCSC. *Genome Research*, 12(6); 996-1006.
- Kerem, B., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M. & Tsui, L.C. (1989) Identification of the cystic fibrosis gene: genetic analysis. *Science*, 245(4922); 1073-1080.
- Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N.A., Tsang, P., Newman, T.L., Tuzun, E., Cheng, Z., Ebling, H.M., Tusneem, N., David, R., Gillett, W., Phelps, K.A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., McKernan, K., Chen, L., Malig, M., Smith, J.D., Korn, J.M., McCarroll, S.A., Altshuler, D.A., Peiffer, D.A., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D.A., Mullikin, J.C., Wilson, R.K., Bruhn, L., Olson, M.V., Kaul, R., Smith, D.R. & Eichler, E.E. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191); 56-64.
- Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature*, 217(5129); 624-626.
- Klareskog, L., Stolt, P., Lundberg, K., Kallberg, H., Bengtsson, C., Grunewald, J., Ronnelid, J., Harris, H.E., Ulfgren, A.K., Rantapaa-Dahlqvist, S., Eklund, A., Padyukov, L. & Alfredsson, L. (2006a) A new model for an etiology of rheumatoid arthritis: smoking may trigger HLA-DR (shared epitope)-restricted immune reactions to autoantigens modified by citrullination. *Arthritis and Rheumatism*, 54(1); 38-46.
- Klareskog, L., Padyukov, L., Lorentzen, J. & Alfredsson, L. (2006b) Mechanisms of disease: Genetic susceptibility and environmental triggers in the development of rheumatoid arthritis. *Nature Clinical Practice Rheumatology*, 2(8); 425-433.
- Kool, M., Koster, J., Bunt, J., Hasselt, N.E., Lakeman, A., van Sluis, P., Troost, D., Meeteren, N.S., Caron, H.N., Cloos, J., Mrcic, A., Ylstra, B., Grajkowska, W., Hartmann, W., Pietsch, T., Ellison, D., Clifford, S.C. & Versteeg, R. (2008) Integrated genomics identifies five medulloblastoma subtypes with distinct genetic profiles, pathway signatures and clinicopathological features. *Public Library of Science One*, 3(8); e3088.
- Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., Taillon, B.E., Chen, Z., Tanzer, A., Saunders, A.C., Chi, J., Yang, F., Carter, N.P., Hurles, M.E., Weissman, S.M., Harkins, T.T., Gerstein,

- M.B., Egholm, M. & Snyder, M. (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318(5849); 420-426.
- Krumsiek, J., Arnold, R. & Rattei, T. (2007) Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*, 23(8); 1026-1028.
- LaFave, M.C. & Sekelsky, J. (2009) Mitotic recombination: Why? When? How? Where?. *Public Library of Science Genetics*, 5(3); e1000411.
- Lam, K.W. & Jeffreys, A.J. (2006) Processes of copy-number change in human DNA: the dynamics of {alpha}-globin gene deletion. *Proceedings of the National Academy of Sciences of the United States of America*, 103(24); 8921-8927.
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczký, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglu, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kasprzyk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J. & International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, 409(6822); 860-921.

- Lander, E.S. (1996) The new genomics: global views of biology. *Science*, 274(5287); 536-539.
- Lebenthal, E. (1987) Role of salivary amylase in gastric and intestinal digestion of starch. *Digestive Diseases and Sciences*, 32(10); 1155-1157.
- Lee, H.S., Korman, B.D., Le, J.M., Kastner, D.L., Remmers, E.F., Gregersen, P.K. & Bae, S.C. (2009) Genetic risk factors for rheumatoid arthritis differ in Caucasian and Korean populations. *Arthritis and Rheumatism*, 60(2); 364-371.
- Lee, J.A., Carvalho, C.M. & Lupski, J.R. (2007a) A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell*, 131(7); 1235-1247.
- Lee, Y.H., Rho, Y.H., Choi, S.J., Ji, J.D., Song, G.G., Nath, S.K. & Harley, J.B. (2007b) The PTPN22 C1858T functional polymorphism and autoimmune diseases--a meta-analysis. *Rheumatology*, 46(1); 49-56.
- Lee, H.S., Remmers, E.F., Le, J.M., Kastner, D.L., Bae, S.C. & Gregersen, P.K. (2007c) Association of STAT4 with rheumatoid arthritis in the Korean population. *Molecular Medicine*, 13(9-10); 455-460.
- Lei, C., Dongqing, Z., Yeqing, S., Oaks, M.K., Lishan, C., Jianzhong, J., Jie, Q., Fang, D., Ningli, L., Xinghai, H. & Daming, R. (2005) Association of the CTLA-4 gene with rheumatoid arthritis in Chinese Han population. *European Journal of Human Genetics*, 13(7); 823-828.
- Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., Lin, Y., MacDonald, J.R., Pang, A.W., Shago, M., Stockwell, T.B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S.A., Busam, D.A., Beeson, K.Y., McIntosh, T.C., Remington, K.A., Abril, J.F., Gill, J., Borman, J., Rogers, Y.H., Frazier, M.E., Scherer, S.W., Strausberg, R.L. & Venter, J.C. (2007) The diploid genome sequence of an individual human. *Public Library of Science Biology*, 5(10); e254.
- Li, Y. & Begovich, A.B. (2009) Unraveling the genetics of complex diseases: susceptibility genes for rheumatoid arthritis and psoriasis. *Seminars in Immunology*, 21(6); 318-327.
- Listing, J., Strangfeld, A., Kekow, J., Schneider, M., Kapelle, A., Wassenberg, S. & Zink, A. (2008) Does tumor necrosis factor alpha inhibition promote or prevent heart failure in patients with rheumatoid arthritis? *Arthritis and Rheumatism*, 58(3); 667-677.
- Locke, D.P., Sharp, A.J., McCarroll, S.A., McGrath, S.D., Newman, T.L., Cheng, Z., Schwartz, S., Albertson, D.G., Pinkel, D., Altshuler, D.M. & Eichler, E.E. (2006) Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *American Journal of Human Genetics*, 79(2); 275-290.
- Lorentzen, J.C., Flornes, L., Eklow, C., Backdahl, L., Ribbhammar, U., Guo, J.P., Smolnikova, M., Dissen, E., Seddighzadeh, M., Brookes, A.J., Alfredsson, L., Klareskog, L., Padyukov, L. & Fossum, S. (2007) Association of arthritis with a gene complex encoding C-type lectin-like receptors. *Arthritis and Rheumatism*, 56(8); 2620-2632.
- Lorentzen, J.C., Glaser, A., Jacobsson, L., Galli, J., Fakhrai-rad, H., Klareskog, L. & Luthman, H. (1998) Identification of rat susceptibility loci for adjuvant-oil-induced arthritis. *Proceedings of the National Academy of Sciences of the United States of America*, 95(11); 6383-6387.

- 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., Chakravarti, A. & Patel, P.I. (1991) DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell*, 66(2); 219-232.
- MacGregor, A.J., Snieder, H., Rigby, A.S., Koskenvuo, M., Kaprio, J., Aho, K. & Silman, A.J. (2000) Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis and Rheumatism*, 43(1); 30-37.
- Macheda, M.L., Rogers, S. & Best, J.D. (2005) Molecular and cellular regulation of glucose transporter (GLUT) proteins in cancer. *Journal of Cellular Physiology*, 202(3); 654-662.
- MacKay, K., Eyre, S., Myerscough, A., Milicic, A., Barton, A., Laval, S., Barrett, J., Lee, D., White, S., John, S., Brown, M.A., Bell, J., Silman, A., Ollier, W., Wordsworth, P. & Worthington, J. (2002) Whole-genome linkage analysis of rheumatoid arthritis susceptibility loci in 252 affected sibling pairs in the United Kingdom. *Arthritis and Rheumatism*, 46(3); 632-639.
- Maher, B. (2008) Personal genomes: The case of the missing heritability. *Nature*, 456(7218); 18-21.
- Mallon, E., Newson, R. & Bunker, C.B. (1999) HLA-Cw6 and the genetic predisposition to psoriasis: a meta-analysis of published serologic studies. *The Journal of Investigative Dermatology*, 113(4); 693-695.
- Maratou, E., Dimitriadis, G., Kollias, A., Boutati, E., Lambadiari, V., Mitrou, P. & Raptis, S.A. (2007) Glucose transporter expression on the plasma membrane of resting and activated white blood cells. *European Journal of Clinical Investigation*, 37(4); 282-290.
- Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F. & Rothberg, J.M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057); 376-380.
- Martinez-Gamboa, L., Brezinschek, H.P., Burmester, G.R. & Dorner, T. (2006) Immunopathologic role of B lymphocytes in rheumatoid arthritis: rationale of B cell-directed therapy. *Autoimmunity Reviews*, 5(7); 437-442.
- McCarroll, S.A., Kuruvilla, F.G., Korn, J.M., Cawley, S., Nemesh, J., Wysoker, A., Shaperro, M.H., de Bakker, P.I., Maller, J.B., Kirby, A., Elliott, A.L., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P.J., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, K.W., Rava, R., Daly, M.J., Gabriel, S.B. & Altshuler, D. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics*, 40(10); 1166-1174.
- McCarthy, M.I. (2009) Exploring the unknown: assumptions about allelic architecture and strategies for susceptibility variant discovery, *Genome Medicine*, 1(7); 66.

- McCarthy, M.I. & Hirschhorn, J.N. (2008) Genome-wide association studies: potential next steps on a genetic journey. *Human Molecular Genetics*, 17(R2); R156-65.
- McKinney, C., Merriman, M.E., Chapman, P.T., Gow, P.J., Harrison, A.A., Highton, J., Jones, P.B., McLean, L., O'Donnell, J.L., Pokorny, V., Spellerberg, M., Stamp, L.K., Willis, J., Steer, S. & Merriman, T.R. (2008) Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis. *Annals of the Rheumatic Diseases*, 67(3); 409-413.
- Medical Look, *Illustration of Rheumatoid Arthritis*. Available from: http://www.medical-look.com/Joint_pain/Rheumatoid_arthritis.html [Accessed August 2010] .
- Meyne, J. & Goodwin, E.H. (1994) Strand-specific fluorescence in situ hybridization for determining orientation and direction of DNA sequences. *Methods in Molecular Biology*, 33141-145.
- Michou, L., Lasbleiz, S., Rat, A.C., Migliorini, P., Balsa, A., Westhovens, R., Barrera, P., Alves, H., Pierlot, C., Glikmans, E., Garnier, S., Dausset, J., Vaz, C., Fernandes, M., Petit-Teixeira, E., Lemaire, I., Pascual-Salcedo, D., Bombardieri, S., Dequeker, J., Radstake, T.R., Van Riel, P., van de Putte, L., Lopes-Vaz, A., Prum, B., Bardin, T., Dieude, P., Cornelis, F. & European Consortium on Rheumatoid Arthritis Families (2007) Linkage proof for PTPN22, a rheumatoid arthritis susceptibility gene and a human autoimmunity gene. *Proceedings of the National Academy of Sciences of the United States of America*, 104(5); 1649-1654.
- Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M. & Yamanaka, S. (2003) The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell*, 113(5); 631-642.
- Mobasheri, A., Bondy, C.A., Moley, K., Mendes, A.F., Rosa, S.C., Richardson, S.M., Hoyland, J.A., Barrett-Jolley, R. & Shakibaei, M. (2008) Facilitative glucose transporters in articular chondrocytes. Expression, distribution and functional regulation of GLUT isoforms by hypoxia, hypoxia mimetics, growth factors and pro-inflammatory cytokines. *Advances in Anatomy, Embryology, and Cell Biology*, 2001 p following vi, 1-84.
- Morgan, A.W., Thomson, W., Martin, S.G., Yorkshire Early Arthritis Register Consortium, Carter, A.M., UK Rheumatoid Arthritis Genetics Consortium, Erlich, H.A., Barton, A., Hocking, L., Reid, D.M., Harrison, P., Wordsworth, P., Steer, S., Worthington, J., Emery, P., Wilson, A.G. & Barrett, J.H. (2009) Reevaluation of the interaction between HLA-DRB1 shared epitope alleles, PTPN22, and smoking in determining susceptibility to autoantibody-positive and autoantibody-negative rheumatoid arthritis in a large UK Caucasian population. *Arthritis and Rheumatism*, 60(9); 2565-2576.
- Mullighan, C.G., Goorha, S., Radtke, I., Miller, C.B., Coustan-Smith, E., Dalton, J.D., Girtman, K., Mathew, S., Ma, J., Pounds, S.B., Su, X., Pui, C.H., Relling, M.V., Evans, W.E., Shurtleff, S.A. & Downing, J.R. (2007) Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature*, 446(7137); 758-764.
- Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. & Erlich, H. (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, 51 Pt 1263-273.

- Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. (2005) A fine-scale map of recombination rates and hotspots across the human genome, *Science*, 310(5746); 321-324.
- Nathans, J., Piantanida, T.P., Eddy, R.L., Shows, T.B. & Hogness, D.S. (1986) Molecular genetics of inherited variation in human color vision. *Science*, 232(4747); 203-210.
- National Rheumatoid Arthritis Society 2010, *The Economic Burden of Rheumatoid Arthritis*.
- Nguyen, D.Q., Webber, C. & Ponting, C.P. (2006) Bias of selection on human copy-number variants. *Public Library of Science Genetics*, 2(2); e20.
- Nielen, M.M., van Schaardenburg, D., Reesink, H.W., van de Stadt, R.J., van der Horst-Bruinsma, I.E., de Koning, M.H., Habibuw, M.R., Vandenbroucke, J.P. & Dijkmans, B.A. (2004) Specific autoantibodies precede the symptoms of rheumatoid arthritis: a study of serial measurements in blood donors. *Arthritis and Rheumatism*, 50(2); 380-386.
- Orozco, G., Eyre, S., Hinks, A., Ke, X., Wilson, A.G., Bax, D.E., Morgan, A.W., Emery, P., Steer, S., Hocking, L., Reid, D.M., Wordsworth, P., Harrison, P., Thomson, W., Barton, A. & Worthington, J. (2009) Association of CD40 with rheumatoid arthritis confirmed in a large UK case-control study. *Annals of the Rheumatic Diseases*, .
- Otero, M. & Goldring, M.B. (2007) Cells of the synovium in rheumatoid arthritis. Chondrocytes. *Arthritis Research & Therapy*, 9(5); 220.
- Pentao, L., Wise, C.A., Chinault, A.C., Patel, P.I. & Lupski, J.R. (1992) Charcot-Marie-Tooth type 1A duplication appears to arise from recombination at repeat sequences flanking the 1.5 Mb monomer unit. *Nature Genetics*, 2(4); 292-300.
- Perry, G.H., Ben-Dor, A., Tsalenko, A., Sampas, N., Rodriguez-Revenga, L., Tran, C.W., Scheffer, A., Steinfeld, I., Tsang, P., Yamada, N.A., Park, H.S., Kim, J.I., Seo, J.S., Yakhini, Z., Laderman, S., Bruhn, L. & Lee, C. (2008) The fine-scale and complex architecture of human copy-number variation. *American Journal of Human Genetics*, 82(3); 685-695.
- Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., Werner, J., Villanea, F.A., Mountain, J.L., Misra, R., Carter, N.P., Lee, C. & Stone, A.C. (2007) Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*, 39(10); 1256-1260.
- Petrone, J. (6 April 2010), 'Disappointing' WTCCC Study Results Could Hasten Development of New CNV Chips, Authors Say. *BioArray News*.
- Pinard, R., de Winter, A., Sarkis, G.J., Gerstein, M.B., Tartaro, K.R., Plant, R.N., Egholm, M., Rothberg, J.M. & Leamon, J.H. (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics*, 7216.
- 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. *Nature Genetics*, 20(2); 207-211.

- Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S., Almeida, J., Bacchelli, E., Bader, G.D., Bailey, A.J., Baird, G., Battaglia, A., Berney, T., Bolshakova, N., Bolte, S., Bolton, P.F., Bourgeron, T., Brennan, S., Brian, J., Bryson, S.E., Carson, A.R., Casallo, G., Casey, J., Chung, B.H., Cochrane, L., Corsello, C., Crawford, E.L., Crossett, A., Cytrynbaum, C., Dawson, G., de Jonge, M., Delorme, R., Drmic, I., Duketis, E., Duque, F., Estes, A., Farrar, P., Fernandez, B.A., Folstein, S.E., Fombonne, E., Freitag, C.M., Gilbert, J., Gillberg, C., Glessner, J.T., Goldberg, J., Green, A., Green, J., Guter, S.J., Hakonarson, H., Heron, E.A., Hill, M., Holt, R., Howe, J.L., Hughes, G., Hus, V., Iglizoi, R., Kim, C., Klauck, S.M., Klevzon, A., Korvatska, O., Kustanovich, V., Lajonchere, C.M., Lamb, J.A., Laskawiec, M., Leboyer, M., Le Couteur, A., Leventhal, B.L., Lionel, A.C., Liu, X.Q., Lord, C., Lotspeich, L., Lund, S.C., Maestrini, E., Mahoney, W., Mantoulan, C., Marshall, C.R., McConachie, H., McDougle, C.J., McGrath, J., McMahon, W.M., Merikangas, A., Migita, O., Minshew, N.J., Mirza, G.K., Munson, J., Nelson, S.F., Noakes, C., Noor, A., Nygren, G., Oliveira, G., Papanikolaou, K., Parr, J.R., Parrini, B., Paton, T., Pickles, A., Pilorge, M., Piven, J., Ponting, C.P., Posey, D.J., Poustka, A., Poustka, F., Prasad, A., Ragoussis, J., Renshaw, K., Rickaby, J., Roberts, W., Roeder, K., Roge, B., Rutter, M.L., Bierut, L.J., Rice, J.P., Salt, J., Sansom, K., Sato, D., Segurado, R., Sequeira, A.F., Senman, L., Shah, N., Sheffield, V.C., Soorya, L., Sousa, I., Stein, O., Sykes, N., Stoppioni, V., Strawbridge, C., Tancredi, R., Tansey, K., Thiruvahindrapduram, B., Thompson, A.P., Thomson, S., Tryfon, A., Tsiantis, J., Van Engeland, H., Vincent, J.B., Volkmar, F., Wallace, S., Wang, K., Wang, Z., Wassink, T.H., Webber, C., Weksberg, R., Wing, K., Wittemeyer, K., Wood, S., Wu, J., Yaspan, B.L., Zurawiecki, D., Zwaigenbaum, L., Buxbaum, J.D., Cantor, R.M., Cook, E.H., Coon, H., Cuccaro, M.L., Devlin, B., Ennis, S., Gallagher, L., Geschwind, D.H., Gill, M., Haines, J.L., Hallmayer, J., Miller, J., Monaco, A.P., Nurnberger, J.I., Jr, Paterson, A.D., Pericak-Vance, M.A., Schellenberg, G.D., Szatmari, P., Vicente, A.M., Vieland, V.J., Wijsman, E.M., Scherer, S.W., Sutcliffe, J.S. & Betancur, C. (2010) Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, 466(7304); 368-372.
- Piotrowski, A., Bruder, C.E., Andersson, R., de Stahl, T.D., Menzel, U., Sandgren, J., Poplawski, A., von Tell, D., Crasto, C., Bogdan, A., Bartoszewski, R., Bebok, Z., Krzyzanowski, M., Jankowski, Z., Partridge, E.C., Komorowski, J. & Dumanski, J.P. (2008) Somatic mosaicism for copy number variation in differentiated human tissues. *Human Mutation*, 29(9); 1118-1124.
- Plenge, R.M., Seielstad, M., Padyukov, L., Lee, A.T., Remmers, E.F., Ding, B., Liew, A., Khalili, H., Chandrasekaran, A., Davies, L.R., Li, W., Tan, A.K., Bonnard, C., Ong, R.T., Thalamuthu, A., Pettersson, S., Liu, C., Tian, C., Chen, W.V., Carulli, J.P., Beckman, E.M., Altshuler, D., Alfredsson, L., Criswell, L.A., Amos, C.I., Seldin, M.F., Kastner, D.L., Klareskog, L. & Gregersen, P.K. (2007) TRAF1-C5 as a risk locus for rheumatoid arthritis--a genome-wide study. *The New England Journal of Medicine*, 357(12); 1199-1209.
- Qiao, Y., Liu, X., Harvard, C., Nolin, S.L., Brown, W.T., Koochek, M., Holden, J.J., Lewis, M.E. & Rajcan-Separovic, E. (2007) Large-scale copy number variants (CNVs): distribution in normal subjects and FISH/real-time qPCR analysis. *BMC Genomics*, 8167.
- Rantapaa-Dahlqvist, S., de Jong, B.A., Berglin, E., Hallmans, G., Wadell, G., Stenlund, H., Sundin, U. & van Venrooij, W.J. (2003) Antibodies against cyclic citrullinated peptide and IgA rheumatoid factor predict the development of rheumatoid arthritis. *Arthritis and Rheumatism*, 48(10); 2741-2749.

- Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., Cho, E.K., Dallaire, S., Freeman, J.L., Gonzalez, J.R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J.R., Marshall, C.R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M.J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D.F., Estivill, X., Tyler-Smith, C., Carter, N.P., Aburatani, H., Lee, C., Jones, K.W., Scherer, S.W. & Hurles, M.E. (2006) Global variation in copy number in the human genome. *Nature*, 444(7118); 444-454.
- Reid, A.G., Tarpey, P.S. & Nacheva, E.P. (2003) High-resolution analysis of acquired genomic imbalances in bone marrow samples from chronic myeloid leukemia patients by use of multiple short DNA probes. *Genes, Chromosomes & Cancer*, 37(3); 282-290.
- Reiter, L.T., Hastings, P.J., Nelis, E., De Jonghe, P., Van Broeckhoven, C. & Lupski, J.R. (1998) Human meiotic recombination products revealed by sequencing a hotspot for homologous strand exchange in multiple HNPP deletion patients. *American Journal of Human Genetics*, 62(5); 1023-1033.
- Remmers, E.F., Plenge, R.M., Lee, A.T., Graham, R.R., Hom, G., Behrens, T.W., de Bakker, P.I., Le, J.M., Lee, H.S., Batliwalla, F., Li, W., Masters, S.L., Booty, M.G., Carulli, J.P., Padyukov, L., Alfredsson, L., Klareskog, L., Chen, W.V., Amos, C.I., Criswell, L.A., Seldin, M.F., Kastner, D.L. & Gregersen, P.K. (2007) STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. *The New England Journal of Medicine*, 357(10); 977-986.
- Renaudineau, Y., Jamin, C., Saraux, A. & Youinou, P. (2005) Rheumatoid factor on a daily basis. *Autoimmunity*, 38(1); 11-16.
- Ribbhammar, U., Flornes, L., Backdahl, L., Luthman, H., Fossum, S. & Lorentzen, J.C. (2003) High resolution mapping of an arthritis susceptibility locus on rat chromosome 4, and characterization of regulated phenotypes. *Human Molecular Genetics*, 12(17); 2087-2096.
- Rice, T.K., Schork, N.J. & Rao, D.C. (2008) Methods for handling multiple testing. *Advances in Genetics*, 60293-308.
- Rieck, M., Arechiga, A., Onengut-Gumuscu, S., Greenbaum, C., Concannon, P. & Buckner, J.H. (2007) Genetic variation in PTPN22 corresponds to altered function of T and B lymphocytes. *Journal of immunology*, 179(7); 4704-4710.
- Risch, N. & Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science*, 273(5281); 1516-1517.
- Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A. & Arnheim, N. (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 230(4732); 1350-1354.
- Sambrook, J. & Russell, D. (2001), *Molecular Cloning: A Laboratory Manual*, 3rd Edition, Cold Spring Harbour Laboratory Press.
- Schellekens, G.A., de Jong, B.A., van den Hoogen, F.H., van de Putte, L.B. & van Venrooij, W.J. (1998) Citrulline is an essential constituent of antigenic determinants recognized by rheumatoid arthritis-specific autoantibodies. *The Journal of Clinical Investigation*, 101(1); 273-281.

- Schmidt, S., Hommel, A., Gawlik, V., Augustin, R., Junicke, N., Florian, S., Richter, M., Walther, D.J., Montag, D., Joost, H.G. & Schurmann, A. (2009) Essential role of glucose transporter GLUT3 for post-implantation embryonic development. *The Journal of Endocrinology*, 200(1); 23-33.
- Schouten, J.P., McElgunn, C.J., Waaijer, R., Zwijnenburg, D., Diepvens, F. & Pals, G. (2002) Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Research*, 30(12); e57.
- Schwartz, D.C. & Cantor, C.R. (1984) Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell*, 37(1); 67-75.
- Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., Leotta, A., Pai, D., Zhang, R., Lee, Y.H., Hicks, J., Spence, S.J., Lee, A.T., Puura, K., Lehtimäki, T., Ledbetter, D., Gregersen, P.K., Bregman, J., Sutcliffe, J.S., Jobanputra, V., Chung, W., Warburton, D., King, M.C., Skuse, D., Geschwind, D.H., Gilliam, T.C., Ye, K. & Wigler, M. (2007) Strong association of de novo copy number mutations with autism. *Science*, 316(5823); 445-449.
- Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T.C., Trask, B., Patterson, N., Zetterberg, A. & Wigler, M. (2004) Large-scale copy number polymorphism in the human genome. *Science*, 305(5683); 525-528.
- Sharp, A.J., Locke, D.P., McGrath, S.D., Cheng, Z., Bailey, J.A., Vallente, R.U., Pertz, L.M., Clark, R.A., Schwartz, S., Segraves, R., Oseroff, V.V., Albertson, D.G., Pinkel, D. & Eichler, E.E. (2005) Segmental duplications and copy-number variation in the human genome. *American Journal of Human Genetics*, 77(1); 78-88.
- Shiozawa, S., Hayashi, S., Tsukamoto, Y., Goko, H., Kawasaki, H., Wada, T., Shimizu, K., Yasuda, N., Kamatani, N., Takasugi, K., Tanaka, Y., Shiozawa, K. & Imura, S. (1998) Identification of the gene loci that predispose to rheumatoid arthritis. *International Immunology*, 10(12); 1891-1895.
- Siggberg, L., Ala-Mello, S., Jaakkola, E., Kuusinen, E., Schuit, R., Kohlhase, J., Böhm, D., Ignatius, J. & Knuutila, S. (2010) Array CGH in molecular diagnosis of mental retardation - A study of 150 Finnish patients. *American Journal of Medical Genetics*, 152A(6); 1398-1410.
- Silman, A.J., Newman, J. & MacGregor, A.J. (1996) Cigarette smoking increases the risk of rheumatoid arthritis. Results from a nationwide study of disease-discordant twins. *Arthritis and Rheumatism*, 39(5); 732-735.
- Silman, A., Kay, A. & Brennan, P. (1992) Timing of pregnancy in relation to the onset of rheumatoid arthritis. *Arthritis and Rheumatism*, 35(2); 152-155.
- Silver, I. (1975) Measurement of pH and ionic composition of pericellular sites. *Philosophical Transactions of the Royal Society*, 271; 261-272.
- Singh-Gasson, S., Green, R.D., Yue, Y., Nelson, C., Blattner, F., Sussman, M.R. & Cerrina, F. (1999) Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nature Biotechnology*, 17(10); 974-978.

- Slavik, J.M., Hutchcroft, J.E. & Bierer, B.E. (1999) CD28/CTLA-4 and CD80/CD86 families: signaling and function. *Immunologic Research*, 19(1); 1-24.
- Slijepcevic, P. (2001) Telomere length measurement by Q-FISH. *Methods in Cell Science*, 23(1-3); 17-22.
- Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A.K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J.P., Gray, J.W., Jain, A.N., Pinkel, D. & Albertson, D.G. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics*, 29(3); 263-264.
- Southern, E.M. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology*, 98(3); 503-517.
- Stastny, P. (1976) Mixed lymphocyte cultures in rheumatoid arthritis. *The Journal of Clinical Investigation*, 57(5); 1148-1157.
- Suzuki, A., Yamada, R., Chang, X., Tokuhira, S., Sawada, T., Suzuki, M., Nagasaki, M., Nakayama-Hamada, M., Kawaida, R., Ono, M., Ohtsuki, M., Furukawa, H., Yoshino, S., Yukioka, M., Tohma, S., Matsubara, T., Wakitani, S., Teshima, R., Nishioka, Y., Sekine, A., Iida, A., Takahashi, A., Tsunoda, T., Nakamura, Y. & Yamamoto, K. (2003) Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. *Nature Genetics*, 34(4); 395-402.
- Symmons, D., Turner, G., Webb, R., Asten, P., Barrett, E., Lunt, M., Scott, D. & Silman, A. (2002) The prevalence of rheumatoid arthritis in the United Kingdom: new estimates for a new century. *Rheumatology*, 41(7); 793-800.
- Symmons, D.P., Bankhead, C.R., Harrison, B.J., Brennan, P., Barrett, E.M., Scott, D.G. & Silman, A.J. (1997) Blood transfusion, smoking, and obesity as risk factors for the development of rheumatoid arthritis: results from a primary care-based incident case-control study in Norfolk, England. *Arthritis and Rheumatism*, 40(11); 1955-1961.
- Takata, Y., Inoue, H., Sato, A., Tsugawa, K., Miyatake, K., Hamada, D., Shinomiya, F., Nakano, S., Yasui, N., Tanahashi, T. & Itakura, M. (2008) Replication of reported genetic associations of PADI4, FCRL3, SLC22A4 and RUNX1 genes with rheumatoid arthritis: results of an independent Japanese population and evidence from meta-analysis of East Asian studies. *Journal of Human Genetics*, 53(2); 163-173.
- The Huntington's Disease Collaborative Research Group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, 72(6); 971-983.
- Thomson, W., Barton, A., Ke, X., Eyre, S., Hinks, A., Bowes, J., Donn, R., Symmons, D., Hider, S., Bruce, I.N., Wellcome Trust Case Control Consortium, Wilson, A.G., Marinou, I., Morgan, A., Emery, P., YEAR Consortium, Carter, A., Steer, S., Hocking, L., Reid, D.M., Wordsworth, P., Harrison, P., Strachan, D. & Worthington, J. (2007) Rheumatoid arthritis association at 6q23. *Nature Genetics*, 39(12); 1431-1433.
- Turesson, C., O'Fallon, W.M., Crowson, C.S., Gabriel, S.E. & Matteson, E.L. (2003) Extra-articular disease manifestations in rheumatoid arthritis: incidence trends and risk factors over 46 years. *Annals of the Rheumatic Diseases*, 62(8); 722-727.

- Turesson, C., Jacobsson, L. & Bergstrom, U. (1999) Extra-articular rheumatoid arthritis: prevalence and mortality. *Rheumatology*, 38(7); 668-674.
- Turnbull, C., Ahmed, S., Morrison, J., Pernet, D., Renwick, A., Maranian, M., Seal, S., Ghoussaini, M., Hines, S., Healey, C.S., Hughes, D., Warren-Perry, M., Tapper, W., Eccles, D., Evans, D.G., Breast Cancer Susceptibility Collaboration (UK), Hooning, M., Schutte, M., van den Ouweland, A., Houlston, R., Ross, G., Langford, C., Pharoah, P.D., Stratton, M.R., Dunning, A.M., Rahman, N. & Easton, D.F. (2010) Genome-wide association study identifies five new breast cancer susceptibility loci. *Nature Genetics*, 42(6); 504-507.
- Turner, D.J., Miretti, M., Rajan, D., Fiegler, H., Carter, N.P., Blayney, M.L., Beck, S. & Hurles, M.E. (2008) Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nature Genetics*, 40(1); 90-95.
- Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J.A., Costa, G., McKernan, K., Sidow, A., Fire, A. & Johnson, S.M. (2008) A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Research*, 18(7); 1051-1063.
- Vang, T., Congia, M., Macis, M.D., Musumeci, L., Orru, V., Zavattari, P., Nika, K., Tautz, L., Tasken, K., Cucca, F., Mustelin, T. & Bottini, N. (2005) Autoimmune-associated lymphoid tyrosine phosphatase is a gain-of-function variant. *Nature Genetics*, 37(12); 1317-1319.
- Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratt, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R.,

- Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. & Zhu, X. (2001) The sequence of the human genome, *Science (New York, N.Y.)*, 291(5507); 1304-1351.
- Waalder, E. (1940) On the occurrence of a factor in human serum activating the specific agglutination of sheep blood corpuscles. *Acta Pathologica, Microbiologica, et Immunologica Scandinavica*, 115(5); 422-38
- Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., Guo, Y., Feng, B., Li, H., Lu, Y., Fang, X., Liang, H., Du, Z., Li, D., Zhao, Y., Hu, Y., Yang, Z., Zheng, H., Hellmann, I., Inouye, M., Pool, J., Yi, X., Zhao, J., Duan, J., Zhou, Y., Qin, J., Ma, L., Li, G., Yang, Z., Zhang, G., Yang, B., Yu, C., Liang, F., Li, W., Li, S., Li, D., Ni, P., Ruan, J., Li, Q., Zhu, H., Liu, D., Lu, Z., Li, N., Guo, G., Zhang, J., Ye, J., Fang, L., Hao, Q., Chen, Q., Liang, Y., Su, Y., San, A., Ping, C., Yang, S., Chen, F., Li, L., Zhou, K., Zheng, H., Ren, Y., Yang, L., Gao, Y., Yang, G., Li, Z., Feng, X., Kristiansen, K., Wong, G.K., Nielsen, R., Durbin, R., Bolund, L., Zhang, X., Li, S., Yang, H. & Wang, J. (2008) The diploid genome sequence of an Asian individual. *Nature*, 456(7218); 60-65.
- Wei, Z., Sun, W., Wang, K. & Hakonarson, H. (2009) Multiple testing in genome-wide association studies via hidden Markov models. *Bioinformatics*, 25(21); 2802-2808.
- Weir, B.A., Woo, M.S., Getz, G., Perner, S., Ding, L., Beroukhi, R., Lin, W.M., Province, M.A., Kraja, A., Johnson, L.A., Shah, K., Sato, M., Thomas, R.K., Barletta, J.A., Borecki, I.B., Broderick, S., Chang, A.C., Chiang, D.Y., Chirieac, L.R., Cho, J., Fujii, Y., Gazdar, A.F., Giordano, T., Greulich, H., Hanna, M., Johnson, B.E., Kris, M.G., Lash, A., Lin, L., Lindeman, N., Mardis, E.R., McPherson, J.D., Minna, J.D., Morgan, M.B., Nadel, M., Orringer, M.B., Osborne, J.R., Ozenberger, B., Ramos, A.H., Robinson, J., Roth, J.A., Rusch, V., Sasaki, H., Shepherd, F., Sougnez, C., Spitz, M.R., Tsao, M.S., Twomey, D., Verhaak, R.G., Weinstock, G.M., Wheeler, D.A., Winckler, W., Yoshizawa, A., Yu, S., Zakowski, M.F., Zhang, Q., Beer, D.G., Wistuba, I.I., Watson, M.A., Garraway, L.A., Ladanyi, M., Travis, W.D., Pao, W., Rubin, M.A., Gabriel, S.B., Gibbs, R.A., Varmus, H.E., Wilson, R.K., Lander, E.S. & Meyerson, M. (2007) Characterizing the cancer genome in lung adenocarcinoma. *Nature*, 450(7171); 893-898.
- Wellcome Trust Case Control Consortium, Craddock, N., Hurles, M.E., Cardin, N., Pearson, R.D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D.F., Giannoulatou, E., Holmes, C., Marchini, J.L., Stirrups, K., Tobin, M.D., Wain, L.V., Yau, C., Aerts, J., Ahmad, T., Andrews, T.D., Arbury, H., Attwood, A., Auton, A., Ball, S.G., Balmforth, A.J., Barrett, J.C., Barroso, I., Barton, A., Bennett, A.J., Bhaskar, S., Blaszczyk, K., Bowes, J., Brand, O.J., Braund, P.S., Bredin, F., Breen, G., Brown, M.J., Bruce, I.N., Bull, J., Burren, O.S., Burton, J., Byrnes, J., Caesar, S., Clee, C.M., Coffey, A.J., Connell, J.M., Cooper, J.D., Dominiczak, A.F., Downes, K., Drummond, H.E., Dudakia, D., Dunham, A., Ebbs, B., Eccles, D., Edkins, S., Edwards, C., Elliot, A., Emery, P., Evans, D.M., Evans, G., Eyre, S., Farmer, A., Ferrier, I.N., Feuk, L., Fitzgerald, T., Flynn, E., Forbes, A., Forty, L., Franklyn, J.A., Freathy, R.M., Gibbs, P., Gilbert, P., Gokumen, O., Gordon-Smith, K.,

- Gray, E., Green, E., Groves, C.J., Grozeva, D., Gwilliam, R., Hall, A., Hammond, N., Hardy, M., Harrison, P., Hassanali, N., Hebaishi, H., Hines, S., Hinks, A., Hitman, G.A., Hocking, L., Howard, E., Howard, P., Howson, J.M., Hughes, D., Hunt, S., Isaacs, J.D., Jain, M., Jewell, D.P., Johnson, T., Jolley, J.D., Jones, I.R., Jones, L.A., Kirov, G., Langford, C.F., Lango-Allen, H., Lathrop, G.M., Lee, J., Lee, K.L., Lees, C., Lewis, K., Lindgren, C.M., Maisuria-Armer, M., Maller, J., Mansfield, J., Martin, P., Massey, D.C., McArdle, W.L., McGuffin, P., McLay, K.E., Mentzer, A., Mimmack, M.L., Morgan, A.E., Morris, A.P., Mowat, C., Myers, S., Newman, W., Nimmo, E.R., O'Donovan, M.C., Onipinla, A., Onyiah, I., Ovington, N.R., Owen, M.J., Palin, K., Parnell, K., Pernet, D., Perry, J.R., Phillips, A., Pinto, D., Prescott, N.J., Prokopenko, I., Quail, M.A., Rafelt, S., Rayner, N.W., Redon, R., Reid, D.M., Renwick, Ring, S.M., Robertson, N., Russell, E., St Clair, D., Sambrook, J.G., Sanderson, J.D., Schuilenburg, H., Scott, C.E., Scott, R., Seal, S., Shaw-Hawkins, S., Shields, B.M., Simmonds, M.J., Smyth, D.J., Somaskantharajah, E., Spanova, K., Steer, S., Stephens, J., Stevens, H.E., Stone, M.A., Su, Z., Symmons, D.P., Thompson, J.R., Thomson, W., Travers, M.E., Turnbull, C., Valsesia, A., Walker, M., Walker, N.M., Wallace, C., Warren-Perry, M., Watkins, N.A., Webster, J., Weedon, M.N., Wilson, A.G., Woodburn, M., Wordsworth, B.P., Young, A.H., Zeggini, E., Carter, N.P., Frayling, T.M., Lee, C., McVean, G., Munroe, P.B., Palotie, A., Sawcer, S.J., Scherer, S.W., Strachan, D.P., Tyler-Smith, C., Brown, M.A., Burton, P.R., Caulfield, M.J., Compston, A., Farrall, M., Gough, S.C., Hall, A.S., Hattersley, A.T., Hill, A.V., Mathew, C.G., Pembrey, M., Satsangi, J., Stratton, M.R., Worthington, J., Deloukas, P., Duncanson, A., Kwiakowski, D.P., McCarthy, M.I., Ouwehand, W., Parkes, M., Rahman, N., Todd, J.A., Samani, N.J. & Donnelly, P. (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464(7289); 713-720.
- Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145); 661-678.
- Weyand, C.M., Goronzy, J.J., Takemura, S. & Kurtin, P.J. (2000) Cell-cell interactions in synovitis. Interactions between T cells and B cells in rheumatoid arthritis. *Arthritis Research*, 2(6); 457-463.
- Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C.L., Irzyk, G.P., Lupski, J.R., Chinault, C., Song, X.Z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D.M., Margulies, M., Weinstock, G.M., Gibbs, R.A. & Rothberg, J.M. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189); 872-876.
- Wong, K.K., deLeeuw, R.J., Dosanjh, N.S., Kimm, L.R., Cheng, Z., Horsman, D.E., MacAulay, C., Ng, R.T., Brown, C.J., Eichler, E.E. & Lam, W.L. (2007) A comprehensive analysis of common copy-number variations in the human genome. *American Journal of Human Genetics*, 80(1); 91-104.
- Worthington, J. (2005) Investigating the genetic basis of susceptibility to rheumatoid arthritis. *Journal of Autoimmunity*, 25 Suppl16-20.
- Wu, X. & Freeze, H.H. (2002) GLUT14, a duplicon of GLUT3, is specifically expressed in testis as alternative splice forms. *Genomics*, 80(6); 553-557.
- Wyman, A.R. & White, R. (1980) A highly polymorphic locus in human DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 77(11); 6754-6758.

- Xie, C. & Tammi, M.T. (2009) CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, 1080.
- Xing, J., Zhang, Y., Han, K., Salem, A.H., Sen, S.K., Huff, C.D., Zhou, Q., Kirkness, E.F., Levy, S., Batzer, M.A. & Jorde, L.B. (2009) Mobile elements create structural variation: analysis of a complete human genome. *Genome Research*, 19(9); 1516-1526.
- Xu, Y. & Song, G. (2004) The role of CD40-CD154 interaction in cell immunoregulation. *Journal of Biomedical Science*, 11(4); 426-438.
- Yamamoto, T., Seino, Y., Fukumoto, H., Koh, G., Yano, H., Inagaki, N., Yamada, Y., Inoue, K., Manabe, T. & Imura, H. (1990) Over-expression of facilitative glucose transporter genes in human cancer. *Biochemical and Biophysical Research Communications*, 170(1); 223-230.
- Yang, Y., Chung, E.K., Wu, Y.L., Savelli, S.L., Nagaraja, H.N., Zhou, B., Hebert, M., Jones, K.N., Shu, Y., Kitzmiller, K., Blanchong, C.A., McBride, K.L., Higgins, G.C., Rennebohm, R.M., Rice, R.R., Hackshaw, K.V., Roubey, R.A., Grossman, J.M., Tsao, B.P., Birmingham, D.J., Rovin, B.H., Hebert, L.A. & Yu, C.Y. (2007) Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *American Journal of Human Genetics*, 80(6); 1037-1054.
- Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J. & Speed, T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4); e15.
- Yatsugi, N., Tsukazaki, T., Osaki, M., Koji, T., Yamashita, S. & Shindo, H. (2000) Apoptosis of articular chondrocytes in rheumatoid arthritis and osteoarthritis: correlation of apoptosis with degree of cartilage destruction and expression of apoptosis-related proteins of p53 and c-myc, *Journal of Orthopaedic Science*, 5(2); 150-156.
- Zendman, A.J., van Venrooij, W.J. & Pruijn, G.J. (2006) Use and significance of anti-CCP autoantibodies in rheumatoid arthritis. *Rheumatology*, 45(1); 20-25.
- Zervou, M.I., Goulielmos, G.N., Castro-Giner, F., Tosca, A.D. & Krueger-Krasagakis, S. (2009) STAT4 gene polymorphism is associated with psoriasis in the genetically homogeneous population of Crete, Greece. *Human Immunology*, 70(9); 738-741.
- Zhernakova, A., Alizadeh, B.Z., Bevova, M., van Leeuwen, M.A., Coenen, M.J., Franke, B., Franke, L., Posthumus, M.D., van Heel, D.A., van der Steege, G., Radstake, T.R., Barrera, P., Roep, B.O., Koeleman, B.P. & Wijmenga, C. (2007) Novel association in chromosome 4q27 region with rheumatoid arthritis and confirmation of type 1 diabetes point to a general risk locus for autoimmune diseases. *American Journal of Human Genetics*, 81(6); 1284-1288.
- Zintzaras, E., Voulgarelis, M. & Moutsopoulos, H.M. (2005) The risk of lymphoma development in autoimmune diseases: a meta-analysis, *Archives of Internal Medicine*, 165(20); 2337-2344.