Allele Specific Gene Expression in the Major Histocompatibility Complex

Katharine Plant

Jesus College

University of Oxford

Hilary Term 2012

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy at the University of Oxford

Abstract

Allele Specific Gene Expression in the Major Histocompatibility Complex

Katharine Plant, Jesus College, Oxford

A thesis submitted for the degree of DPhil in Hilary Term 2012

The Major Histocompatibility Complex (MHC) is a highly polymorphic region of the genome located on chromosome 6p21 in which genetic diversity has been associated with susceptibility to many autoimmune, infectious and other common diseases. Despite strong associations between disease and variation in the MHC that have been identified initially from serological testing and more recently by genome-wide association studies, functional insights into how specific variants may be altering disease susceptibility remain poorly understood in most cases. It is predicted that gene expression will play a significant role in the modulation of disease susceptibility and so further understanding of allele specific gene expression in the MHC will be necessary to help define the function of disease associated variants in this region.

This thesis aimed to define allele specific gene expression in the MHC by characterising specific candidate genes together locus-wide approaches in order to try and resolve functional variants. Gene expression was analysed in both lymphoblastoid cell lines (LCLs) and primary human peripheral blood mononuclear cells (PBMCs). Data is presented validating a novel haplotype-specific MHC microarray and fine mapping putative local, likely cis-acting, regulatory variants. This was done by expression quantitative trait mapping for two cohorts of healthy volunteers. A transcription factor ZFP57, encoded in the MHC, was found to show significant differential allelic expression relating to specific single nucleotide polymorphisms (SNPs) and possession of HLA-type. This provided new insights into reported disease associations, notably HIV-1 infection and cancer. The function of ZFP57 was further investigated in terms of genome-wide DNA binding sites by ChIP-seq together with its binding co-factor KAP1.

Allele-specific gene expression was also demonstrated for several classical HLA genes including the *HLA-C* and HLA-DQ genes, fine mapping specific putative regulatory variants. This provided new insights into disease association, notably variants of *HLA-DQB1* and susceptibility to leprosy. The applicability and sensitivity of the technique of RNA sequencing (mRNA-seq) for allele-specific quantification of gene expression was investigated for different allelic ratios of RNA from LCLs homozygous for sequence across the MHC. Significant challenges were identified in successful application of this technique to MHC genes while high levels of accuracy were observed dependent on read depth in non-MHC genes.

This thesis provides new insights into the extent and nature of allele-specific gene expression in the MHC, experimental approaches that can be used and insights gained into disease susceptibility for this important genomic region.

Acknowledgements

First and foremost, I am extremely grateful to my supervisor Julian Knight, who has constantly supported my work and provided invaluable guidance in a very patient and enthusiastic manner! This thesis would not have been possible without his supervision and encouragement. I am lucky to have worked in the Knight lab for the duration of my project, and I am truly grateful to all members of the lab past and present for their help and support with experiments and analysis, and also their friendship. I would especially like to thank Seiko Makino and Ben Fairfax for all the work on my part of the volunteer collection project and their advice and help on its analysis, as well as allowing me to be part of this exciting project. I would like to particularly thank Emma Davenport for generally being a sounding board to my musings!

Several people have assisted me with the analysis of the data I am presenting. I am grateful to The High Throughput Genomics group at the WTCHG for their handling of all my Next Generation Sequencing samples, particularly Lorna Gregory. Analysis of the mRNA-seq data and ChIP-seq data would not have been possible without Zamin Iqbal, Stephen Sansom and Andreas Heger and I cannot thank them enough! I am also grateful to Stephen Leslie and Alexander Dilthey for performing the imputation of the HLA-haplotypes and opening up this avenue of investigation to me, as well as Alexander Kanapin and Gerton Lunter for their help with the mRNA-seq data alignments.

I would also like to thank my examiners, Kirk Rockett and Timothy Vyse, for all their careful reading of my thesis, and their many helpful suggestions and comments.

Finally I would like to thank my family and friends for listening to me talk incessantly about gene expression, and for always encouraging me when I was dejected about my research. Last but definitely not least, thank you Will for all your support, I wouldn't have coped without you.

Contents

Declaration

I declare that unless otherwise stated all work presented in this thesis is my own. Several aspects of this project relied upon collaboration where part of the work was conducted by others. In particular Next Generation Sequencing was performed, mapped and variants called by the Wellcome Trust Centre High-throughput Genomics Group. Aspects of mRNA-seq data analysis were performed by Dr Gerton Lunter, Dr Alexander Kanapin and Dr Zamin Iqbal, and ChIP-seq data analysis was performed by Dr Stephen Sansom and Dr Andreas Heger of the CGAT Department. The collections and sample processing for the second 288 volunteers project were undertaken jointly by myself, Dr Ben Fairfax and Miss Seiko Makino. Imputation of the volunteer haplotypes was performed by Dr Stephen Leslie and Dr Alexander Dilthey. Dr Ben Fairfax had previously collected the first cohort of 96 initial volunteer samples and generously allowed the use of this material for the experiments detailed in Chapter 2. Genotyping the Indian leprosy cohort, which is described in Chapter 4 and subsequent haplotype analysis, was performed by Dr Tom Parks.

Abbreviations

AMD	Age-related Macular Degeneration
ASE	Allele-Specific Expression
BMI	Body Mass Index
BR	Biological Replicate
BSA	Bovine Serum Albumen
BS-seq	Bisulphite Sequencing
ChIP	Chromatin Immuno-Precipitation
CDCV	Common Disease Common Variant
CNV	Copy Number Variation
CD	Crohn's Disease
DHS	Dnase Hypersensitive Site
dNTP	Deoxynucleotide Triphosphate
ddNTP	Dideoxynucleotide Triphosphate
DMSO	Dimethyl Sulphoxide
EDTA	Ethylene Diamine Tetra-acetic Acid
EGTA	Ethylene Glycol Tetra-acetic Acid
EST	Expressed Sequence Tag
eQTL	Expression Quantitative Trait Loci
FACS	Fluorescent-Activated-Cell-Sorting
FAIRE	Formaldehyde-Assisted Isolation of Regulatory Elements
FCS	Fetal Calf Serum
GWAS	Genome-wide association Studies
HBS	Hanks Buffered Saline
HIV	Human Immunodeficiency Virus
INDEL	Insertion/ Deletion
INS-IGF2 VNTR	Insulin-IGF2 Variable Tandem Repeat
LB	Lysogeny Broth
LCL	Lymphoblastoid cell line
LD	Linkage Disequilibrium
MALDI	Matrix-assisted laser desorption/ionisation
MBP-seq	Methyl-Binding Protein sequencing
MeDIP-seq	Methylated DNA immunoprecipitation sequencing
MEM	Minimal Essential Media
MHC	Major Histocompatibility Complex

MS	Multiple Sclerosis
NGS	Next Generation Sequencing
NK	Natural Killer
OR	Odds Ratio
PBMC	Peripheral-Blood Mononuclear Cell
PMA	Phorbal 12-myristate 13 acetate
QC	Quality Control
RA	Rheumatoid Arthritis
RACE	Rapid Amplification of cDNA Ends
(m)RNA-seq	(Messenger) RNA Sequencing
RPMI	Roswell Park Memorial Institute medium
RSN	Robust Spline Normalisation
RT-PCR	Real-time Polymerase Chain Reaction
SNP	Single Nucleotide Polymorphism
SSIII	Super Script III
SLE	Systemic Lupus Erythematosus
T1D	Type 1 Diabetes
TAR	Transcriptionally Active Region
TF	Transcription Factor
TND	Transient Neonatal Diabetes
TOF-MS	Time-of-Flight Mass Spectrometry
TSS	Transcription Start Site
UTR	Untranslated Region
WTCCC	Wellcome Trust Case Control Consortium

List of Manufacturers

Abcam	AbcamPLC, Cambridge, UK
Agilent	Agilent Technologies, UK Ltd., West Lothian, UK
Beckman	Beckman Dickenson Ltd., High Wycombe, Buckinghamshire UK
Becton Dickinson	Becton Dickenson UK Ltd., Oxford Oxfordshire UK
Bioline	Bioline UK Ltd., London UK
Bio-Rad	Bio-Rad Laboratories Ltd., Hemel Hempstead, Herts UK
Coriell	Coriell Cell Repositories Camden, New Jersey 08103, USA
Clontech	Cambridge Bioscience, Cambridge, Cambridgeshire, UK
Dako	Dako Denmark A/S, Denmark
Diagenode	Diagenode s.a., Liege, Belgium
Dynal	Dynal, UK Ltd., Bromsborough, Merseyside, UK
eBioscience	eBioscience Ltd., Hatfield, UK
ECACC	European Collection of Cell Cultures, Salisbury, UK
GE Healthcare	GE Healthcare UK Ltd., Buckinghamshire, UK
Gibco	Life Technologies, Paisley, Scotland
Illumina	Illumina United Kingdom, Essex, UK
IHW	International Histocompatibility Working Group, Seattle, USA
Invitrogen	Invitrogen, Leek, The Netherlands
Miltenyi	Miltenyi Biotec GmbH, Bergisch Gladbach, Germany
Molecular Devices	Molecular Devices LLC, California, USA
Kodak	Kodak, New York, USA
NEB	New England Biolabs (UK) Ltd., Hitchin, Hertfordshire, UK
PAA	PAA Laboratories GmbH, Pasching, Austria
Qiagen	Qiagen Ltd., Dorking, Surrey UK
Roche	Roche Ltd., Beaconsfield, Buckinghamshire, UK
Santa Cruz	Santa Cruz Biotechnology, Santa Cruz, California, USA
Sigma	Sigma-Aldrich Company Ltd, Poole, Dorset, UK
Source Bioscience	Source Bioscience PLC, Nottingham, UK
Thermo Scientific	Thermo Fisher Scientific, Massachusetts, USA
Xograph	Xograph Healthcare Ltd., Tetbury, UK
5Prime	5Primer GmbH, Hamburg, Germany

Chapter 1 - General Introduction

Regulatory genetic variation plays a key role in the differences seen in gene expression throughout the human population. Much of what was once considered "junk" DNA is now known to be vital in the regulation of gene expression, including through maintenance of DNA structure or through affecting recruitment of regulatory proteins or polymerases directly. As Genome-wide Association Studies (GWAS) have shown, these non-genic, regulatory regions are often found to be disease associated, particularly in complex traits. Genetic variation in the Major Histocompatibility Complex (MHC) has been repeatedly implicated in susceptibility to disease, in particular autoimmune conditions and infectious disease but defining causal variants and mechanisms has been challenging. Differential gene expression and its regulation in the MHC is therefore likely to be of great importance in human health and disease, and highlights the need for a greater understanding of how individual genes and variants could contribute to pathogenesis of disease phenotypes.

This chapter reviews the contribution of genetic variation to regulation of gene expression regulation, focussing on methods used to determine differential gene expression and their application to disease. I will assess the role of the MHC in complex disease and how its astonishing diversity contributes to disease association. I will also attempt to demonstrate how study of genetic variation and differential gene expression is necessary for further resolution of complex traits including the many common diseases associated to the MHC.

1.1 Genetic Variation

It has long been established in the field of genetics that possession of particular genetic factors leads to particular phenotypic traits. Genetic variation is implicated in many diseases as a cause, both where a single gene is responsible and where many genes have been implicated. An example of a single disease causing variant sufficient to cause disease is the *IT15* gene, encoding Huntingtin, and Huntingdon's Disease. If the repeated CAG motif occurs more than 36 times in the *IT15* gene clinical symptoms develop (Roze 2010). Linkage and association studies combined with sequencing of whole genes, or more recently exomes, has led to identification of many diseases caused by a single gene,

where multiple variants are described at one locus. Sequencing of the *CFTR* gene for example, discovered more than 850 mutant alleles that lead to cystic fibrosis (Zielenski 2000). While single gene effects in disease are more widely understood, though by no means universally described, complex polygenic disease has been harder to study, as many disease related quantitative traits exist that all have a small magnitude of effect (Hirschhorn and Daly 2005, Altshuler 2008). For example Type I Diabetes (T1D) where over 40 loci have been identified as potentially important in pathogenesis of the disease, yet not all heritability is explained (Cooper 2008, Barrett 2009a). The number of associated loci continues to rise, with recent meta-analysis of 6 different T1D cohorts showing more than 50 associated loci (Bradfield 2011). Alleles that are extremely common are much less likely to be the cause of a single gene disorder, and when they are associated with disease have low to modest penetrance, while Mendelian disease is associated with much rarer alleles, that in turn have high penetrance (see Figure 1.1) (McCarthy 2008).

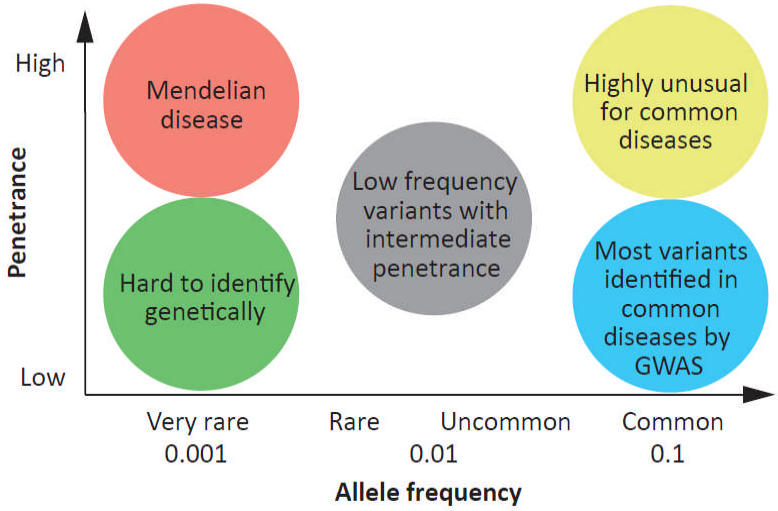


Figure 1.1: Schematic showing relationship between penetrance and variant allele frequency. Redrawn from (McCarthy 2008). Rare variants traditionally associated with high penetrance Mendelian disease are contrasted with common variants that are generally low penetrance and are discovered by GWAS. Extremely rare variants that confer a low penetrance are difficult to determine experimentally either through linkage analysis or association studies due to the large sample sizes required.

1.1.1 Types of genetic variation

Genomic variations have many different forms and sizes. A single nucleotide variant (SNV) most

commonly involves the substitution of one nucleic acid base for another in the sequence. SNVs occurring in more than 1% of a population are termed single nucleotide polymorphisms (SNPs). Insertions and deletions (INDELS) range from one base pair to large sections of a gene or chromosome. Copy number variations (CNVs) have also been implicated in phenotypic variation and disease, both as coding variants and involving regulatory sequences located away from the gene they are associated with. Where a whole gene is duplicated or deleted in its entirety, this will usually have a large effect causing increased or decreased expression of the gene (Stranger 2007, Conrad 2009, Wain 2009). The sequence variations shown in Figure 1.2 can be applied to much larger sections of the genome, for example where whole genes are inverted or parts of different chromosomes are translocated (Knight 2009a).

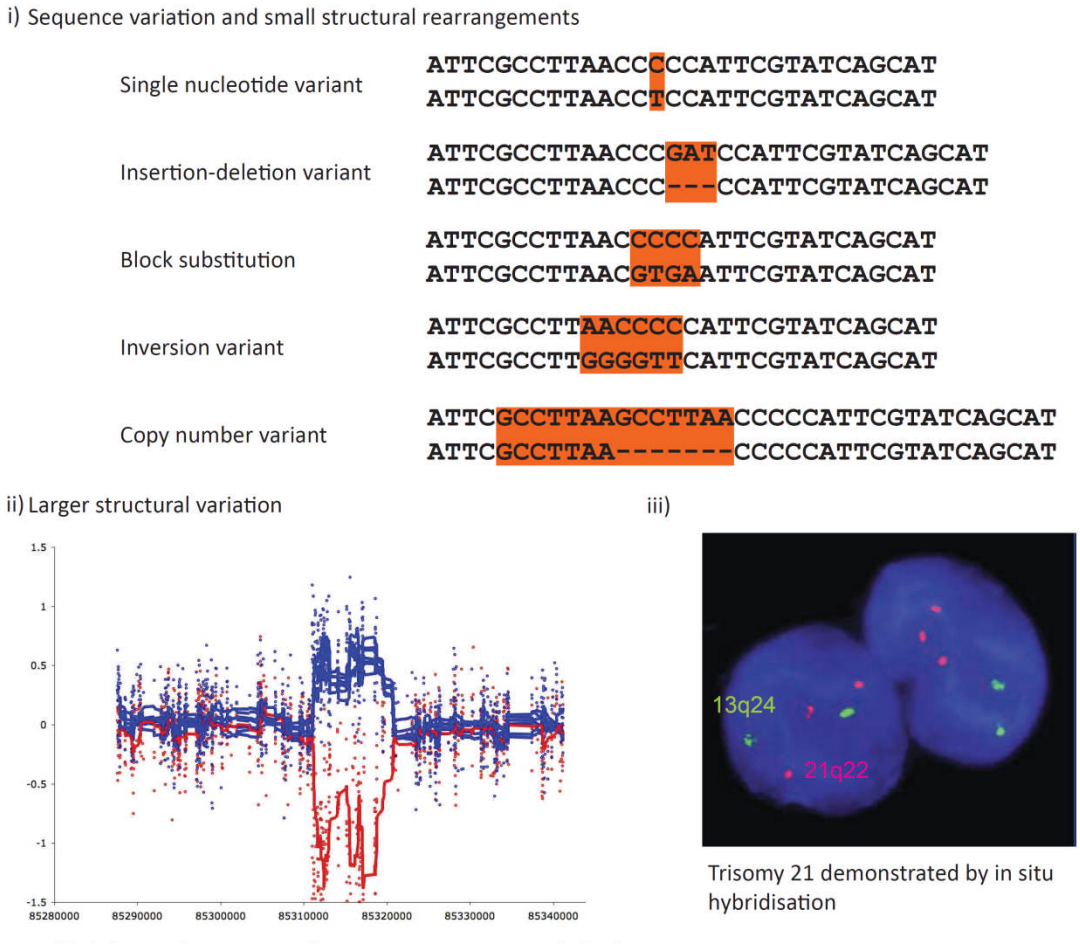


Figure 1.2: Different types of genetic variation. Redrawn from (Frazer 2009, Knight 2009a)
Genetic variation has many different types and varies from changes that affect only a single nucleotide to structural changes that alter whole chromosomes. i) Small scale sequence variation is detailed, in contrast to larger structural events and their detection using ii) comparative genome hybridisation and iii) in situ hybridisation.

1.1.2 Functional consequences of genetic variation

Due to degeneracy of the genetic code, not all variants cause changes in the amino acid code. These variants are known as synonymous or silent mutations. Where a genetic variant causes a change in the amino acid code, it is known as a non-synonymous substitution or mutation. Evolutionary pressure selects against deleterious non-synonymous, hence fewer of these mutations are found than would be expected through chance. Traditionally, research has focussed on non-synonymous mutations and their role in genetic disease. In recent years however, more emphasis has been placed on differences in promoter sequences, enhancers, introns and what was once termed "junk" DNA: the DNA that has no readily explained function (Hudson 2003). These have been implicated in control of gene expression changes as regulatory sequences; for example transcription of the human dopamine transporter is affected by intronic enhancer variation (Greenwood and Kelsoe 2003). Synonymous mutations have also been interrogated as they may not have as "silent" an effect as previously thought. They have been implicated in changes in mRNA stability and control of splicing, both of which have disease association (Chamary 2006).

Changes in gene expression levels occur in many diseases, though it is often hard to differentiate between causal variation driven by genetic differences and that which arises due the disease itself (Leek and Storey 2007). Regulatory variants may affect gene expression in several ways: for example, modulation of transcription factor (TF) binding sites, alteration of a promoter or enhancer, or creation of a *de novo* promoter or enhancer (Stranger and Dermitzakis 2005). Over the whole genome, CNVs account for more sequence diversity in terms of total base coverage than SNPs (Redon 2006). CNVs can affect gene expression a large distance away from the transcription start site (TSS) of a gene. This may be caused by disruption of enhancer sites and other modulating factors due to alterations in the structure of the DNA (Stranger 2007).

Study of the effect and incidence of regulatory polymorphisms can be carried out using expression quantitative trait loci (eQTL) mapping to identify sites of genetic variation associated with gene expression changes. Regulatory polymorphisms may occur in *cis,* where a SNP is found close to the gene it is associated with on the same chromosome, or in *trans,* where the SNP is located on a

different chromosome or much further from the associated gene (see Figure 1.3). A *trans* association could modulate expression by affecting differential expression of a transcription factor. Thus a protein encoded on a different chromosome affects the expression of a gene. Alternatively, interchromosomal configuration could allow interactions between two chromosomes, termed transvection (Williams 2010).

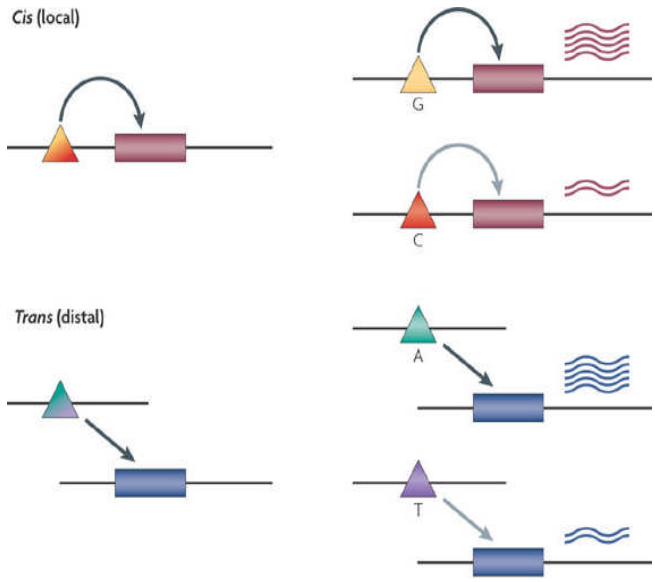


Figure 1.3: *Cis* and *trans* acting variants and their affect on gene expression (Reproduced from (Cheung and Spielman 2009)) Different allelic forms of either *cis* or *trans* acting variants have different effects on gene expression. The *cis* example shows an upstream regulatory element affecting gene expression further downstream. The *trans* example shows a regulatory variant found at a distant locus affecting gene expression.

Cis and *trans* associations have been reported in many different species from single cellular organisms such as yeast to multicellular organisms like mice and humans (Brem 2002, Cowles 2002, Pastinen and Hudson 2004, Gibson and Weir 2005, Maston 2006, Ge 2009). Comparison of two different populations found that Europeans expressed the gene *UGT2B17* an average of 22 times more than Japanese or Chinese individuals. This was linked to a deletion of *UGT2B17* in *cis* that was

much more common in the Japanese and Chinese populations (Spielman 2007). A *trans* association has been demonstrated affecting of human foetal haemoglobin. Human foetal haemoglobin is regulated by BCL11A, and regulatory variants affecting expression of *BCL11A* then impact human foetal haemoglobin (Sankaran 2008). Variants in *cis* have been more successfully replicated in their associations with gene expression and have been found more frequently than those in *trans* (Stranger 2005). How methodology or sample size limits the discovery of *trans* variants is still unclear however further *trans* associations are predicted to be revealed as techniques improve and sampling sizes increase as greater experimental power is likely to result in detection of more associations (Morley 2004, Dixon 2007). A recent study using gene knock-down to validate *trans* associations and RNA-seq to validate *cis* associations has found many previously unreported *trans* associations (Cheung 2010).

Alternative splicing, where different mRNA transcripts are created from the same DNA sequence, is also a heritable genetic trait (Hull 2007, Kwan 2007). Different isoforms created by alternative splicing are truncated versions of the full length transcript, through exon skipping or multiple polyadenylation sites, or include normally untranscribed regions through intron retention and alternative splice sites (Kim 2008). The different isoforms increase diversity of the translated protein and help modulate gene expression. Different isoforms of transcripts have different stabilities, and inclusion or exclusion of introns and exons may also lead to premature stop codons that can alter the rate of mRNA degradation (Kim 2008).

1.1.3 Haplotypes, haplotype blocks and tagging SNPs

Particular alleles or genetic markers found in an individual at one specific area of a chromosome are referred to as a haplotype. Haplotype structure can be determined for populations, defining alleles or markers that are likely to be inherited together. This can be due to strong linkage disequilibrium (LD), or non-random association of alleles, between the markers. Low recombination in an area or proximity of a recombination hotspot ensures that nearby alleles are frequently co-inherited. As distance between markers increases across the genome, the LD tends to decrease as the likelihood of recombination between the two markers becomes higher. Exact LD between markers will be

affected by the population and chromosome studied; however, due to recombination hotspot location some SNPs are consistently inherited together. Haplotypes are therefore conserved and historically preserved within populations, making them important to consider when studying variation in human genetic markers. Different haplotypes may confer specific advantages or cause susceptibility to various diseases which gives an insight into the function of the genetic variants (Gabriel 2002a). For example, a specific haplotype associated with Crohn's disease (CD) is located near to immunoregulatory cytokines (found on chromosome 5q31, between *IL3* and *IL4*), implicating them in disease pathogenesis (Rioux 2001). African population haplotypes have greater diversity compared to haplotypes found within other populations, consistent with the "Out of Africa" theory surrounding evolution of the human race. Migrating groups from Africa are thought to have become genetically isolated in different locations thus allowing population bottle necks to form (Templeton 2002). Diversity could be lost due to specific selective pressures in these locations; however, the original population they migrated from would not lose genetic variation, and would continue to accumulate variation.

The MHC on chromosome 6p21 has been considered by many to be a paradigm in the dissection of haplotypic structure, LD and recombination, as the term haplotype came about to describe "the combination of individual antigenic [MHC] determinants that are positively controlled by one allele" (Mungall 2003). It was the first region of the genome to be studied extensively in terms of its LD (Vandiedonck and Knight 2009). Ancestral or conserved extended haplotypes can extend over the entirety of the 3.6Mb classical MHC and are particularly common in European populations (Degli-Esposti 1992, Yunis 2003). This may be due to selection, with striking disease associations reported (see Section 1.4.6).

"Recombination Hotspots" are locations in the genome where meiotic recombination occurs at a higher than average frequency. These form boundaries of "haplotype blocks"; genomic areas with low recombination and strong LD (Daly 2001). In addition to studies on the MHC regarding linkage and haplotypes, the MHC was also one of the first regions to be studied regarding haplotype structure due to the previously existing characterisation of familial crossovers (Cullen 1997). With

this information and analysis of LD over a 200kb region, six hotspots were defined that seemed to explain all crossovers in the region. These hotspots had similar morphology showing they were non-random in their location and could define the LD of a region (Jeffreys 2001). The following year, study of 51 autosomal genomic regions in African, European and Asian populations showed that the average haplotype block size was 11kb in the African population, and 22kb in the European and Asian populations (Gabriel 2002a). This also confirmed that haplotypes were commonly shared among the three populations, but where a unique haplotype was found it predominantly occurred in the African population. Despite the name of "ancestral haplotypes" used in the context of the MHC, it is clear that recombination hotspots rather than population history have a greater impact on the LD patterns seen. A study showed that in three separately arising populations a high divergence in haplotypes was seen but similar patterns of LD structure occurred (Kauppi 2003).

As further advances in sequencing were made, a set of tagging SNPs for association studies was suggested. These would be used to define variants for GWAS. Gabriel and colleagues proposed that between 300,000 and 1,000,000 SNPs would be needed to define the haplotypic diversity in a population for the purposes of an association study, depending on the population of study (Gabriel 2002a).

1.1.4 The International HapMap Project

With this in mind, the International HapMap Project was launched in 2002. Its aim was to produce a genome-wide map of one million or more SNPs, found at a frequency of more than 5%, in diverse populations, to establish haplotypic structure across different human populations (TIHMP 2003). 270 individuals from African (YRI), European (CEU) and Asian (Chinese (CHB) and Japanese (JPT)) populations comprising of either unrelated individuals (CHB/JPT) or parent-child trios (CEU/YRI) were recruited to Phase I of the HapMap project. This identified more than one million SNPs in 269 individuals and recombination rates and hotspots were identified over the whole genome at an average of one hotspot per 57kb of sequence (Altshuler 2005). Phase II of the HapMap project led to identification of more than three million mapped variants using the original 270 individuals (Frazer 2007). This study suggested that for GWAS an informative panel of SNPs for the CEU population

should be around 500,000, while 1,000,000 would be necessary for an YRI population (Kruglyak 2008) to accurately reflect common genomic variation. Phase III included seven new populations and 1,184 samples in total from all 11 populations. They were genotyped at 1.6 million common SNPs. 100 10kb regions were sequenced from 692 of the individuals to allow characterisation of lower frequency variants (Altshuler 2010). This suggested that dense SNP genotyping and imputation in a well characterised haplotype background could be used to successfully interrogate low frequency variation.

1.1.5 The 1000 Genomes Project (www.1000genomes.org)

To exhaustively catalogue human genetic variation the 1000 Genome Project was set up in 2008 to perform whole genome sequencing of multiple individuals (Siva 2008). Pilot data from this project became available in 2010. In total, 179 individuals were sequenced at low coverage, 2 parent-child trios were sequenced at high coverage and 697 individuals were sequenced with exome targeting. Around 15 million SNPs, one million short INDELs and 20,000 larger structural variants were identified. The majority of variants found were previously unknown and tended to be found in only one population, likely to be because any novel variant discovered following previous characterisation was disproportionately rare. It was also predicted that every individual will have 250-300 loss of function mutations in coding sequences, as well as 50-100 loci that are heterozygous for alleles associated with inherited diseases (Durbin 2010). The pilot study concluded that over 95% of human genetic variation was now catalogued, as most common variation had been described. The production phase of The 1000 Genome Project aims to fully sequence 2500 individuals from 5 different large area populations (Europe, East Asia, South Asia, West Africa, The Americas) at a depth of 4x, taking a step towards complete description of human genetic variation to help analyse human evolution and inherited traits. Collection of parent/ child trios will allow the chromosomal phase of parents to be determined from the child after high density genotyping (Via 2010).

Imputation of genotypes, where genotyped variants are used to infer variants at non-genotyped locations has been improved by data both from the 1000 Genome and HapMap projects (Altshuler 2010, Durbin 2010). Several algorithms for imputation are now available and can accurately

determine genotypes when a good-quality genotyping data cohort is available (Hao 2009). Imputation significantly increases power in GWAS where a good haplotype map for the population exists as well as helping to define boundaries of significantly associated haplotypes (Anderson 2008). As sequencing costs fall, and demand for personalised medicine increases, exome and low coverage whole genome sequencing will become more common. This will identify more rare variants, especially insertions and deletions, leading to an even more complete understanding of human genetic variation (Nielsen 2010).

1.1.6 Genome-wide Association Studies (GWAS)

GWAS attempt to locate variants that are important in the development of particular traits by using sets of SNP markers that define genotypes at particular regions of the genome (Hirschhorn and Daly 2005, Pastinen 2005). Many studies have discovered regions of the genome associated with disease by comparing cohorts of samples from patients and disease-free controls (van Heel 2007, Lettre and Rioux 2008). In some cases, particularly earlier candidate gene association studies, findings from these case-control comparisons were not reproducible and concerns were raised over population stratification (Hutchison 2004) especially where ethnicity was not taken into account (Marchini 2004, Wang 2005). This led to large scale GWAS where frequencies of hundreds of thousands of marker SNPs are compared between case samples and carefully chosen phenotyped controls in an effort to combat these problems (WTCCC 2007, Barrett 2008). The most well known of these large scale studies was the first Wellcome Trust Case Control Consortium study (WTCCC), which compared 14,000 case samples from 7 common diseases with 3,000 matched controls that were shared between the different diseases (WTCCC 2007). The studies have become larger and more exhaustive as sequence information grows. The second Wellcome Trust Case Control Consortium study (WTCCC2) investigated associated loci with 13 new diseases as well as the genetics of reading and maths ability in children, and the pharmacogenetics of statin response. Results of this study are already emerging, for instance the implication of peptide handling in ankylosing spondylitis patients carrying the HLA-B*27 haplotype and the identification of new susceptibility loci in Parkinson's disease (Evans 2011, IPDGC and Vincent Plagnol 2011). Wellcome Trust Case Control Consortium

study Three (WTCCC3) will use the same controls and 40,000 further patient samples to study another four disease conditions; primary biliary cirrhosis, anorexia nervosa, and pre-eclampsia, as well as host control of HIV-1 and interactions between donor and recipient DNA related to early and late renal transplant dysfunction (http://www.wtccc.org.uk/ccc3/). GWAS research has also focussed on the choice of SNPs used for genotyping (Barrett and Cardon 2006), in addition to the number required for accurate whole genome coverage in a given population (Kruglyak 2008) to maximise findings of a study.

Generally GWAS have proven informative, identifying new loci and pathways that contribute to disease. Interestingly, many disease susceptibility loci are shared between different diseases, particularly in autoimmune diseases. However, effect sizes identified by GWAS have been relatively small, and explain a minority of heritability. For example, the extensive meta-analysis undertaken on T1D has not explained all inherited risk despite the identification of more than 50 susceptibility loci (Bradfield 2011). Larger GWAS studies are allowing greater numbers of variants that affect disease susceptibility to be identified, especially those with very low frequency, or those that have ever smaller effect sizes.

One of the best reported successes of GWAS has been the study of Crohn's Disease. By 2008, more than 30 loci had been identified from several different studies that showed good concordance between their results (Barrett 2008, Mathew 2008). Identified genes were involved in a wide range of immunological processes, but some associations were located in gene deserts. Strikingly, none of the associations reported had large effect sizes. As sample sizes became larger, power to detect association increased, with an odds ratio of around 1.5 being detected by initial GWAS and around 1.2 by the meta-analysis. *NOD2* was previously known as a susceptibility locus for CD so its implication by GWAS was unsurprising. However, GWAS has highlighted the role of IL23 signalling and autophagy in CD (Mathew 2008). The number of loci identified as associated with CD susceptibility has reached over 70, and this is likely to increase as larger studies take place (Franke 2010). However, predictions show that perhaps only 25-50% genetic of variance can be detected by

GWAS (see Figure 1.4), suggesting that rarer variants with higher penetrance and other low frequency variants must exist (see Section 1.2).

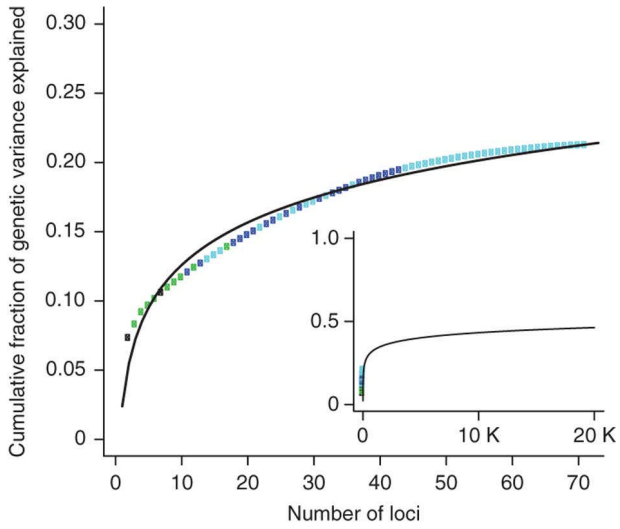


Figure 1.4: Cumulative explained genetic variance associated with CD, reproduced from (Franke 2010). The loci are ordered according to overall effect, from larger effect size loci to smaller. Pre-GWAS associations (*NOD2* and *IBD5*) are coloured black, initial GWAS hits in green, dark blue points were first identified in meta analysis from 2008 and light blue points were first identified in meta analysis from 2010. The inset logarithmic graph shows that in an extreme case where 20,000 variants are hypothetically associated with CD, only 50% of variance is explained.

Association of late onset Alzheimer's Disease and APOE ε4

ε4, the isoform of ApoE had been associated with late onset Alzheimer's disease by linkage studies, showing that there was a locus associated with familial late onset Alzheimer's disease located on chromosome 19 (Pericak-Vance 1991). Shortly after this finding, it was noted that the APOE ε4 isoform was significantly enriched in familial Alzheimer's patients compared to controls and was also seen in sporadic Alzheimer's patients (Saunders 1993, Strittmatter 1993a). APOE ε4 was noted to bind to the amyloid β peptide, part of the Aβ dimer in an isoform specific manner (Strittmatter 1993b). The Aβ dimer is the main component of amyloid plaques found in Alzheimer's patients. As APOE ε4 allele number increased, Alzheimer's risk increased from 20% to 90%, and age of onset decreased from 84 to 68, showing a dosage effect (Corder 1993). The ε4 isoform is hypothesised to alter the clearance and aggregation of Aβ, thus contributing to Alzheimer's Disease (Kim 2009). Despite functional evidence of association, the APOE ε4 allele was neither necessary nor sufficient to

cause onset of Alzheimer's disease, therefore additional genetic variants were determined using GWAS and candidate gene studies (Waring and Rosenberg 2008). Along with many other susceptibility loci, rs440638 was identified as being associated with Alzheimer's Disease, unsurprising as it was in direct LD with the APOE ε4 variant (Coon 2007). This association would have been missed had rs440638 not been typed on the 502,000 Affymetrix panel of SNPs used in genotyping. This highlights the limitations of GWAS when using SNP coverage to determine associated loci. While it is easy to point out near misses when an association in question is documented, how many more must occur when an association is not previously known?

CD and Alzheimer's disease show there is still a long way to go before all genetic susceptibility to complex disease is explained. Deep sequencing of whole genomes to discover new variants and pinpoint disease related genes could improve this. For example, a recent study has determined the causative gene in the Mendelian disorder Miller Syndrome by sequencing the whole genomes of a family of four: two unaffected parents and two affected children (Roach 2010).

Common Disease Common Variant Hypothesis

Following publication of the draft human genome sequence (Lander 2001, Venter 2001) there was an expectation that association of genetic variants with traits and disease would become common-place and the heritability of previously undetermined conditions would become apparent (Risch and Merikangas 1996). The "common disease common variant" (CDCV) hypothesis states that a common polygenic trait is dependent on possession of particular common variants, i.e. variants that are found in at least 1% of the human population (Reich and Lander 2001, Schork 2009). However, as GWAS were carried out common variants were found to play a much smaller role than expected in heritability of traits, even where much of the overall heritability was found in total (Goldstein 2009). Most GWAS found many common genetic variants playing a role in heritable risk, where each loci had a comparatively small effect size (Plomin 2009).

Missing heritability in complex traits has led to questions about the accuracy of the CDCV hypothesis, suggesting rare variants may play a much larger role than previously expected. Alternatively,

common variants have such low penetrance or such a small effect that GWAS studies could not accurately define them (Maher 2008). Another theory to explain missing heritability suggests a spectrum of different effects at each locus, with both common and rare variants playing a role but rare variants exerting a larger effect (Manolio 2009). Gene-gene and gene-environment interactions also make the task of identifying genes underlying these complex traits more difficult (Glazier 2002). Studies that seek to determine the separate impact of nature and nurture on complex traits are relatively new, and as yet no definitive analysis has been carried out (Ionita-Laza 2009, Grundberg 2011).

1.1.7 Complexity in heritable determinants of complex traits

Sequence variation, splicing and environmental factors help to explain many complex traits and diseases; however, it is likely that additional factors also influence these phenotypes. Epigenetic variation, defined as transfer of information in a heritable manner which does not require an underlying DNA sequence change, could also be important (Richards 2006). This can encompass many changes, such as DNA methylation, histone modifications, post-translational modifications and non-coding RNAs. However, the mechanism of inheritance is currently only well described for methylation (Bell and Beck 2010). Differences in DNA methylation are associated with schizophrenia and bipolar symptoms in humans. Additionally, lifetime anti-psychotic drug use was linked specifically to methylation changes at the *MEK1* promoter (Mill 2008). Structure of chromosomes themselves may also affect genetic regulation; looping of chromosomes bringing regulatory elements into closer proximity has been shown in *cis* (Majumder 2008) and there is evidence that this may also occur in *trans* (Williams 2010).

The tissue used in a study may also cause bias when investigating genetic determinants of a complex trait (Heinzen 2008). Tissue specific gene expression and tissue specific splicing have both been described, in addition to variable responses to environmental stimuli depending on tissue type (Wang 2008b, Nica 2011). Common regulatory variation is known to act in a cell-type dependent manner, therefore regulatory variation may be missed if a study does not focus on a suitable tissue (Dimas 2009). A 2008 study used adipose tissue and blood for expression analysis when investigating

Body Mass Index (BMI) and obesity. Analysis of gene expression traits in both tissues found more obesity related biometric traits (such as BMI) in the adipose tissue when compared to the blood samples, illustrating the importance of matching the tissue to the complex trait being studied (Emilsson 2008).

1.2 Genetics of Gene Expression

1.2.1 Expression Quantitative Trait Loci (eQTL) Studies

In order understand the functional consequences of genetic variation, eQTL studies have become more prominent (Franke and Jansen 2009). As many GWAS hits occur in areas of non-coding DNA it is assumed that these loci play a regulatory role (Manolio 2010). Studies aiming to define genetic determinants of gene expression have analysed the whole genome using expression microarray and genotyping arrays (Cheung 2005, Goring 2007). Many loci found by eQTL studies overlap with disease or other phenotype associated loci thus making them good candidates for further functional study. However, many of these overlapping loci could be due to coincidence because of strong LD structure throughout parts of the genome and the abundance of eQTL signals (Nica 2010). Another difficulty with microarray based eQTL analysis is that SNPs can cause bias in the results if probe hybridisation is affected (Alberts 2007). The platform used for detection can cause bias; only 30-40% of transcript detection is conserved across different platforms, whether considered as relative levels or at an absolute level (Cookson 2009).

RNA sequencing (RNA-seq) may help to overcome these issues and has been employed in two recent eQTL studies (Montgomery 2010, Pickrell 2010). Both use lymphoblastoid cell lines (LCLs) of European or African descent extensively analysed by the HapMap project. These studies have found higher numbers of eQTLs than array based studies and the genome-wide coverage of RNA-seq allows previously unknown transcripts to be analysed. RNA-seq can be used to directly identify rare regulatory variants, for example comparing 1000 Genome Project variants with RNA-seq data. Putative causal SNPs for rare allele-specific expression (ASE) events were identified, with a median of four identified for each allele-specific event studied (Montgomery 2011).

1.2.2 Allele-specific expression

One of the most striking examples of allelespecific expression (ASE) is allelic silencing, for example autosomal silencing (Morison 2005) or X-chromosome inactivation (Plath 2002). Often allele silencing is mediated by imprinting, particularly in developmental settings where imprinted marks are passed down in a heritable manner from one or both parents (Reik and Walter 2001, Lewis and Reik 2006, Kacem and Feil 2009, McEwen and Ferguson-Smith 2010). Often imprinted genes are found clustered together under control of an imprinting control region (Reik and Walter 2001). Some of the most commonly cited examples of imprinting are found in a cluster on chromosome 15. Here, normal gene expression is mediated by differential imprinting depending on the parent of origin. Two associated syndromes are reported where expression of either an allele from the mother or the father is lost. In Prader-Willi syndrome a paternal copy of the critical region at chromosome 15 including the *SNRPN* gene is lost, generally due to a deletion, or in more unusual cases, a structural rearrangement of the area resulting in loss of epigenetic control (Ledbetter 1981, Buiting 1995). As the maternal copy is generally imprinted, gene expression is lost and Prader-Willi syndrome develops. Conversely, Angelman syndrome develops when expression of a maternal allele in the region is lost. Maternal *UBE3A* expression is lost through deletion (15q11-q13), mutation or other translocation (Fang 1999) and the paternal allele is imprinted so expression of *UBE3A* is completely stopped, leading to abnormal development (Nicholls 1993).

Imprinting is not the only form of ASE seen in humans. Variable levels of expression of both alleles have been described for many genes as a phenomenon of normal gene expression (see Figure 1.5), and have also been observed in other organisms such as mice and maize (Yan 2002, Cheung 2003, Schadt 2003, Buckland 2004, Knight 2004, Morley 2004). Variable gene expression of different alleles can also be inherited and can be studied as heritable traits (Darvasi 2003, Monks 2004). In one of the earliest demonstrations of ASE, 13 different genes with a heterozygous SNP within them were studied in LCLs derived from individuals in the CEPH families. ASE was seen in 3-30% of LCLs at these genes (Yan 2002). Because the LCLs used had known pedigrees, the heritability of these expression traits could be determined to show that ASE was a heritable trait. Primary cells from humans were

also shown to exhibit ASE in a study analysing RNA from post-mortem brain tissue in 60 individuals (Bray 2003). Genome-wide studies of allele specific protein-DNA interactions have shown that RNA polymerase binds differentially at both imprinted genes and for other ASE events (Maynard 2008) consistent with the hypothesis that SNPs or other genetic variants could be responsible for variation in expression (Brem 2002, Cheung and Spielman 2002).

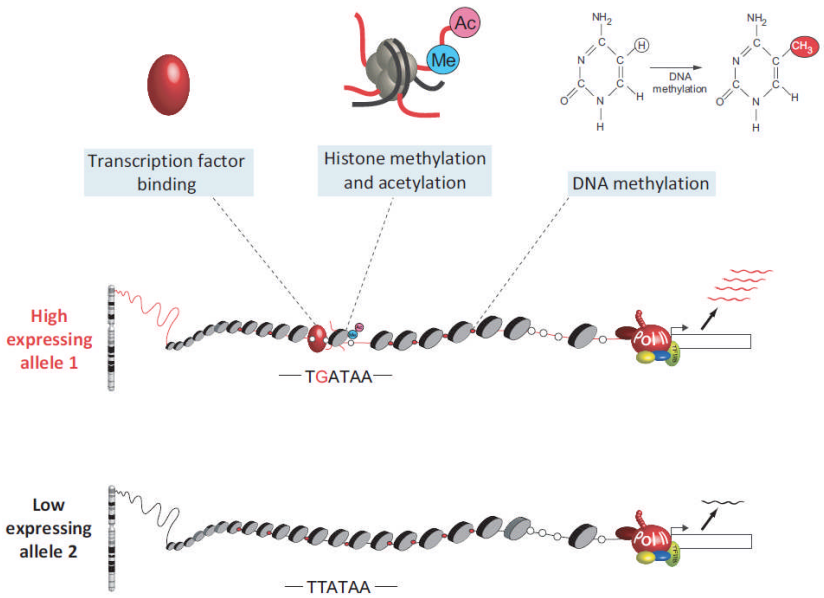


Figure 1.5: Allele-specific expression from two chromosomes. Reproduced from (Knight In Press). The regulatory sequence of the gene is different between 2 alleles and affects the expression of the gene, for example by affecting the binding of TFs or altering the structure of the chromatin via histone modifications or the DNA by methylation.

Regulatory polymorphisms that are associated with disease have also been shown to modulate gene expression, for example the insulin IGF2 variable tandem repeat (INS-IGF2 VNTR) (Knight 2005b). The INS-IGF2 VNTR is thought to be responsible for variation in levels of proinsulin and insulin in the thymus, where reduction in expression could reduce self-tolerance and increased autoimmune responses, increasing T1D susceptibility (Pugliese 1997). Another example of regulatory variation leading to disease was reported where a SNP between the alpha-globin genes and their upstream regulatory elements created a new promoter. This new promoter interfered with normal gene expression and lead to development of thalassemia (De Gobbi 2006).

1.2.3 Studying allele-specific gene expression

ASE has a great deal of promise in terms of elucidating the effect of regulatory polymorphisms on gene expression. In a heterozygous individual, studying relative expression of the two different alleles within a cell has exactly the same conditions and environment so many confounding factors are removed. This has substantial advantages over using an averaged cell population however does not resolve the complex genetic background, or each individual cell having a unique expression pattern (Levsky 2002).

Originally, the measurement of ASE required using a fluorescence based dideoxy-terminator (ddNTP) for primer extension in a PCR reaction (Yan 2002). This was laborious as each gene had to be interrogated individually. Further primer extension protocols that determined ASE using either matrix-assisted laser desorption/ionisation (MALDI) or time-of-flight mass spectrometry (TOF-MS) were similarly limited by the need for an exonic marker (Knight 2003, Jurinke 2005). Wider scale analysis of ASE has been carried out using genome-wide microarray analysis, initially in foetal tissue where an estimated 50% of the 602 genes tested showed evidence of ASE (Lo 2003). Following this study several others have interrogated a wide range of cell lines and primary tissue for ASE using array-based technology (Pant 2006, Bjornsson 2008, Dimas 2008, Serre 2008, Verlaan 2009). These studies show that ASE across the whole genome is thought to occur in around 10-20% of all genes and is composed of a mixture of modest expression differences between alleles as well as the monoallelic expression thought to be the result of imprinting. Conducted a couple of years later than the study by Lo et al., they tended to give lower estimations of genome-wide ASE. This may have been due to an over-estimation of ASE in the first studies due to the technical challenges associated with determination of differential expression (Bell and Beck 2009). Custom arrays have striven to reduce the impact of bias occurring at the PCR amplification stage (Bjornsson 2008) and over time longer probes have been used to reduce cross-hybridisation that may have impacted the earlier studies causing the over-estimation. Again, these methods relied upon an exonic SNP to define the two alleles for analysis. Intronic SNPs found in pre-spliced mRNA have been used for this purpose but as pre-spliced mRNA is found at a low abundance this is only suitable for analysis of highly

expressed genes (Pastinen and Hudson 2004, Serre 2008).

A method avoiding the need for a transcribed SNP marker in analysis focusses on the relative abundance of RNA polymerase on each of the two alleles (Knight 2003). Phosphorylated RNA polymerase is immunoprecipitated while crosslinked to DNA. The crosslinks are then reversed and DNA is quantified in an allele-specific manner to determine if the different alleles cause a bias in polymerase loading, for example by MALDI/ TOF-MS as previously discussed (Knight 2006). In this technique, termed HaploChIP (haplotype-specific chromatin immunoprecipitation), the marker SNP used to distinguish the alleles may be up to 2kb away from the gene of interest (Knight 2005a). HaploChIP can also be used in a genome-wide manner, where RNA polymerase ChIP products are hybridised to a SNP genotyping array to locate evidence of AS binding (Maynard 2008).

As RNA-seq becomes more affordable and reliable as a technique, it will likely revolutionise the study of ASE. In combination with genotyping assays it will be possible to determine how expression levels vary according to allele possession in terms of both variants within a gene and those that are affecting its expression from a distance (Majewski and Pastinen 2011).

1.2.4 Allele-specific alternative splicing

It has already been noted that alternative splicing creates increased diversity across the whole genome, adding to phenotypic variety (Blencowe 2006, Moore and Silver 2008). Aberrant regulation of alternative splicing has been implicated in many human diseases. Variations in alternative splicing can cause disease, including cancer, and also modulate how severe disease is or the susceptibility to a disease (Wang and Cooper 2007, Tazi 2009).

Each transcript is unlikely to be capable of being expressed as every possible isoform in every case, and it has been shown that 6-21% of alternatively spliced genes had varying isoform abundance depending on the possession of particular alleles (Nembaware 2004). By using Expressed Sequence Tags (EST), Exon array data and computational identification of likely SNPs to regulate alternative splicing, a genome-wide picture of allele specific alternative splicing has been generated. This implies the overall heritability of alternative splicing is greater than that of differential gene expression and suggests that allele specific splicing is a powerful source of genetic variation

(Nembaware 2008). Alternative splicing has also been studied over the whole genome using a second methodology, where genome-wide SNP analysis was combined with expression data from an exon targeted array (Kwan 2008). LCLs were analysed for isoform-specific expression; it was estimated that around 50-55% of the variation in gene expression in an individual is isoform based. The full extent of alternative splicing is debated, especially as a bias based on the exonic probes and the variation in transcript abundance has been described elsewhere (Sorek 2004, Gaidatzis 2009). However, other studies have also estimated alternative splicing in the genome and have suggested that up to 74% of multi-exon genes may be alternatively spliced (Johnson 2003). As with the study of ASE, next generation sequencing (NGS) should help to solve some of the uncertainty around the prevalence of alternative splicing and will allow more accurate assessment of how the two phenomena are linked.

1.3 The MHC

1.3.1 MHC structure and polymorphism

The MHC is found on the short arm of chromosome 6 and is one of the most gene rich regions of the genome with 224 gene loci and exhibits remarkably high levels of polymorphism (Horton 2004). Early studies noted this high diversity of sequence in human MHC genes and as early as 1981 it was suggested that each combination could be unique (Dausset 1981). The classical MHC spans 3.6Mb and contains three regions where genes that are involved with different aspects of the immune system are often clustered together according to their roles (Figure 1.6).

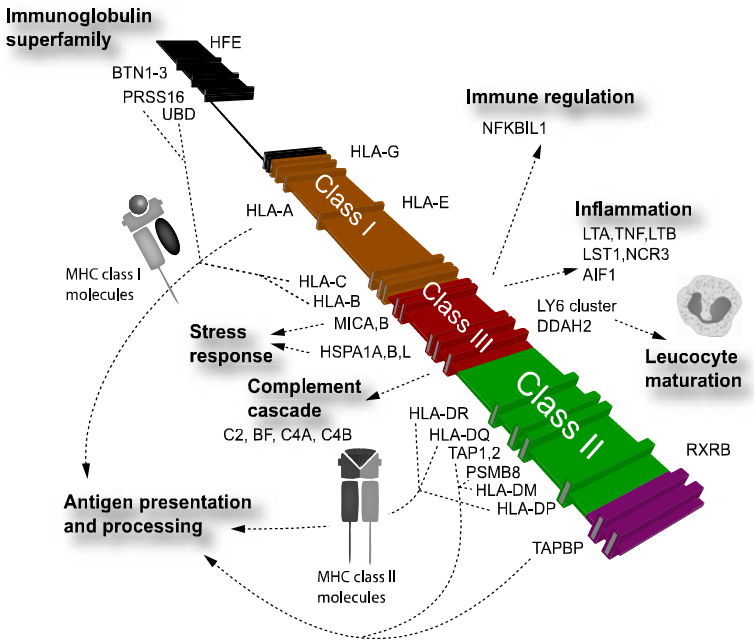


Figure 1.6: Immune related genes encoded on the MHC. Reproduced from (Knight 2009b). The three regions of the classical MHC are shown in a schematic. Positions of some of the immune related genes encoded by them are indicated. Various cellular processes that particular genes are involved in are also indicated.

The MHC has been shown to control the immune response towards particular antigens (Benacerraf 1981), with the class I and class II regions encoding molecules that are important in the presentation of antigen to T cells (Braciale 1987). Antigen presentation is detailed in Figure 1.7.

The MHC class I region contains the HLA-class I supercluster of the classical MHC genes *HLA-A*, *-B*, and *-C*, as well as many others and encodes molecules that are associated with presentation to CD8+ T cells. The class II molecules are associated with presentation to CD4+ T cells and are encoded by the class II region. This cluster contains both classical class II genes, such as *HLA-DQA*, *-DQB* and *-DR*, that encode α and β chains of cell surface antigen presenting molecules and non-classical class II genes that are not expressed on the cell surface. Instead, non-classical class II molecules, such as *HLA-DM* and *-DO*, are involved with intracellular peptide exchange and loading onto the antigen presenting complexes (Horton 2004). The high polymorphism in these regions involving antigen presentation is thought to be selected for by the heterozygote advantage, where a more diverse immune response against infection can be mounted than that from a homozygous individual (Carrington 1999).

The MHC class III region is the most dense gene area of the genome with 62 expressed genes (Xie 2003). Many of these are involved in the humoral immune response (Vandiedonck 2004), the stress and inflammatory responses via the complement cascade, heat shock proteins (Milner and Campbell 1990) and the tumour necrosis factor family and systemic inflammation (Bayley 2004, Vandiedonck and Knight 2009). The regions either side of class I and class II are referred to as the extended MHC due to their similar functions and the LD of the region, making the total MHC 7.6Mb in length (Stephens 1999). The MHC includes many different clusters of genes, such as the zinc finger supercluster, the histone supercluster, the olfactory-receptor supercluster and the butyrophilin supercluster (Horton 2004). A summary of pathways the MHC contributes to and the genes involved is shown in Figure 1.6.

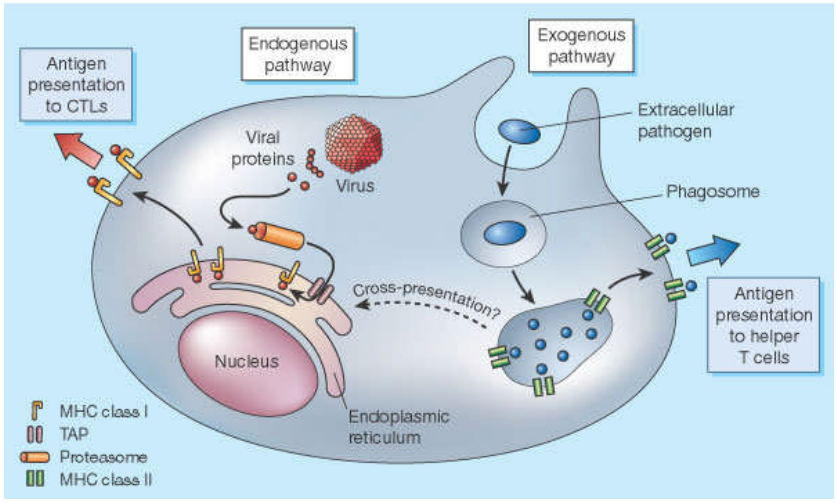


Figure 1.7: Antigen presentation involving the MHC classes I and II proteins. Reproduced from (Roy 2003). The endogenous pathway, involving MHC class I molecules, and the exogenous pathway, involving MHC class II molecules, are shown in an antigen presenting cell. In the endogenous pathway, intracellular pathogens such as viruses are broken down by proteasomes and shuttled to the endoplasmic reticulum by the TAP protein. Antigenic peptides are then displayed on the cell surface by the MHC class I complexes, formed from proteins encoded in the MHC class I region. In the exogenous pathway, extracellular pathogens are engulfed by phagosomes and degraded to be presented at the cell surface by the MHC class II complexes, formed from proteins encoded on the MHC class II region.

In addition to genetic diversity presumed to be conferred by the heterozygote advantage, the presence of rare alleles in the MHC could also be conferring an evolutionary fitness by allowing presentation of pathogenic peptides that are not represented by more commonly found MHC alleles (Borghans 2004). This is thought to be important in generation of the high diversity seen in the MHC class I gene *HLA-B* (Prugnolle 2005). Both possession of rare alleles and heterozygote advantage are likely to be influential in maintenance of this diversity; for example, HIV-1 mutates quickly and adapts to common alleles meaning possession of rarer MHC alleles with the additional increased diversity of a heterozygote confers an advantage to the host (Carrington 1999, Trachtenberg 2003).

1.3.2 Role in disease and immunity

The MHC was identified as having a crucial role in transplantation via histocompatibility antigens (Dausset 1972). The role of the MHC in autoimmune disease was first discovered in the 1970s based on serological testing and subsequently based on association analysis using genetic markers. For example among Rheumatoid Arthritis (RA) patients particular HLA-DRB1 alleles were identified as being significant in disease susceptibility (Stastny 1976). The MHC has been implicated in human disease many times in both Mendelian and complex disorders. This was highlighted by the initial WTCCC GWAS of seven complex human diseases (see Figure 1.8) when the MHC was associated strongly with T1D and RA, and less strongly with CD (WTCCC 2007).

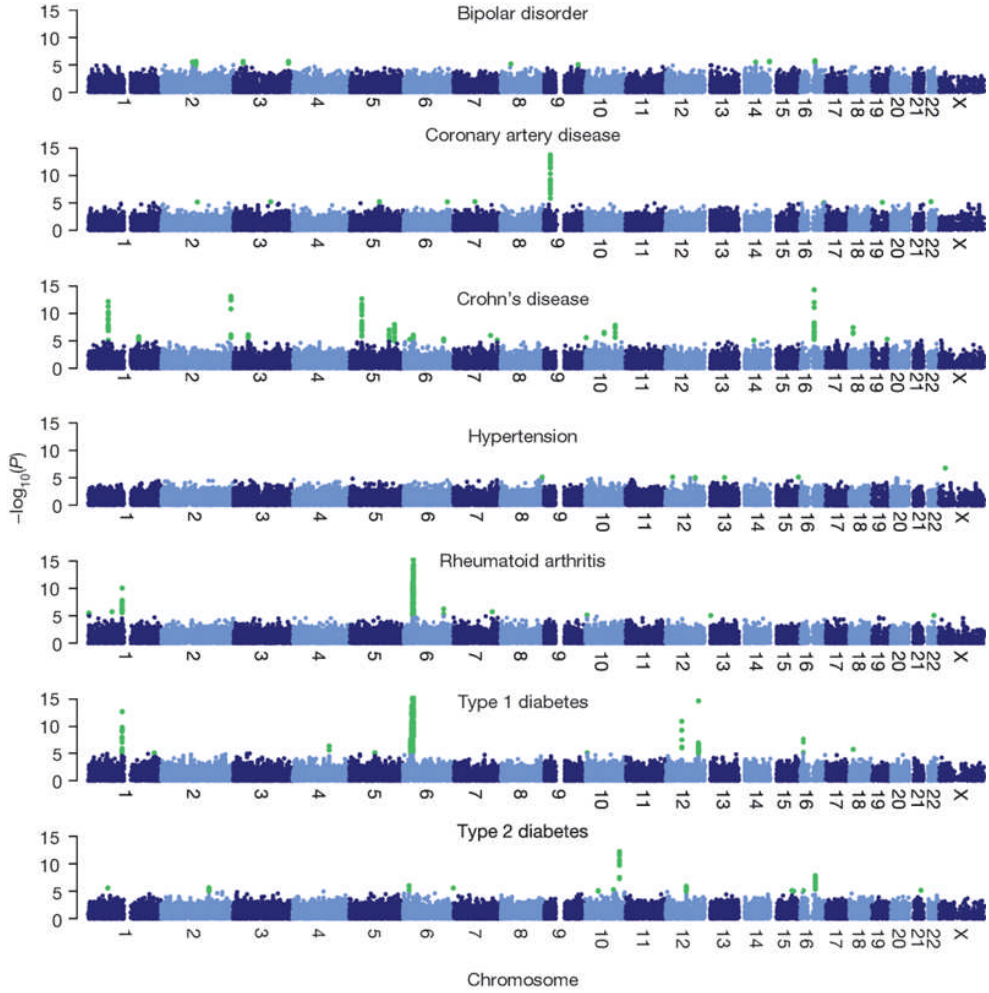


Figure 1.8: Manhattan plots showing association of genomic loci with disease susceptibility. Reproduced from (WTCCC 2007). The MHC is strongly associated with two complex disease traits investigated by the WTCCC. Statistically significant SNPs associated with each disease are indicated at their chromosomal position in green. The strongest associations between the MHC and disease are seen with the autoimmune conditions T1D and RA, with associations also seen in CD.

Due to the high polymorphism and strong LD in the MHC, specific causal variants have been harder to define. However the MHC has emerged as the most important risk locus throughout the whole genome for many autoimmune diseases, notably T1D, Multiple Sclerosis (MS), RA, psoriasis, systemic lupus erythematosus (SLE) and coeliac disease (Trembath 1997, Sawcer 2002, Aly 2006, Fernando 2008b, Hunt 2008, Chanda 2009, Eleftherohorinou 2009, Jiang 2009, Mathew 2009, Rioux 2009, Viken 2009).

Many of these diseases, such as T1D and CD have been intensively studied, combining several large cohorts with meta-analysis and identifying large numbers of susceptibility loci, as previously noted

(Franke 2010, Bradfield 2011). Other autoimmune diseases associated with the MHC such as SLE present problems when they are investigated due to low cohort sizes, heterogeneity within a cohort of patients and treatment complications (Rhodes and Vyse 2008). SLE is not found at high prevalence, particularly in the European population, thereby affecting the availability of patients to recruit for a study (Rhodes and Vyse 2010). Despite this, genetic determinants of SLE have been successfully identified as the incidence of SLE aggregates in families. Monozygotic twins have a concordance of disease of between 24-57%, compared with 2-5% in dizygotic twins. RA is also suggested to genetically overlap with SLE, and this may also be the case for many other immune related diseases (Alarcon-Segovia 2005). The association of the MHC with SLE has been more precisely defined, implicating loci in all three classes of the MHC as significant in susceptibility to SLE including HLA-DRB1*0301, alleles of *HLA-DRA* and *TRIM31* (Rioux 2009). Looking at individual diseases such as SLE in the context of genetic associations could lead the way to personalised medicine and targeted therapy for these diseases (Rhodes and Vyse 2010), but also for other immune related conditions where overlapping susceptibilities have been observed.

In general, associations of the MHC with common disease tend to be at a haplotype level. Functional variants are less commonly described and while GWAS and other studies consistently show the MHC as being associated with a particular susceptibility, mechanisms for these associations are far from clear. For example, the particular variants of HLA-DRB1 were linked to susceptibility to RA in the 1970s. A shared epitope theory suggested that the functional variant was a common set of amino acids in the DRB1 binding groove shared across diseases that were associated with the *HLA-DRB1* allele (Gregersen 1987). However, though there is a strong and reproducible genetic association between RA and possession of this "shared epitope" it is neither necessary nor sufficient for disease incidence. The full mechanism of disease susceptibility is likely to involve several of the risk loci and remains unclear even in the "post-GWAS era" (Perricone 2011).

1.3.3 Structural Variants

Antigen presentation clearly plays a role in MHC associated complex disease. Most of the genes encoding antigen presenting molecules are found within the class II region that is often implicated in

disease susceptibility. The high genetic diversity in this region in particular allows for differential peptide binding to the antigenic molecules and was suggested as a mechanism for disease susceptibility several years before GWAS studies highlighted its likely strong effect (Todd 1988). In an example of this, MHC molecules that have been associated with MS (HLA-DR) have been shown to have a particular structure that favours the presentation of shorter parts of antigenic peptides in contrast to most other classical MHC presenting molecules which present longer peptides. This may induce disease-causing cross-reactivity by reacting to presentation of an antigen that is not an immunogen (Jones 2006). Possession of the HLA-B*27 haplotype and peptide handling has also been linked with ankylosing spondylitis incidence and pathogenesis. Polymorphisms in *ERAP1*, an aminopeptidase, also are found to only affect ankylosing spondylitis risk when they are found in conjunction with the HLA-B*27 allelle (Evans 2011).

1.3.4 Gene expression and the MHC

Since gene expression was defined as a heritable and variable trait, differing levels of gene expression in the MHC in relation to disease susceptibility have been an area of interest for study. In LCLs derived from 400 children with asthma. 206 different families showed that genes involved in the immune response demonstrated highly heritable expression patterns and most variation associated with this was found within the MHC (Figure 1.9) (Dixon 2007). This implied that gene expression levels may have as profound an impact on the susceptibility to disease as structural variation in the MHC (Dixon 2007). Other studies have also implicated gene expression as a phenotype in the context of disease, making the MHC an attractive target for further study (Cookson 2009).

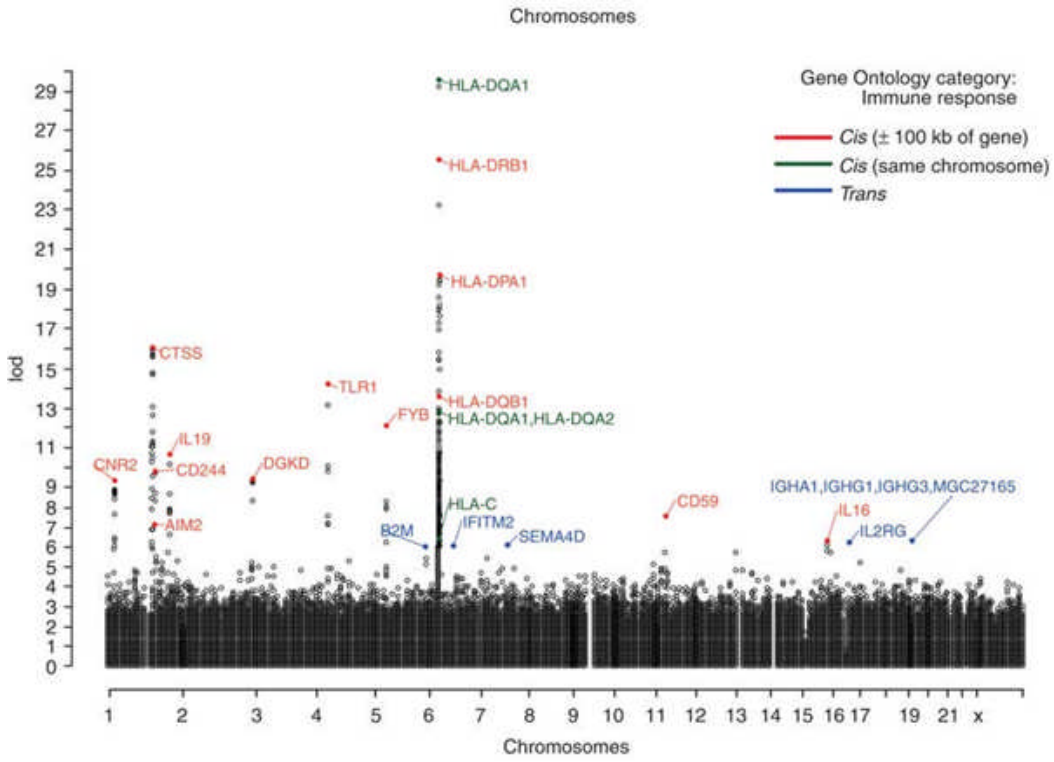


Figure 1.9: Gene Ontology Analysis identifies genes that are significantly enriched in heritable traits. Reproduced from (Dixon, 2007). Manhattan plot showing the logarithm of the odds of linkage (LOD) of SNPs associated with enriched genes involved in immune response. The MHC shows the highest number of associated SNPs.

For example, studies of coeliac disease have indicated that both altered peptide binding capacities and different levels of gene expression could be important in the presentation of a disease phenotype. The HLA-DQ region in the MHC has long been associated with coeliac disease and is thought to play a role through antigen presentation (Di Sabatino and Corazza 2009). Variation in the HLA-DQ region is neither necessary nor sufficient to cause disease. Many of the general population carry the risk allele, suggesting other loci must also be important, and other immune-related genes have been implicated in coeliac susceptibility (Hunt 2008). An eQTL analysis of variants affecting immune-gene expression showed several immune regulatory variants co-localised with coeliac risk loci (Dubois 2010). Functional variants have yet to be completely defined, however it is clear that complex disease will have several mechanisms which cause a disease phenotype, and that the interplay between gene expression and gene function will be crucial to the final phenotype.

1.3.5 Sequencing the MHC for eight different common disease associated haplotypes

In an effort to better understand the associations of ancestral haplotype with complex disease, the MHC Haplotype Project was initiated using LCLs homozygous for these haplotypes. The diversity of the MHC and the similar sequences of several of the genes within it have made defining one reference sequence for the MHC a challenge. Although the MHC was completely sequenced first in 1999, high levels of variation meant that further sequencing projects by the MHC sequencing consortium were necessary (Beck 1999). In 2004 two common ancestral haplotypes that are found in the consanguineous homozygous LCLs COX and PGF were sequenced completely (Stewart 2004). The PGF LCL was the reference sequence for the human genome sequence and contains the "7.1 ancestral haplotype", HLA-A3-B7-Cw7-DR15. This haplotype has been previously associated with protection from T1D, susceptibility to MS and SLE (Barcellos 2003, Larsen and Alper 2004). The COX LCL contains the "8.1 ancestral haplotype", HLA-A1-B8-Cw7-DR3, which has been linked to susceptibility to T1D, SLE, myasthenia gravis, common variable immunodeficiency and several other diseases including increased susceptibility to infectious disease (Price 1999). The full sequence of the MHC in the QBL homozygous LCL followed in 2006 (Traherne 2006). In 2008 the full sequence for eight haplotypes from homozygous LCLs became available through the efforts of the sequencing consortium, discovering over 44,000 variations between the 8 haplotypes (Horton 2008). An extensive map of the MHC showing LD and haplotypic structure has also been published, allowing tagging SNPs for specific haplotypes of the MHC to be identified for use in study of disease associations (de Bakker 2006).

1.3.6 A haplotype specific MHC gene expression array

Although RNA-seq is becoming more widely used for analysis of gene expression, it remains expensive, and may be unsuitable for the study of ASE due to mapping bias and lack of sensitivity for genes with low expression (Degner 2009). Despite eQTL analysis localising regulatory variation over the MHC these have yet to assess the extended haplotypic nature of the MHC in their design. Most focus on single regulatory variants rather than co-inherited extended haplotypes that have different expression properties over a combination of variants. How much an extended haplotype alters the

expression of genes on the MHC has not been fully addressed. Because of this a novel hybrid tiling/ splice junction array over the entire MHC has been developed in the Knight lab to produce a high resolution, strand-specific transcriptional map of the MHC (Vandiedonck 2011). Probes for this array were designed to be allele specific so that ASE could be accurately defined with no confounding results caused by variants within probes affecting probe hybridisation (Walter 2007). Genes showing ASE detected by this array will become candidate loci for areas harbouring genetic variants that modulate gene expression, and thus may possess interesting functional properties. Genes showing ASE that have previously been associated with disease may prove especially interesting. However, as the MHC is implicated so strongly in human disease, any gene showing variable expression is potentially important in terms of a disease or complex phenotypes. This gives an opportunity to investigate how gene expression affects susceptibility to and severity of disease.

Three LCLs homozygous across the classical MHC (hereafter referred to as "MHC homozygous") that had been previously sequenced by the MHC sequencing consortium were analysed using the custom array, PGF, COX and QBL. All three contained extended haplotypes that had been associated with autoimmune disease; the PGF and COX LCLs as previously described and the QBL LCL containing the "18.2 ancestral haplotype", HLA-A26-B18-Cw5-DR3-DQ2, which is associated with susceptibility to Grave's disease and T1D (Johansson 2003).

1.4 Aims of this thesis

The overall aim of this thesis is to analyse ASE within the MHC and associate the relationship with complex disease. Specific aims are listed below and are expanded upon in the introduction to each individual chapter:

1. **To validate the results of the MHC haplotype specific array in the analysis of allele specific gene expression.** Specifically, to confirm and validate the incidences of significant allele-specific gene expression seen between the LCLs analysed using qPCR (Chapter 3). Genes will be selected that show differential expression and re-analysed using a different method of gene-expression-level quantification to confirm the ASE between different haplotypes. These genes will then be investigated in primary peripheral blood cells to confirm the ASE is not a result of the EBV immortalisation of cells or other artefact (Chapter 3).

2. **To functionally characterise candidate loci within the MHC showing evidence of ASE.** Genes that show haplotypic expression both in LCLs and in primary cells will be investigated to localise the regulatory variants likely to be important in the mediation of the ASE (Chapters 4 and 5). Where there is little known about the function of these genes, especially regarding their expression levels, their function will be explored to hypothesise how their expression levels may be affecting cellular processes. Experiments, including ChIP-seq and bioinformatic analysis to determine their function will be attempted and the results analysed in the context of their differential expression (Chapter 4, 5 and 7).

3. **To analyse ASE in the context of human disease.** Where a previous disease association to the gene showing ASE is known through GWAS investigations, the likelihood of differential gene expression contributing to disease will be investigated (Chapters 4 and 5). Where no known disease association exists or where a known association seems unlikely to be due to ASE the contribution of differential gene expression to any pathogenic phenotype will be discussed.

4. **To assess the feasibility of using RNA-seq for the detection of ASE especially in regard to application for genes in the MHC.** As eQTL and GWAS results are analysed further the need for high depth sequencing to categorise regulatory and causal variants for gene expression to in turn be associated with complex traits will only increase. The MHC is well documented as being a susceptibility locus in many complex traits, particularly diseases, and will become increasingly important to study as further RNA-seq experiments are carried out. Currently mapping bias in RNA-seq and the highly polymorphic nature of the MHC suggest that a definitive analysis to confirm both the suitability of RNA-seq for gene expression analysis in the MHC and the ratios of ASE that can be called with confidence in the MHC and beyond would be a topical and insightful experiment to perform (Chapter 6).

Chapter 2 – Methods

This chapter includes details of cell culture and harvesting of RNA and DNA as well as the collection and processing of a large primary peripheral blood cell collection through the establishment of a healthy volunteer cohort. Specific experimental techniques such as Western blotting, Fluorescent Activated Cell Sorting (FACS) and cloning are described. Techniques where specific services or data analysis has been employed are also detailed. In this chapter, Sigma is the default supplier. Other manufacturers of chemicals, consumables and instruments are detailed, and full addresses can be found in the Key to Manufacturers (page xi).

2.1 Cell Culture

2.1.1 LCLs

Epstein-Barr Virus immortalised B cells from the HapMap cell line collection (CEU and YRI lines) (TIHMP 2003) were obtained from Coriell Cell Repositories. COX, APD, MANN and DBB were obtained from The International Histocompatibility Working Group (http://www.ihwg.org/index.html) (IHW, ref 0922, 9291, 9050 and 9052) PGF and QBL were purchased from ECACC (www.hpacultures.org.uk/collections/ecacc.jsp) (ref 94050342 and 94070713).

Maintenance

LCLs were initially cultured in RPMI-1640 supplemented with 20% FCS (PAA) and 2mM L-Glutamine. Once cells were established, FCS concentration was reduced to 10%. Cells were incubated at 37˚C with 5% CO_2 in a humidified environment. The density of the cells was maintained between $5x10^5$ and $8x10^5$ cells/ml. Frozen cells were stored in aliquots of 1ml containing 5×10^6 cells in 90% FCS 10% DMSO . They were kept in long term storage in liquid nitrogen, having been first brought down to -80˚C slowly in a sealed polystyrene box.

2.1.2 HEK cells

Human Embryonic Kidney (HEK) 293 cells were obtained through the generosity of Dr Simon Fisher (Wellcome Trust Centre for Human Genetics, Oxford, UK).

Maintenance

HEK cells were maintained in DMEM supplemented with 10% FCS and 2mM L-Glutamine. Cells were incubated at 37˚C with 5% CO_2 in a humidified environment. Cells were detached from coated flask walls with trypsin when confluent and were resuspended at a density of $5x10^4$cells/ml.

A haemocytometer was used where the sample (Trypan blue stain, PBS, and cells to a 10µl volume) was added to each chamber and live cells counted under a microscope. The results of the two chambers were averaged to give the cell concentration in suspension after correction for dilution factors. Alternatively, the Countess machine (Invitrogen) was used following manufacturer's instructions. Briefly, a cell suspension was combined with Trypan blue and electronically counted using specialist counting chamber slides (Invitrogen).

2.1.3 Transfection

HEK cells were transfected using a lipofection-based protocol. Lipofectamine LTX with the Plus reagent (Invitrogen) was used according to manufacturer's instructions. Cells were plated in 6-well plates 24 hours before transfection and 2.5µg endo-free prep plasmid DNA was used per well with 7µl of lipofectamine and 2µl of Plus reagent. Opti-MEM (Gibco) was used as the serum-free media for transfection. Control cells were treated with Opti-MEM, and lipofectamine LTX with Plus reagent only. 24 hours following transfection cell lysates were harvested. All other procedures were carried out as for LCLs.

2.2 RNA and DNA and cell lysate isolation

2.2.1 Harvesting RNA

Total RNA was isolated from the cell culture mid-log phase, 24 hours following feeding. Alternatively, in the volunteer cohort, total RNA was isolated from purified PBMCs, B cells or monocytes.

Cell culture

For each biological replicate (BR), a flask containing 200 million cells in 200ml media was split into unstimulated (RNA harvested at time point=0 hours) and stimulated samples (RNA harvested at time point=6 hours) (stimuli described below). Cells were centrifuged at 500 x g for 5 minutes, washed in PBS, lysed in RLT buffer (Qiagen) and stored at -80˚C or immediately processed for RNA extraction. Cell lysate samples stored in RLT Buffer were homogenised using a 19 gauge needle. RNA was extracted using the Qiagen RNeasy Midi kit (Qiagen) as described by the manufacturer. Briefly, ethanol was added to lysed and homogenized samples creating binding conditions for the RNeasy column. An on-column DNase digestion was carried out, and RNA was eluted using RNase free water. RNA concentration was determined by spectrophotometry using a NanoDrop ND-1000 (Thermo Scientific) at 260nm and 280nm wavelengths. RNA was aliquoted and stored at -80˚C.

Stimulation with PMA/ionomycin

PMA/ionomycin stimulation was carried out before harvesting for RNA where appropriate. Cells were incubated for 6h at 37˚C with 5% CO_2 in a humidified environment following addition of PMA (to a final concentration of 200nM) and ionomycin (to a final concentration of 125nM).

Volunteer cohorts

PBMCs were isolated from the volunteer blood samples as described below. The samples stored in RLT were homogenised using a Qiashredder (Qiagen) and RNA was purified using the RNeasy Mini Kit (Qiagen) following the same principles as the Midi kit described previously.

2.2.2 Harvesting gDNA

Cell culture

The Puregene kit for purification of genomic DNA (Qiagen) was used to isolate gDNA from 10 million cells according to the manufacturer's instructions. Briefly, cells were resuspended in lysis solution and were treated with RNase. Proteins were precipitated out of solution and DNA was precipitated using alcohol. The DNA was dried and then resuspended in TE or water.

Volunteer cohort

gDNA extraction was performed by Miss Seiko Makino. DNA was extracted using the Gentra

PureGene extraction kit (Qiagen) from 1-2ml of whole blood. The DNA was quantified by PicoGreen dsDNA quantification assay (Invitrogen).

2.2.3 Isolation of cell and nuclear lysates

Isolation of cell lysate

Cultured cells were washed with PBS, and lysed at a density of 5 million cells/ml (100mM Tris-HCl, 150mM NaCl, 10mM EDTA, 0.2% Triton-X 100 , 1x Protease Inhibition Cocktail (Roche), 1% PMSF) for 10 minutes. The resulting lysate was collected following centrifugation (16000xg, 4˚C) and stored at -80˚C.

Isolation of Nuclear Lysate

Isolation of nuclear lysate was performed based on a previously published method (Schreiber 1989). Cultured cells were washed with cold PBS and lysed using 1ml lysis buffer A (10mM Hepes pH8, 10mM KCl, 0.1mM EDTA, 0.1mM EGTA, 1x Protease Inhibition Cocktail (Roche)) on ice for 10 minutes. 67µl 10% NP-40 (Roche) detergent was added and the sample was vortexed and incubated on ice for a further 5 minutes. Nuclei were collected by centrifugation (16000xg, 4˚C) washed in lysis buffer and resuspended in three volumes of high-salt cold Nuclear-Lysate-buffer B (20mM Hepes pH8, 0.4mM KCl, 1mM EDTA, 1mM EGTA, 1x Protease Inhibition Cocktail (Roche)). Concentration was determined using a NanoDrop ND-1000 Spectrophotometer (Thermo Scientific).

2.3 Applications of RNA and gene expression study

2.3.1 cDNA synthesis

The Superscript III (SSIII) Reverse Transcriptase kit (Invitrogen) was used. In each case, 1-5µg of RNA in a final volume of 8µl (balance water) was reverse transcribed using random hexamers. RNA, primers and dNTPs were incubated at 65˚C to denature RNA secondary structure, before addition of 10XPCR Buffer, 50mM $MgCl_2$, DTT, RNase OUT and Super Script III (SSIII). RT negative samples had water added instead of SSIII. The samples were incubated at 50˚C for reverse transcription, and then inactivated by heating to 85˚C. RNase H was added to the samples and they were incubated at 37˚C. The final reaction volume of 21µl was diluted with H_20 to give 100µl final volume.

2.3.2 mRNA-seq

Total RNA was submitted to the High-throughput Genomics core at the WTCHG for library preparation. Quality control (QC) was carried out using the Qubit (Invitrogen) and the Lab 901 Tapestation (Agilent) to confirm concentration and RNA quality (>9 RNA Integrity Number). RNA was reverse transcribed using OligodT primers and the resulting cDNA was fragmented using chemical hydrolysis. PCR amplification was carried out, and overly large fragments were removed through size selection. Final library QC was carried out using the Qubit (Invitrogen), Bioanalyser 2100 (Agilent) and a qPCR was carried out to determine the exact concentration. Sequencing was carried out using the GAIIx platform before alignment to the Human Hg37 sequence using Stampy (Lunter and Goodson 2011). Reference-free mapping was carried out using Cortex (Iqbal 2012).

2.3.3 Quantitative Polymerase Chain Reaction (qPCR)

Icycler qPCR for cell line gene expression analysis

The assay was performed using a Biorad Icycler using 96-well optical plates (BioRad). Each PCR reaction used the SYBR green mastermix (BioRad), forward and reverse primer at 250nM and between 25-100ng of cDNA template giving a final reaction volume of 25µl. Primer sequences can be found in Table A.1. Two or three technical replicates for each sample were performed to reduce error. Cycling conditions are specified in Section A.2.

A melting curve was determined for all reactions and analysed to ensure the presence of only one amplification product. Analysis of the reactions was carried out using Biorad IQ5 software. Relative gene transcript levels were determined using the ΔΔCt method.

CFCX qPCR for volunteer study gene expression analysis

The assay was performed on the CFCX cycler (BioRad) using 96-well white low-background plates (BioRad). Each PCR reaction used the SYBR green mastermix (BioRad), a forward and reverse primer at 250nM and 2 cDNA template giving a final reaction volume of 12.5µl. Reactions were performed in duplicate to minimise errors. Primer sequences can be found in Table A.1. Cycling conditions are specified in Section A.2.

The CFCX software (BioRad) was used to analyse the data and determine melting curves for the PCR products. Relative gene transcript levels were determined using the ΔΔCt method.

2.3.4 Rapid Amplification of cDNA Ends (RACE) PCR

The GeneRacer kit with Superscript III and TOPO cloning kit (Invitrogen) was used with minor modifications. Briefly, 1µg COX RNA was treated by dephosphorylating, decapping and ligating Generacer RNA Oligo to the mRNA before reverse transcription (for 5' cDNA preparation) or beginning at the reverse transcription step (for 3' cDNA preparation). Touchdown PCR was carried out using 2µl of the RT template; cycling parameters and primer sequences can be found in Section A.2 and Table A.2. A re-amplification using 1µl of the initial reaction and the same cycling parameters was carried out. The final products were analysed by agarose-gel electrophoresis and visible bands were excised. PCR products were purified by gel extraction using a PCR gel extraction kit (Qiagen). The TOPO TA cloning kit included in the Generacer kit was used to clone the PCR products. Briefly, PCR products were ligated into the cloning vector (pCR®4-TOPO®) and used to transform competent TOP10 *E. Coli* which were left to recover in Super Optimal broth with Catabolite repression then grown on Lysogeny Broth (LB) agar plates containing Carbenicillin at 50mg/ml overnight at 37˚C.

Colonies were picked to inoculate mini-cultures which were incubated overnight at 37˚C with shaking in selective LB media. Plasmids were purified using a Spin Miniprep kit (Qiagen) according to the manufacturer's instructions. Briefly, plasmid DNA is isolated by size and is bound to a column, washed and then eluted in a low salt buffer (TE). Miniprep DNA concentration was determined by NanoDrop as described previously.

2.4 Sanger sequencing

2.4.1 Sequencing PCR

300ng miniprep plasmid DNA was used in the sequencing PCR reaction. The reaction was set up using 1µl of BigDye premix v3.1 with of 1.5µl BigDye v3.1 Sequencing Buffer (all Invitrogen) and

3.2pmol of forward or reverse primer in a total volume of 10µl. Cycling parameters can be found in the Appendix, Section A.2.

2.4.2 Ethanol clean-up

1µl Sodium Acetate pH5.2, 1µl 0.125M EDTA and 25µl 100% ethanol were added to each sequencing reaction and incubated removed from light at room temperature for 15 minutes. The precipitation products were centrifuged at 3000x g, 4˚C, for 30 minutes. The supernatant was removed and each pellet was washed with freshly made 70% ethanol and was centrifuged at 1650x g, 4˚C, for 15 minutes. The final precipitated product pellet was air dried and submitted to Source Bioscience (Oxford) or the Zoology Department (University of Oxford) for sequencing. Resulting files were aligned with the web programme ClustalW allignments and Jalview (zeon.well.ox.ac.uk/git-bin/clustalw.cgi) and sequence quality was checked using BioEdit (Hall 1999).

2.4.3 Sanger sequencing for gDNA genotyping

30ng gDNA was amplified using a 10µl PCR reaction containing 1µl 10x Buffer (Invitogen) 3mM $MgCl_2$, 0.2mM dNTP (Invitrogen), 250nM primer (each direction) and 1U Platinum Taq (Invitrogen). Cycling conditions are described in Section A.2 and primer sequences can be found in Table A.4. 2.5µl of the PCR reaction were cleaned using 1µl ExoSAP (GE Healthcare) incubated at 37°c for 15 minutes then inactivated at 85°c for 15 minutes. The full 3.5µl was used for the sequencing PCR using cycling conditions as previously described, followed by an ethanol precipitation cleanup also as previously described.

2.5 Western Blotting

Protein concentration of cell lysate samples was ascertained by bicinchoninic acid assay (Thermo Scientific). Briefly, known standards and samples are incubated with the assay reagents and protein concentration is quantified using the SoftMax plate reader (Molecular Devices), reading at 562nm. 10µg protein was boiled with Reducing and Sample buffers (Invitrogen) for five minutes in a total volume of 20µl. Samples and protein pre-stained protein standards (Invitrogen) were run on a

NuPage Novex Precast gel (4-12%) (Invitrogen) in MOPS buffer (Invitrogen) for 1 hour at 200V. Protein was transferred from the gel to a membrane (Millipore) for one hour, 30V using transfer buffer (Invitrogen). The membrane was washed in TBST and the blocked in 5% Milk (one hour). Primary antibodies were used at the concentration suggested by the manufacturer in 1% milk and were incubated overnight at 4˚C with shaking:

i) Rabbit Polyclonal α-ZFP57 (ab50944, Abcam)

ii) Rabbit Polyclonal α-KAP1 (ab10483, Abcam)

iii) Mouse monoclonal α-c-Myc (s1826, Clontech)

To detect the protein, the membrane was washed 3 times in TBST for 10 minutes and the secondary antibody Donkey α-Rabbit HRP conjugated (ab16284, Abcam) or Sheep α-Mouse HRP conjugated (ab6808, Abcam) was added at a dilution of 1:4000 for one hour in 1% milk at room temperature with shaking. The membrane was washed three times in TBST and the secondary antibody was detected using Novex ECL reagent (Invitrogen). The blot was visualised using film (Kodak) and the Compact X4 Automatic Processor (Xograph). If more than one primary antibody was used on one blot, the membrane was washed in TBST and then stripped for 15 minutes at room temperature using 5ml ReBlot Plus Stripping Solution (Millipore). The method then continues from the point of the membrane being blocked in 5% milk for one hour.

2.6 FACS

2.6.1 Staining for cell surface markers

50ml whole blood was used for isolation of PBMCs by ficoll gradient (see below). PBMCs were used either as a complete cell mixture or 10 million cells were separated using Magnetic-activating cell-sorting (MACS, Miltenyi) to isolate Monocytes and B-cells (see below). Samples of mixed PBMCs were surface stained to identify the different cell types using fluorescent-conjugated antibodies:

i) α-CD19-FITC for B-cells (eBioscience)

ii) α-CD14-PeCy7 for Monocytes (eBioscience)

iii) α-CD56-Pe for NK cells (Miltenyi)

iv) α-CD4-Pe and α-CD8-APC for T cells (eBioscience)

2.6.2 Staining for ZFP57 expression

One million PBMCS per replicate, and all available purified B-cells and Monocytes per replicate were fixed and permeabilised for detection of ZFP57 expression. The primary α-ZFP57 antibody (ab50944, Abcam) was detected by a F(ab)2 Donkey α-Rabbit IgG PE-conjugate (eBioscience). Cells were fixed using 1% paraformaldehyde (PFA) and washed with PBS containing 1% FBS and EDTA. The samples were blocked using Flow-Cytometry Staining-Buffer (eBioscience) at room temperature for 15 minutes before being permeabilised with 0.1% Triton-X 100 in Flow Cytometry Staining Buffer with the primary antibody in a 1:100 dilution supplemented with 5% normal Rabbit Serum (Invitrogen). The samples were incubated at room temperature for 30 minutes before being washed with PBS containing 5% FBS and EDTA. The secondary antibody was introduced in a 1:200 dilution in Flow-Cytometry Staining-Buffer supplemented with 0.1% Triton-X 100 and was incubated at room temperature for 20 minutes. Following antibody incubations the cells were washed three times in PBS with 5% FBS and EDTA and were finally resuspended in 500µl PBS with EDTA and 0.1% PFA for analysis. Alternatively, following cell surface staining the Fix and Perm Kit (Invitrogen) was used according to the kit protocol, where the step with addition of reagent B with the antibody was performed with both the primary and secondary antibodies.

2.6.3 Analysis of protein expression using FACS

Protein expression was then analysed using the 9-colour equipped flow-cytometry machine CyAn-ADP (Dako) where appropriate gates were introduced to prevent dead cells or cell fragments from being included in the analysis. A minimum of 10,000 events were included in the analysis. Samples incubated with secondary antibody only were used as negative controls.

2.7 Healthy Volunteer Cohort Study

2.7.1 Recruitment

Existing Healthy Volunteer Samples (denoted cohort 1, n=96)

PBMCs were isolated using a Ficoll-Paque (GE Healthcare) gradient and were cultured for 24 hours before RNA extraction. Samples were collected from healthy volunteers recruited by Dr Ben Fairfax and Dr Fred Vannberg (Fairfax 2009) and were established and available to the lab prior to 2008, the commencement of this DPhil project.

New Healthy Volunteer Samples (denoted cohort 2, n=288)

Ethical approval was obtained from the Oxfordshire Research Ethics Committee (ref: 06/Q1605/55) and informed written consent was obtained from each donor. Volunteers were recruited over a 5 month period in the Oxfordshire area. 50ml whole blood was taken in 10ml EDTA vacutainer blood tubes (Becton Dickinson) from 288 individuals of European ancestry who were self-reported non-smokers. An additional 2ml of blood was taken in a 5ml EDTA vacutainer blood tubes (Becton Dickinson) to enable gDNA purification from whole blood. Sex, age and ethnicity were noted and the samples were made anonymous. The age of volunteers ranged from 18-62 years with a median age of 29 years and an average age of 33.1 years (Male 31.1 years, Female 34.6 years).

2.7.2 Cell purification

PBMC purification

Blood samples were taken in the morning, (7.30am-11am). PBMCs were purified as soon as possible after collection using a Ficoll-Paque (GE Healthcare) gradient. Blood was diluted 1:2 in Hanks Buffered Saline Solution (HBSS) without Mg^{2+} or Ca^{2+} and layered on Ficoll-Paque in falcon tubes. The falcon tubes were then centrifuged at 400g for 30min at room temperature with no acceleration or break. The PBMC "buffy coat" layer was aspirated using a Pasteur pipette and was washed twice in HBSS. PBMCs were resuspended in RLT buffer and were stored at -80˚C.

Cell sorting

Magnetic-activating cell-sorting methods (MACS, Miltenyi) were used to positively separate CD14 positive cells (monocyte population) and CD19 positive cells (B cell population) according to the

manufacturer's instructions. Briefly, the cells were incubated with magnetic beads pre-coated with a relevant antibody allowing cells expressing particular cell surface markers to be separated once bound to the beads over a magnetic column. Following separation the monocytes and B cells were resuspended in RLT buffer and stored at -80˚C.

2.7.3 Genotyping and haplotype information

Genome-wide genotyping for the volunteer samples (cohort 2) was determined using the Illumina Infinium high-density genotyping bead arrays (Illumina HumanOmniExpress-12v1.0 Beadchips, NCBI36 Build) (Illumina) resulting in determination of 733,202 genetic variants. Genotyping was performed by the Wellcome Trust Centre High Through-put Genetics Group.

Following standard QC (SNPs showing abnormal chip intensities and cluster plots, or those that deviated from Hardy-Weinberg were removed) 651,210 markers were left for analysis. The average of the sample call rates was 99.72%. PLINK, the whole genome association analysis toolset was used for QC and analysis (Purcell 2007). Genotyping results correctly identified the sex of all volunteers.

In order to determine the genetic stratification of our population multidimensional scaling was calculated with the published HapMap populations (Frazer 2007). The SNP call rate for minor allele frequency (MAF) at > 5% was set at > 98% and that for MAF between 1 to 5% was set at > 99%. 5 individuals were excluded due to biased heterozygosity caused by a low genotyping rate (sample call rate 94%) or mixed ethnicity origin. Two samples identified as having a familial relationship by the pair-wise identity by descent test led to the random exclusion of one from further analysis.

Imputation of the *ZFP57* and the *HLA-DQ* regions was performed using IMPUTE2 (Howie 2009) and data from the 1000 genomes project (www.1000genomes.org). Co-ordinates (Hg19) used were chr6:26400239-32199517 for the *ZFP57* region and chr6:29600054-35639688 for the *HLA-DQ* region.

Haplotype Imputation

2-digit and 4-digit HLA haplotypes occurring in the healthy volunteer cohort of 288 individuals were inferred from the known genotypes. A set of informative SNPs that had been genotyped in the cohort were selected and used to define known haplotypes in a training data set of around 3000

samples before being used to define the HLA-haplotypes of the volunteer cohort (Leslie 2008). This was performed by Dr Stephen Leslie and Dr Alexander Dilthey.

2.7.4 Gene expression and eQTL analysis

Gene expression array

200ng total RNA from monocyte and B-cells from each volunteer sample was submitted for gene expression array Illumina HumanHT-12 v4 BeadChip platform (Illumina) with 48,804 probes. This was performed by the Sanger Institute Microarray Service. Background levels in the data were adjusted using the R packages lumi and limma by the Sanger Institute DNA Services. Raw data was transformed and normalised by the robust spline normalisation method (RSN) by Dr. Jayachandran Radhakrishnan. All probes found to map to more than one location (using hg18 BLAST search) were removed from the analysis. 29022 probes were used to define normalised expression data and of these 26766 uniquely mapped to 19347 RefSeq genes and 2256 uniquely mapped to Unigene transcripts.

eQTL data analysis

eQTL analysis was performed using PLINK association tests following a linear model. Significance at a genome-wide level was taken to be $p<5x10^{7}$ for local likely cis-acting associations (WTCCC 2007). *HLA-DQB1* expression association between the two separate groups was analysed with a "case-control" format using a χ^2 test.

Haplotype construction over genomic areas of interest was carried out using Haploview (Barrett 2005) with the haplotype blocks defined using the confidence interval definition. Manhattan plots were also generated using Haploview.

Recombination plots were constructed using a script in R available from the BROAD Institute (http://www.broadinstitute.org/files/shared/diabetes/scandinavs/assocplot.R) (Saxena 2007).

2.8 Chromatin Immunoprecipitation (ChIP)

2.8.1 Cross-linking material

LCLs were cultured in suspension in a 650ml flask as previously described to a final concentration of 50-100 million cells in 100ml culture volume. DNA and protein were cross-linked by addition of 10ml formaldehyde buffer (0.1M NaCl, 1mM EDTA, 0.5mM EGTA, 50mM HEPES pH8, 11% formaldehyde) directly to suspension cell culture medium to give a final formaldehyde concentration of 1%. The suspension was incubated for 15 minutes as room temperature with rocking. The cross-linking reaction was quenched by addition of glycine to 0.125M concentration. Cells were pelleted, 500x g, 4˚C, 10 minutes and washed twice in cold PBS-10% FCS. Pellets were stored at -80˚C or processed immediately for sonication.

2.8.2 Nuclei Preparation

10ml of Lysis buffer 1 (50mM Hepes-KOH, pH 7.5, 140mM NaCl, 1mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100, 1x Complete Protease Inhibitors (Roche)) was added to 50-100 million cells and cells were incubated at 4˚C for 10 minutes with rocking. Cells were pelleted at 1000xg (4˚C, 5 minutes), and were resuspended in Lysis Buffer 2 (10mM Tris-HCl, pH 8.0, 200mM NaCl, 1mM EDTA, pH 8.0, 0.5mM EGTA, pH 8.0, 1x Complete Protease Inhibitors (Roche)). Cells were again incubated for 10 minutes at 4˚C. Nuclei were pelleted at 1000xg (4˚C, 5 minutes), then washed twice in Shearing Buffer (1% Sodium Dodecyl Sulfate (SDS), 10mM EDTA and 50mM Tris, pH 8.1, 1x Complete Protease Inhibitors (Roche)). Nuclei were finally resuspended in Shearing Buffer for sonication (concentration $3x10^6$cells/ml).

2.8.3 Sonication

Sonication was performed using a Bioruptor™ Twin sonication device (Diagenode) with 1.5ml tubes according to manufacturer's instructions. 12 Cycles of 30 seconds on, 30 seconds off at High Intensity (320W) were used on a volume of 300µl per tube, employing the controlled cooling system to ensure the sample remained at 4˚C. Following the initial sonication, the samples were adjusted to 0.5% Sarkosyl and were incubated at room temperature for 10 minutes with rocking. The samples

were centrifuged at 10000x g, 4˚C for 10 minutes, in order to remove cell debris. Sonicated chromatin was quantified using a nanodrop as previously described and stored at -80˚C or used immediately for ChIP. Fragment size was estimated by visualisation on an agarose gel following cross-link reversal overnight at 65˚C and phenol-chloroform extraction using Phase Lock Gel Light 2ml tubes (5Prime) and was precipitated using ethanol.

2.8.4 Immunoprecipitation and DNA purification

Sheep-raised anti-Rabbit Dynabeads M-280 (Dynal) were incubated overnight at 4˚C with the antibody of interest (raised in rabbit) or alone for a no-antibody control. Unbound antibody was removed by washing the Dynabeads with PBS supplemented with BSA at 5mg/ml. Incubations were set up using 200µg sheared chromatin, 400µl IP Buffer (10mM Tris HCl pH 8, 1mM EDTA, 2% Triton X-100, 0.2% sodium deoxycholate, 2x complete PI, 10µg/ml pepstatin) made up to 750µl using Tris-EDTA. ChIP samples were incubated overnight at 4˚C with rocking. Before washing, input chromatin was removed from the no-antibody control. Beads were washed eight times with Radio-Immunoprecipitation assay buffer (10mM Tris HCl pH 8, 1mM EDTA, 1% NP-40 (Roche), 0.7% sodium deoxycholate , 0.5M LiCl , 1x complete PI (Roche)) and once with TE. They were then resuspended in elution buffer (10mM Tris pH 8, 1mM EDTA, 1% SDS) and incubated at 65˚C for 16 minutes, vortexing the tubes every 2 minutes. Samples were centrifuged and the supernatant was removed to a fresh tube. Crosslinks were reversed overnight at 65˚C and samples were then treated with Proteinase K (Roche) in TE buffer for 2 hours at 55˚C and RNase A (Roche) for 2 hours at 37˚C. DNA was extracted using phenol-chlorophorm with Phase Lock Gel Light 2ml tubes (5Prime) and was precipitated using ethanol.

qPCR for ChIP was performed on the CFCX machine (BioRad) as previously described with 18S used as a normalising control for background and levels expressed relative to those observed with ChIP input DNA (De Gobbi 2006). Primer sequences can be found in Table A.5. Cycling conditions can be found in Section A.2.

2.8.5 ChIP-seq sequencing

ChIP-seq material prepared as described was submitted to the High-throughput Genomics Group (WTCHG) for sequencing. For each cell line antibody-incubated material was submitted with an input control for comparison. Universal adaptors for amplification were ligated to the material and amplification was carried out. Size selection was performed to remove fragments of length greater than 500bp. Library QC was carried out using the Qubit (Invitrogen), Bioanalyser 2100 (Agilent) and a qPCR was performed to assess the final concentration. Single-lane, single-end sequencing was performed on the GAIIx platform.

2.8.6 ChIP-seq Data analysis

Sequence alignment was carried out using Bowtie (http://bowtie-bio.sourceforge.net/index/shtml) (Langmead 2009) to the Human reference sequence (Hg37). Binding sites for the TFs analysed were defined using 3 separate algorithms, MACS (Zhang 2008), PeakSeq (Rozowsky 2009) and the SPP pipeline (Kharchenko 2008). All were followed in a standard manner using default settings, and defining the likely size of sonicated fragments to be 1000bp. An FDR of 5% and a more stringent cut-off of 1% were used to identify the likely true peaks. Binding-site detection was performed by Dr Stephen Sansom and Dr Andreas Heger.

2.9 Cloning

2.9.1 Amplification

ZFP57 was amplified from COX cDNA (50ng) using a PCR reaction (1X Hi-Fi Buffer, 3mM $MgS0_4$, 0.2mM dNTP, 250nM each primer, 1U High-Fidelity Polymerase (All Invitrogen)) where primers were designed with a NotI site at the 5' end of the gene and a KpnI site at the 3'end. Primer sequences can be found in Table A.3 and cycling conditions can be found in Section A.2. PCR reaction products were purified using gel extraction. Single bands of around 1650bp, the length of the *ZFP57* gene, were excised using a sterile scalpel. DNA was extracted from the gel bands using a gel extraction kit (Qiagen) according to the manufacturer's instructions. Briefly, gel fragments were solubilised in the

presence of a DNA-binding buffer. DNA was isolated on a column and washed, then eluted in Tris elution buffer. Purified gel fragments were incubated with Platinum Taq (Invitrogen) at 72˚C for 15 minutes to create T-overhangs at the end of the fragments to enable TOPO cloning.

2.9.2 TOPO cloning

Purified gel fragments containing the PCR product of *ZFP57* DNA were inserted into the TOPO vector pCR®II (Invitrogen) and the associated cloning kit according to the instructions of the kit unless stated otherwise. The ligation reaction used 0.5µl pCR®II TOPO vector mastermix, 3µl PCR product, 0.5µl Salt solution and was incubated at room temperature for 5 minutes. 2.5µl of the ligation reaction was added to 50µl of Alpha Select Bronze Efficiency chemically competent *E. Coli,* (Bioline) and incubated on ice for 30 minutes. Following this incubation the cells were heat shocked for 30 seconds at 42˚C before being returned to the ice for 2 minutes. 250µl SOC (Invitrogen) was added and the cells were incubated at 37˚C with shaking for 1 hour. After this incubation they were plated onto selective LB agar plates supplemented with Kanamycin at 50mg/ml and incubated at 37˚C overnight.

The following day, single isolated colonies were picked and grown in 3ml of LB media supplemented with Kanamycin at 50mg/ml overnight at 37˚C with shaking. Mini-preps (Qiagen) were carried out on 2ml of the culture media according to the manufacturer's instructions. Briefly, *E. Coli* were pelleted, resuspended and lysed. Plasmid DNA was captured using a DNA binding column and was washed with ethanol and eluted in Tris elution buffer. Mini-prep DNA was analysed using restriction digests to detect plasmids with the correct *ZFP57* insertion. Sanger sequencing was used to confirm the sequence as described.

2.9.3 Cloning into the pML5-Myc vector

pML5-Myc vector and *ZFP57* in TOPO were both digested using KpnI and NotI (1µl KpnI-HF, 1µl NotI-HF 3µl NEB Buffer 4 (all NEB Biolabs), 3µl 10X BSA, water and DNA made up to 25µl to give 30µl total reaction volume) for 3 hours at 37˚C. Following digestion reaction were run out on a 1% agarose gel for 1 hour at 80V. The linearised vector and the isolated ZFP57 fragment were purified using the gel

extraction kit (Qiagen) as previously described. The insert and plasmid were ligated in a 15µl reaction (50ng pML5-Myc linearised, 105ng ZFP57 insert, 1µl T4 ligase (NEB Biolabs), 1.5µl T4 ligase buffer (NEB Biolabs), made up to 15µl with water) at room temperature for 1 hour. 5µl of the ligation reaction was used to transform Alpha Select Bronze Efficiency chemically competent *E. Coli* as previously described. Cells were plated onto selective plates supplemented with 50mg/ml Carbenicillin and mini-prep cultures were grown in Carbenicillin selective LB media. Mini-preps and sequence analysis were carried out as previously described.

Midi-preps of the pML5-Myc-ZFP57 plasmid were carried out using a midi kit with endotoxin-free buffers (Qiagen) according to the manufacturer's instructions.

Chapter 3 - Validation of differential gene expression observed in MHC-homozygous LCLs

3.1 Introduction

3.1.1 Design of the MHC expression array

As outlined in Chapter 1, the MHC has been difficult to study in terms of gene expression using conventional microarrays and RNA-seq due to its polymorphism and repetitive structure (Walter 2007). The custom MHC tiling microarray (hereafter termed the "MHC array") aimed to overcome these issues by including alternate allele probes that can be individually selected in order to study gene expression for different haplotypes (Vandiedonck 2011). The MHC array was designed across 3.5Mb of chromosome 6 and covered 230 known genes, with 2755 exons covered in total. A tiling probe set of overlapping probes 25 nucleotides in length spanned the whole region on both strands, from chr6: 29,748,239–33,231,091 (hg18), comprising 398,626 probes in total. Additionally, a set of probes comprising 15,348 in total were designed against all junction sites for known or predicted splice events in order to monitor alternative splicing in the MHC class III region, giving an average of 12 probes at every known or predicted splice junction. Known SNPs according to the current dbSNP release at the time of design (dbSNP 126) were included, as alternate probes were designed for every SNP or segmental duplication. Finally, 10,572 probes shared with the Affymetrix Exon 1.0 ST array were included in the design to allow comparison of the same samples hybridised to another microarray platform.

3.1.2 Validation of Array detection of gene expression and splice junctions

LCLs homozygous across the classical MHC were cultured in standard conditions and RNA harvested from these LCLs was hybridised to the MHC array and the Affymetrix Exon 1.0 ST array in three biological replicates. As described in Chapter 1, these LCLs were the PGF, COX and QBL LCLs, chosen due to their possession of particular ancestral haplotypes (COX HLA-A1-B8-Cw7-DR3; PGF HLA-A3-B7-Cw7-DR15; and QBL HLA-A26-B18-Cw5-DR3-DQ2) (Horton et al.2008) and association with particular autoimmune conditions. Differences in signal intensity between replicates were low, giving a Pearson correlation coefficient value between 0.83 and 0.91. When differences in intensity were observed between the

LCLs, they were found on both array platforms which suggests a biological cause such as haplotypic variation rather than inconsistent hybridisation or sample preparation. Differences between signal intensity for the biological replicates for each cell line were extremely low, giving Pearson correlation coefficient value between 0.96 and 0.98.

In order to determine whether the alternate probes could be used to successfully discriminate between different haplotypic backgrounds, the signal intensity of PGF specific probes were compared with those of COX and QBL specific probes for the PGF RNA samples hybridised to the array. The intensity of the PGF specific probes was significantly higher, showing that allele specific probes were acting in a selective manner. Junction specific probes were analysed by comparing the relative quantities of two isoforms of both the *CD79A* and *CD79B* genes using both the array and RT-PCR. In both cases the ratio found between the longer and shorter isoforms was similar when detected with either method.

3.1.3 Identification of Transcriptionally Active Regions (TARs) in the MHC

Identifying regions of the human genome that are transcribed has been a problem in the past, with gene prediction programs not exhaustively identifying all transcribed regions and experimental validation using reverse transcription PCR being time consuming and labour intensive. In contrast to many other organisms, exons in mammals may be widely spaced between long introns. A tiling array allows this to be examined over both the sense and antisense strands to determine where transcription is occurring (Bertone 2004). The signal intensities were smoothed and then overlapping windows of 51 nucleotides were defined. Any windows with an above-median signal intensity were defined as transcriptionally active, then overlapping windows were merged together to show the TARs on the sequence. In total, 6% of the whole 3.5Mb sequence studied was actively transcribed, with 2% of transcription occurring on sense and antisense strands at the same location, 2% occurring only on the sense strand and 2% occurring only on the antisense strand (Vandiedonck 2011).

Known genes were determined to be transcribed when at least one TAR occurred in the intragenic sequence defined by Vega genes. Across the three LCLs analysed over 92% of the known Vega genes, and over 70% of known pseudogenes, were transcribed (Vandiedonck 2011). The proportion of known genes transcribed was consistent across the different haplotypes and a schematic showing expression

over the whole MHC in all three LCLs and as an aggregate is shown in Figure 3.1. A large number of the TARs however did not map to a location of a known gene, and these were split roughly equally between sections that appeared close to known genic regions, perhaps indicating previously unknown exons or other alternative elements, and sections that were located further away from known genes.

Haplotypic expression was defined where expression appeared only in one cell line; 9%, 4.6%, and 11.1% of the TARs on PGF, COX, and QBL sequences, respectively were expressed only in that cell line (Vandiedonck 2011). Quantitative differences between the LCLs expression were also interrogated. To accomplish this, the uniquely mapping probes for each cell line sequence were selected so that no confounding influence by SNPs could bias the analysis.

Genes showing the greatest fold-changes in expression between the different haplotypes were ranked starting from the greatest fold-change associated with one particular haplotype. This gave a gene list of genes expressed in a haplotype-specific manner. In total, 96 genes were deemed to show evidence of haplotype or ASE, the top 30 of which are shown below in Table 3.1. The most significant differentially expressed gene was *ZFP57* (Figure 3.2).

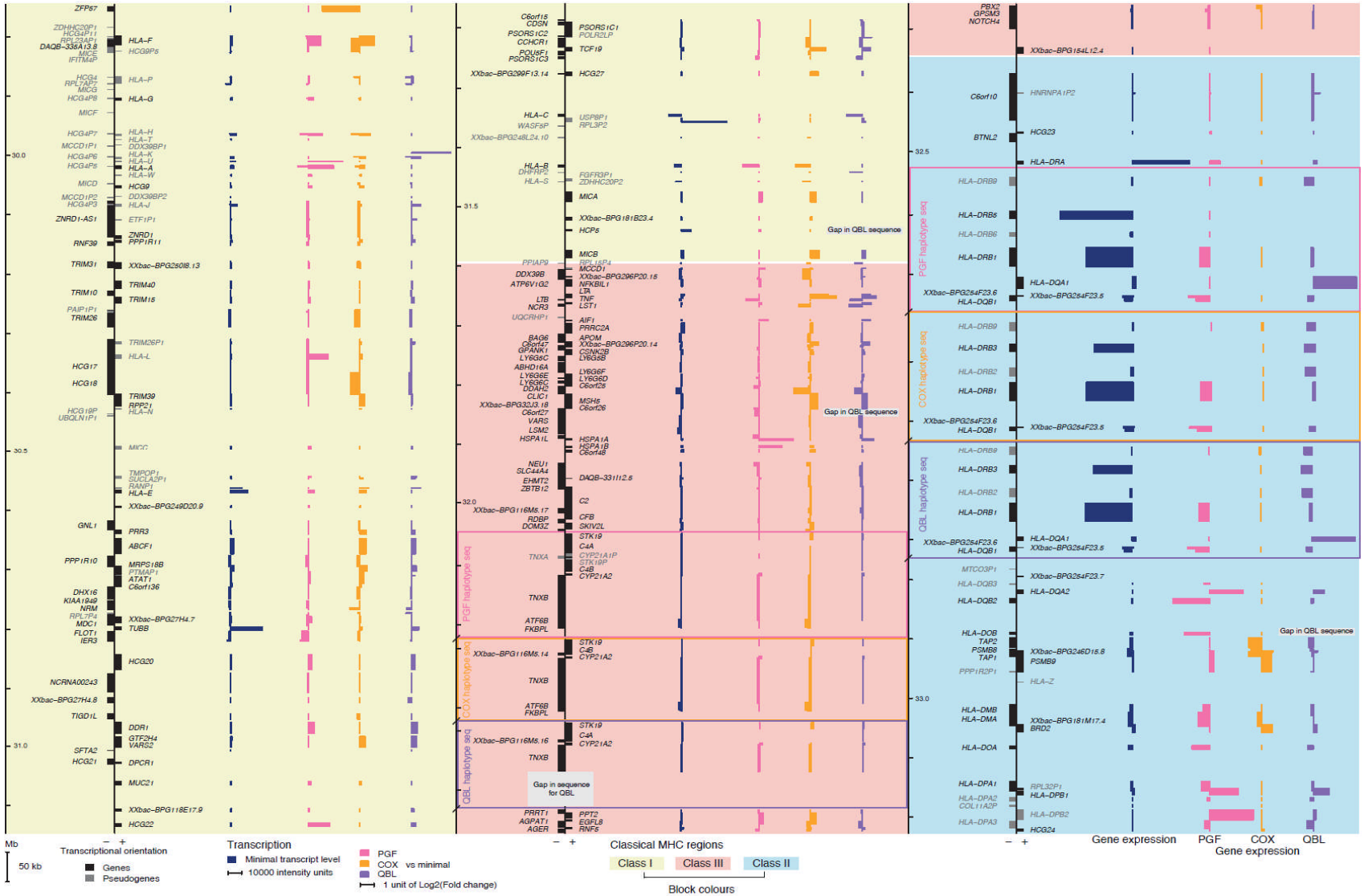


Figure 3.1: Haplotype-specific gene expression in the MHC. Reproduced from (Vandiedonck , 2011). Expression levels are shown in relation to the minimum expression seen for each gene (blue). Expression in each cell line is represented as pink (PGF), orange (COX) and purple (QB L). Expression on the anti-sense strand is indicated by a leftward bar and on the sense strand by a rightward bar.

Gene name	Class	$\log_2$ (fold change)			Adjusted *P*-value
		COX vs. PGF	QBL vs. PGF	QBL vs. COX	
ZFP57	I	2.77	0.00	−2.76	1.22×10^{-14}
HLA-DPB2[a]	II	−3.19	−3.02	0.17	2.89×10^{-12}
HLA-DQA2	II	−2.45	−1.62	0.82	1.91×10^{-11}
HLA-DQB2	II	−2.74	−2.58	0.16	3.21×10^{-11}
HLA-U [a]	I	−2.52	0.36	2.87	1.32×10^{-10}
TNF	III	1.90	1.03	−0.87	4.79×10^{-10}
HLA-DPB1	II	−2.08	−0.90	1.18	6.44×10^{-10}
RPL32P1 [a]	II	−1.52	−1.19	0.33	2.07×10^{-09}
HLA-B	I	−0.06	−1.19	−1.13	6.59×10^{-09}
HLA-A	I	−1.51	−1.86	−0.35	2.30×10^{-08}
HLA-L [a]	I	−1.29	−1.47	−0.18	2.30×10^{-08}
XXbac-BPG254F23.6	II	−1.59	−1.59	0.00	2.50×10^{-08}
HCG22	I	−1.56	−1.26	0.30	2.96×10^{-08}
XXbac-BPG254F23.5	II	−1.42	−1.61	−0.19	1.33×10^{-07}
LTA	III	1.32	0.57	−0.75	2.04×10^{-07}
NCR3	III	0.87	0.95	0.08	4.95×10^{-07}
HLA-F	I	0.15	−0.90	−1.05	4.95×10^{-07}
HLA-DOA	II	−1.32	−0.89	0.43	5.07×10^{-07}
TAP1	II	0.97	0.08	−0.89	6.86×10^{-07}
LTB	III	−0.95	−0.06	0.89	7.02×10^{-07}
LST1	III	−0.18	0.48	0.66	9.42×10^{-07}
DAQB-335A13.8	I	0.61	−0.02	−0.63	1.12×10^{-06}
TCF19	I	1.11	0.62	−0.49	1.49×10^{-06}
CLIC1	III	1.22	0.57	−0.66	1.49×10^{-06}
HLA-DMA	II	−0.57	−0.89	−0.33	3.52×10^{-06}
BRD2	II	0.78	0.27	−0.51	3.60×10^{-06}
NRM	I	0.77	0.39	−0.38	4.48×10^{-06}
HLA-C	I	0.05	1.11	1.06	4.98×10^{-06}
PSMB9	II	0.42	−0.29	−0.71	6.05×10^{-06}
HCG27	I	0.56	0.06	−0.50	7.01×10^{-06}

Table 3.1: Variation of gene expression between haplotypes. Reproduced from (Vandiedonck, 2011). Top 30 genes for which expression differed significantly between different haplotypes analysed using the MHC array. Each gene level intensity value is calculated from uniquely mapping probes to the relevant haplotype. Pseudogenes are notated by suffix 'a' after the gene name. The p-value was calculated using the Limma package (utilising linear and Bayesian methods) and was adjusted using the Benjamini-Hochberg correction to control the FDR.

DNase hypersensitivity was also investigated alongside gene expression with the MHC array (Freidin et al unpublished data), allowing the haplotypic differences in gene expression to be compared to variation seen between the LCLs in terms of DNase Hypersensitive Sites (DHS) (Figure 3.2). Where increased gene expression occurs it is often accompanied by an increase in DHSs as the chromatin has adopted a more relaxed conformation to allow the transcriptional machinery access to the DNA (Weintraub and Groudine 1976, Elgin 1988).

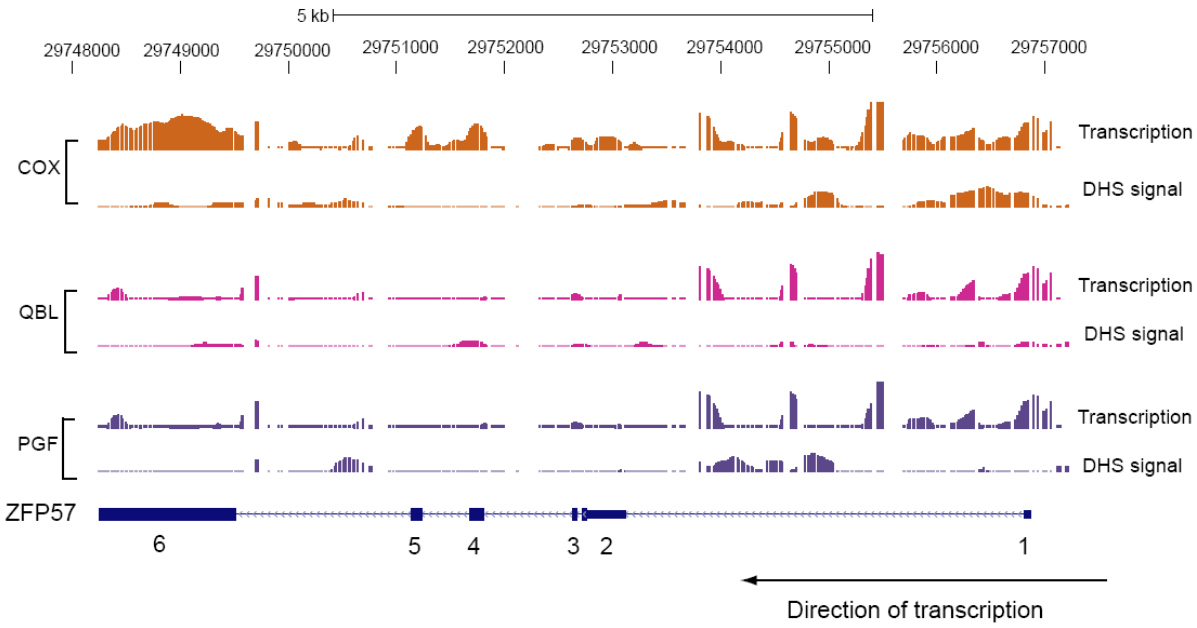


Figure 3.2: *ZFP57* expression and DNase Hypersensitive Sites. Array data showing transcription of *ZFP57* in COX, QBL and PGF cell lines and DHSs across the 3 cell lines. The transcription intensity signal generated by uniquely mapping probes to the region in each cell line is shown above the DHS signal seen in the same region. COX shows higher expression of *ZFP57* especially at the 3' end of the gene. Increased DHS signal is seen at the 5' end of *ZFP57* in COX suggesting a more open chromatin structure consistent with active transcription.

3.2 Aims

In this chapter I will seek to validate the haplotype-specific gene expression found using the MHC custom array for selected genes of interest. Expression in LCLs and primary cells will be analysed to confirm that the differential expression seen is not an artefact of cell immortalisation. Candidate genes that have shown differential expression across the MHC in other studies will also be selected for comparison and I will validate their differential expression in the same manner. Specific aims of the chapter are:

1. To validate the observed differences in gene expression between LCLs using an independent methodology, real-time quantitative PCR.

2. To investigate whether differences relate to underlying genetic variation and are present in primary cells by performing expression-quantitative-trait mapping in peripheral-blood mononuclear cells (PBMCs) from healthy volunteers.

3.3 Results

3.3.1 Selection of genes for validation

From the gene list generated by interrogation of the COX, QBL and PGF LCLs using the MHC array (Table 3.1), five genes were chosen for further analysis to validate the haplotype-specific gene expression in LCLs. The genes chosen for this purpose were the most significant genes showing differential expression across the haplotypes that were not pseudogenes. While the MHC array highlights several genes of potential interest, it seemed prudent to widen the search for functionally interesting genes with varied expression as only three extended haplotypes were analysed using the array. Because of this, four additional genes located in the MHC were selected for gene expression analysis due to their reported association with *cis* variants on eQTL analysis of genes involved in the immune response (Dixon 2007) (Table 3.2).

Gene Name	Rank from MHC array	Source of data
ZFP57	1	MHC array
HLA-DQA2	3	MHC array/ Dixon et al. 2007
HLA-DQB2	4	MHC array
TNF	6	MHC array
HLA-DPB1	7	MHC array
HLA-DQA1	-	Dixon et al. 2007
HLA-DQB1	-	Dixon et al. 2007
HLA-C	28	MHC array/ Dixon et al. 2007
HLA-DPA1	-	Dixon et al. 2007

Table 3.2: Genes chosen for validation of ASE. The top five non-pseudogenes shown to have differential gene expression between the 3 MHC homozygous cell lines using the MHC custom array were selected for analysis using qPCR. Four further genes with heritable SNP associations to gene expression of immune genes (Dixon 2007) were also selected for real-time qPCR analysis. Primer sequences for gene expression analysis can be found in Table A.1.

3.3.2 Choice of cell type for validation

Expression levels were quantified in the three homozygous LCLs used in the initial array-based study and in three further EBV transformed LCLs homozygous over the MHC for which full DNA sequence was available through the MHC Haplotype Project (Horton 2008): APD, DBB and MANN (See Table 3.3).

Haplotype	Length (bp)	Gaps	HLA-A	HLA-B	HLA-C	HLA-DQA1	HLA-DQB1	HLA-DRB1
PGF	4754829	0	A*03010101	B*070201	Cw*07020103	DQA1*010201	DQB1*0602	DRB1*150101
COX	4731878	0	A*01010101	B*080101	Cw*070101	DQA1*050101	DQB1*020101	DRB1*030101
QBL	4249272	5	A*260101	B*180101	Cw*050101	DQA1*050101	DQB1*020101	DRB1*030101
APD	4160965	16	A*01010101	–	–	–	–	–
DBB	2330101	28	A*02010101	–	Cw*06020101	DQA1*0201	DQB1*030302	DRB1*070101
MANN	4191014	10	A*290201	B*440301	Cw*160101	DQA1*0201	DQB1*0202	DRB1*070101

Table 3.3: Haplotype sequence length, number of gaps and HLA allele types from the MHC homozygous LCLs used in the validation of the MHC array. Reproduced from (Horton, 2008). Sequence length (bp) and number of gaps in each haplotype sequence, together with the HLA gene types obtained by BLAST against the IMGT/HLA database. Dashes indicate the absence of a gene owing to a sequence gap.

Additionally, expression levels of all genes selected were measured in PBMCs from healthy volunteers. This allowed analysis of the genes in non-EBV immortalised cells and for differences in gene expression to be mapped as a quantitative trait by using a cohort of 96 volunteers. This cohort was already established within the Knight lab (Fairfax et al. 2010).

3.3.3 Experimental approach

Primers were designed against exon-spanning SNP-free locations in each gene of interest giving a product of around 100bp in length (for primer sequences see Table A.1 in the Appendix). RT-PCR was used to analyse cDNA, normalised using the housekeeping genes *GAPDH* and *ACTB*. From the cohort of healthy volunteers gene expression analysis was not possible in all individuals for all genes, and in some cases genotyping information was unavailable for particular SNPs. The final number of individuals used in the gene-expression of each gene is indicated in the figure legend. Genome-wide genotyping of the healthy volunteer cohort and homozygous LCLs at 45,237 SNPs was performed using the humanCVD bead array (Keating 2008) which covers around 2000 genes implicated in immune and inflammatory responses at a high density. These genes have specific relevance to inflammatory disease states, vascular pathology and metabolic disorders. Genotyping over the *HLA-C* locus was sparse and Sanger sequencing was performed to genotype a SNP previously associated with *HLA-C* expression, rs9264942 (Thomas 2009). Expression-quantitative-trait mapping was performed by genotype using PLINK (Purcell 2007) with four maximum per-SNP missing genotypes (GENO 0.1) and MAF 0.03. For each SNP, PLINK generates a phenotypic mean expression for the

three genotypic states found in the volunteer cohort and compares these means using the Wald test statistic to generate a P-value. The Wald test does not require that the data fit a normal distribution, as it uses an estimate of the sample to assess the true value of an association. Thus expression values can be tested for association without forming a continuous curve. The Vega Genome Browser (http://vega.sanger.ac.uk/Homo_sapiens/Info/Index) was used to determine the genotype of the LCLs at various SNPs shown to be most associated with expression of particular genes in the volunteers.

3.3.4 *ZFP57*

Expression of *ZFP57* as quantified by RT-PCR was found to occur only in the COX cell line and not in PGF and QBL, consistent with the data from the MHC custom array (Figure 3.3). To investigate this further, expression of *ZFP57* was quantified in PBMCs from 93 healthy volunteers and expression-associated SNPs mapped using 45,237 SNPs. When analysed using PLINK, this showed a highly significant association between expression of *ZFP57* and rs29228 ($p=1.23x10^{-14}$), a SNP located 16.8 kb downstream from *ZFP57* and found close to the gene *MOG* (Figure 3.3). The COX cell line is homozygous for the minor allele at rs29228.

Expression of *ZFP57* was quantified in three further LCLs homozygous for the MHC. This showed that only those homozygous for the minor allele at rs29228 expressed *ZFP57* (Figure 3.3). The SNP rs29228 is also found to be in complete LD with rs3129073, a SNP which was found to be significantly associated with *ZFP57* expression in a separate dataset ($p=5.4x10^{-30}$), comprising expression in LCLs established from a familial asthma cohort (Dixon et al. 2007).

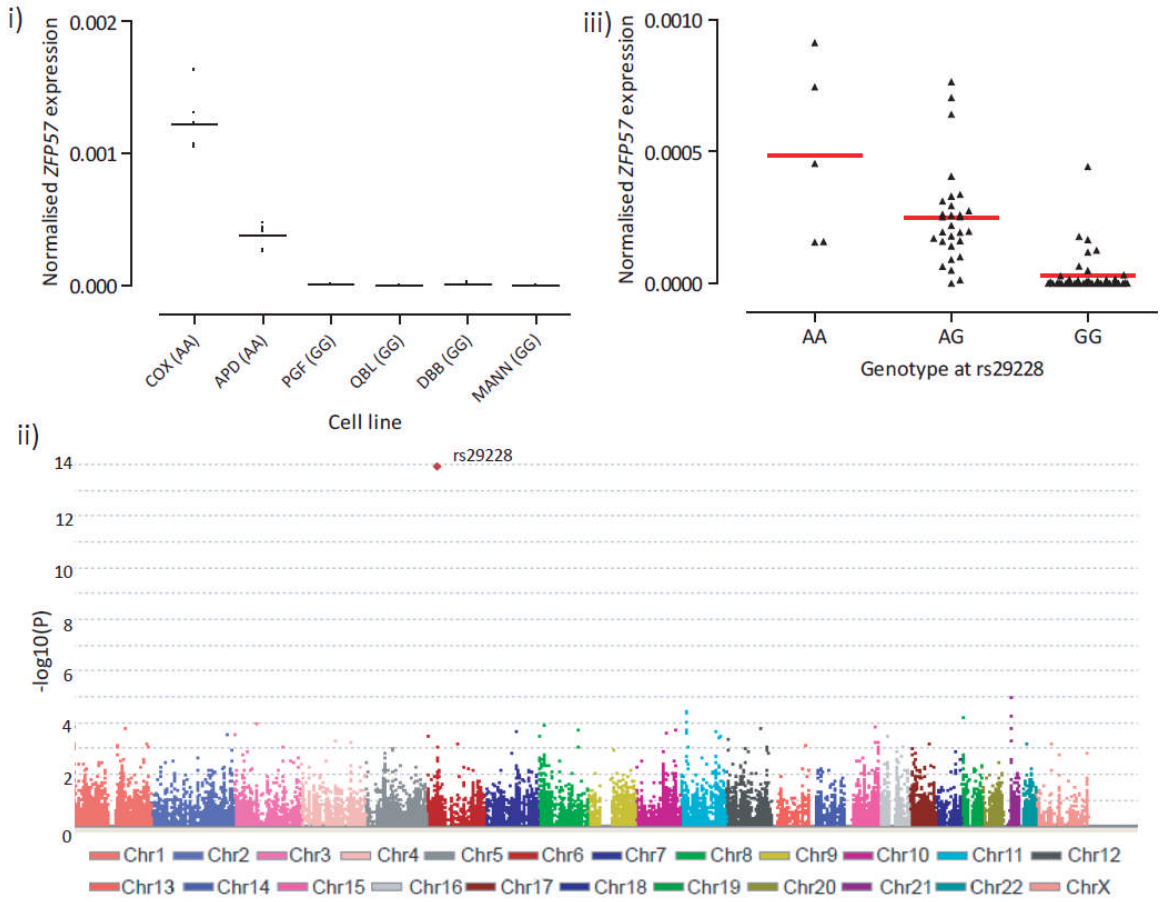


Figure 3.3: Genetic determinants of *ZFP57* expression. (i) Expression of *ZFP57* in six LCLs homozygous for the MHC determined by qPCR and normalised against *GAPDH*. Average expression from six replicates is shown. Genotype for rs29228 is shown in brackets after the name of the cell line (Ancestral allele G); (ii) Genome-wide Manhattan Plot showing the association of SNPs with *ZFP57* expression in 93 healthy volunteers. rs29228 (indicated with a larger diamond), located upstream of neighbouring gene *MOG*, is found to be highly significantly associated with *ZFP57* expression ($p=1.2x10^{-14}$); (iii) *ZFP57* expression by genotype for rs29228 in 93 healthy volunteers shows differential expression according to genotype when analysed using a Kruskal-Wallis test ($p<0.0001$), mean expression for each genotype is indicated by the red bar.

This analysis was repeated using the gene *ACTB* as the house-keeping gene used in normalisation for both the LCLs and the healthy volunteer PBMCs. The same SNP rs29228 was identified as being highly associated with *ZFP57* expression in the volunteer cohort ($p=9.6x10^{-19}$) and expression was also shown to segregate by genotype of this SNP in the LCLs.

3.3.5 *HLA-DQA2*

Expression of *HLA-DQA2* was shown to be significantly higher in the PGF cell line than any of the other LCLs. A peak of association with *HLA-DQA2* expression was found within the MHC locus when association analysis is performed using the volunteer cohort (see Figure 3.4), however these are not at genome-wide significance ($p<5x10^{-7}$). The most significantly associated SNP in the MHC to *HLA-DQA2* expression was rs2269423 and there was a significant difference found between mean expression of *HLA-DQA2* according to genotype at this SNP when analysed with a Kruskal-Wallis test (p=0.0015).

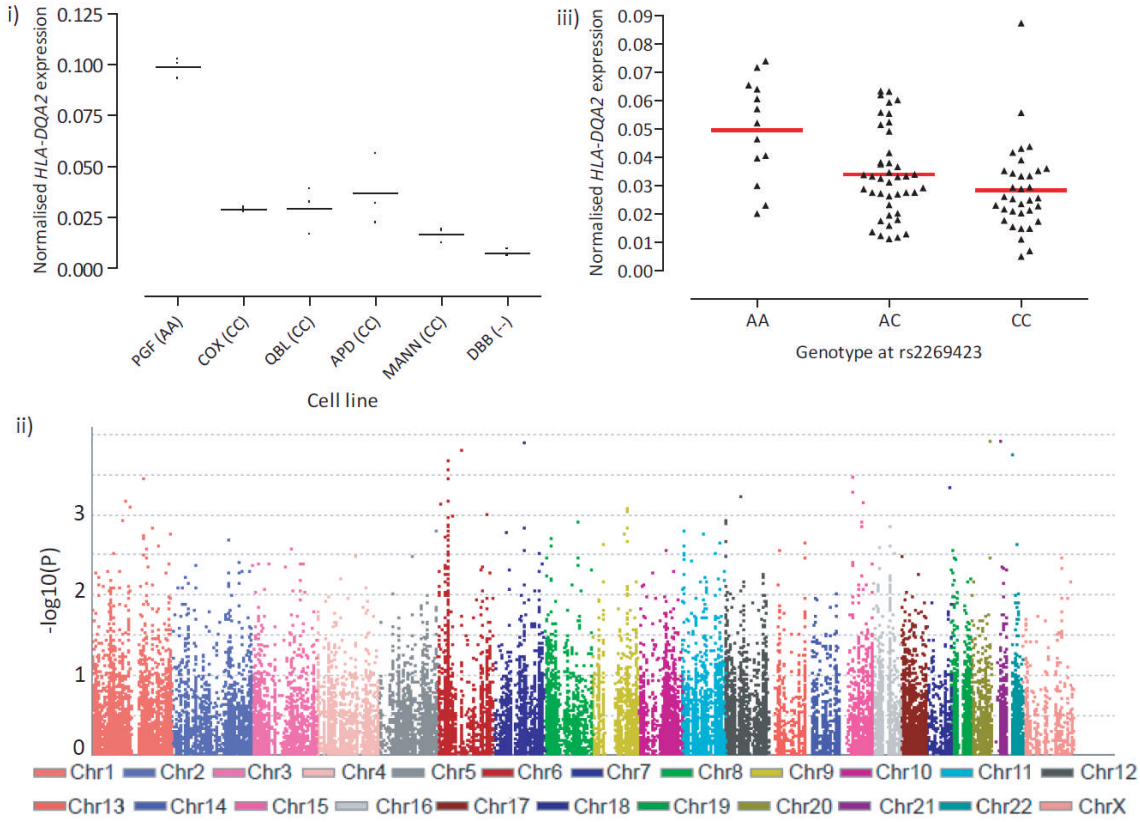


Figure 3.4: Genetic determinants of *HLA-DQA2* expression. (i) Expression of *HLA-DQA2* in six LCLs homozygous for the MHC determined by qPCR and normalised against *ACTB*. Mean expression from six replicates is shown. Genotype for rs2269423 is shown in brackets after the name of the cell line (Ancestral allele C); (ii) Genome-wide Manhattan Plot showing the association of SNPs with *HLA-DQA2* expression in 89 healthy volunteers; (iii) *HLA-DQA2* expression by genotype for rs2269423 in 89 healthy volunteers, mean expression for each genotype is indicated by the red bar.

3.3.6 *HLA-DQB2*

Expression of *HLA-DQB2* was shown to be significantly higher in the PGF cell line than any of the other LCLs as predicted by the MHC array. However, the genotype at rs9469220 does not seem to lead to a difference in expression level in the LCLs. A modest peak of association with *HLA-DQB2* expression was found at the MHC, with the most associated SNP being rs9469220, however these are not at genome-wide significance (Figure 3.5). When expression of *HLA-DQB2* in the volunteer cohort was examined, a significant difference in expression is found according to genotype (p=0.0007 by Kruskal-Wallis test).

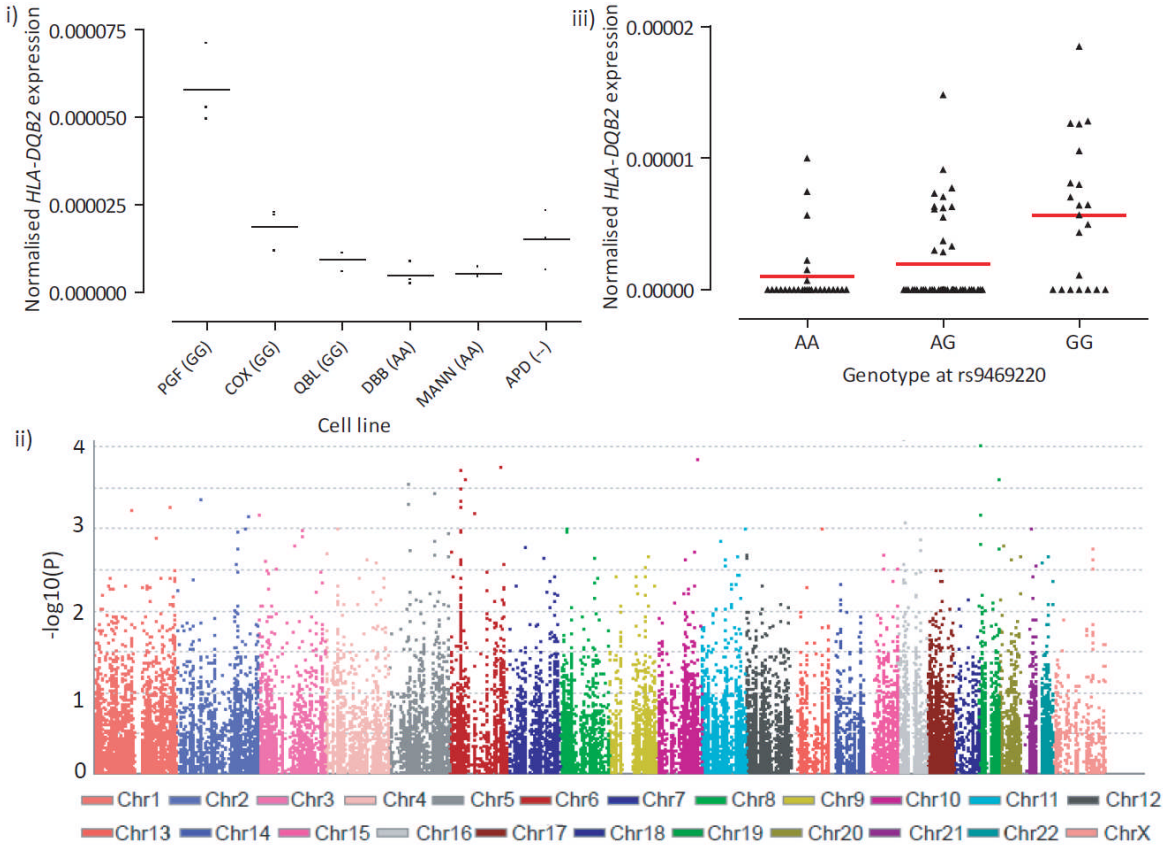


Figure 3.5: Genetic determinants of *HLA-DQB2* expression. (i) Expression of *HLA-DQB2* in six LCLs homozygous for the MHC determined by qPCR and normalised against *ACTB*. Genotype for rs9469220 is shown in brackets after the name of the cell line (Ancestral allele A) and mean expression from six replicates is indicated; (ii) Genome-wide Manhattan Plot showing the association of SNPs with *HLA-DQB2* expression in 94 healthy volunteers; (iii) *HLA-DQB2* expression by genotype for rs9469220 in 94 healthy volunteers, mean expression for each genotype is indicated by the red bar.

3.3.7 *TNF*

Although differential expression of *TNF* was seen in the healthy volunteer cohort, SNPs around the MHC region seem not to be any more associated with *TNF* expression than SNPs over the rest of the genome (Figure 3.6). SNP rs1150476 found on chromosome 5 was seen to be associated with *TNF* expression at genome-wide significance ($p=1.38 \times 10^{-8}$) and may be indicating an association in *trans*. When expression by genotype at rs1150476 was investigated, a significant difference is found ($p=0.0158$ by Kruskal-Wallis test), however due to the relatively small size of the healthy volunteer cohort and the low incidence of the SNP it may be a false positive result (Figure 3.6).

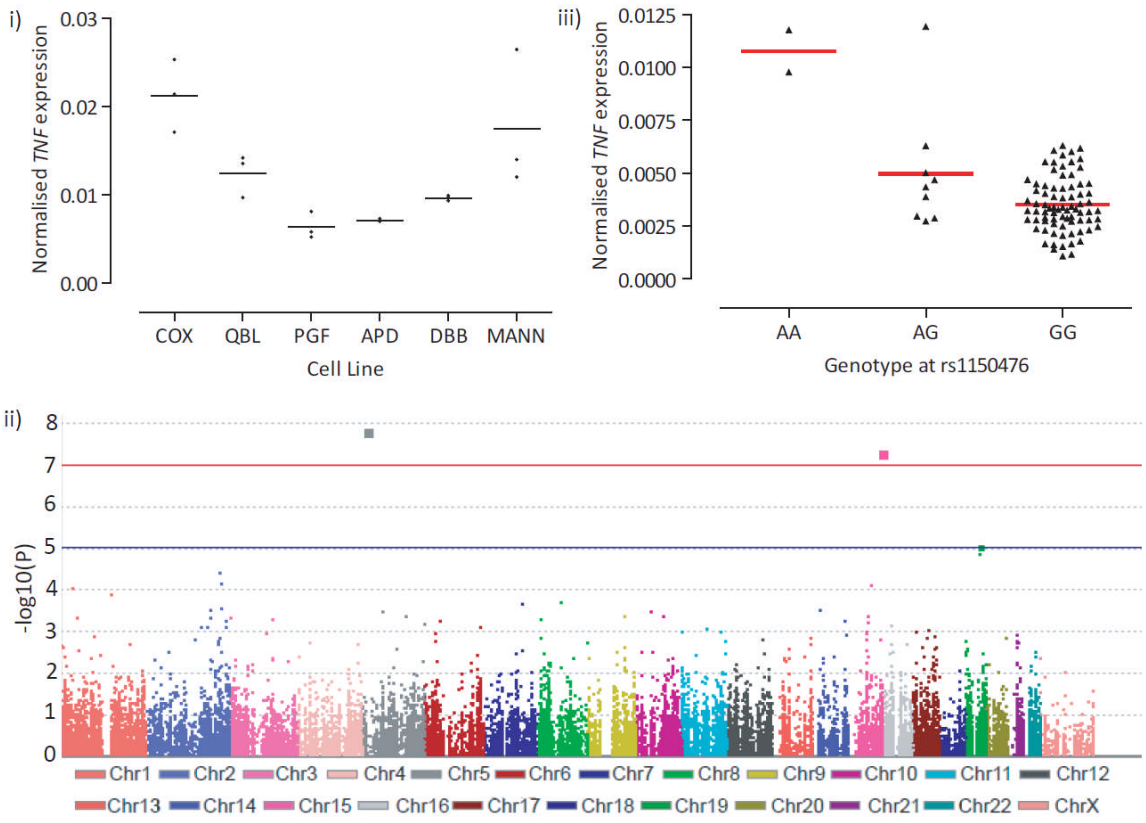


Figure 3.6: Genetic determinants of *TNF* expression. (i) Expression of *TNF* in LCLs homozygous for the MHC determined by qPCR and normalised against *ACTB*, mean expression from six replicates is indicated; (ii) Genome-wide Manhattan Plot showing the association of SNPs with *TNF* expression in 94 healthy volunteers. Suggestive association at genome-wide significance is indicated with a blue line ($p=1x10^{-5}$), while genome-wide significance ($p=5x10^{-7}$) is indicated with a red line. (iii) *TNF* expression by genotype for rs1150476 (Ancestral allele A), mean expression for each genotype is indicated by the red bar.

3.3.8 *HLA-DPB1*

Gene expression of *HLA-DPB1* was found to be highest in the PGF cell line as predicted by the MHC array results (Figure 3.7). There was no clear association found with expression of *HLA-DPB1* and genetic variation in the MHC. The SNP genotype most associated with differential expression in this region in the healthy volunteer cohort (rs77465553) did not associate with expression differences in the LCLs. However, there was a significant difference between gene expression of *HLA-DPB1* by genotype at rs77465553 in the healthy volunteer cohort (see Figure 3.7).

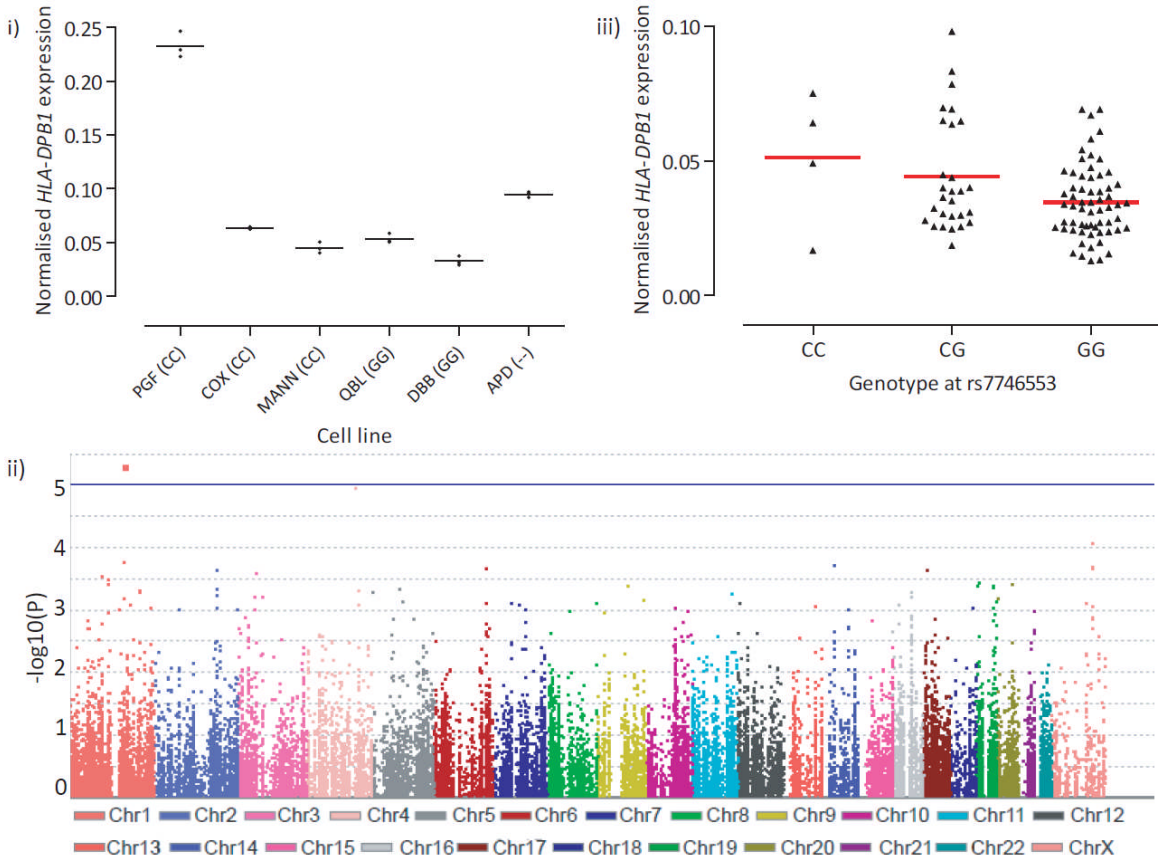


Figure 3.7: Genetic determinants of *HLA-DPB1* expression. (i) Expression of *HLA-DPB1* in six LCLs homozygous for the MHC determined by qPCR and normalised against *ACTB*, mean expression from six replicates is indicated. Genotype for rs7746553 is shown in brackets after the name of the cell line (Ancestral allele C); (ii) Genome-wide Manhattan Plot showing the association of SNPs with *HLA-DPB1* expression in 94 healthy volunteers. Suggestive association at genome-wide significance is indicated with a blue line ($p=1x10^{-5}$). (iii) *HLA-DPB1* expression by genotype for rs7746553, mean expression for each genotype is indicated by the red bar.

3.3.9 *HLA-DQA1*

SNPs associated with *HLA-DQA1* expression from the volunteer cohort peaked around the MHC, specifically the *HLA-DQA1* locus, although the p values for these associated SNPs were not at genome-wide significance (Figure 3.8). The most strongly associated SNP to *HLA-DQA1* expression in the volunteers, rs9272346, was found to be associated with a significant change in expression in the LCLs ($p=0.0128$ using a Mann-Whitney test that does not assume normal distribution of expression). The PGF cell line had the highest level of *HLA-DQA1* expression, and possesses the rs9272346-GG genotype that was shown to be associated with higher *HLA-DQA1* expression in the volunteers ($p<0.0001$ determined by Kruskal-Wallis test) (shown in Figure 3.8).

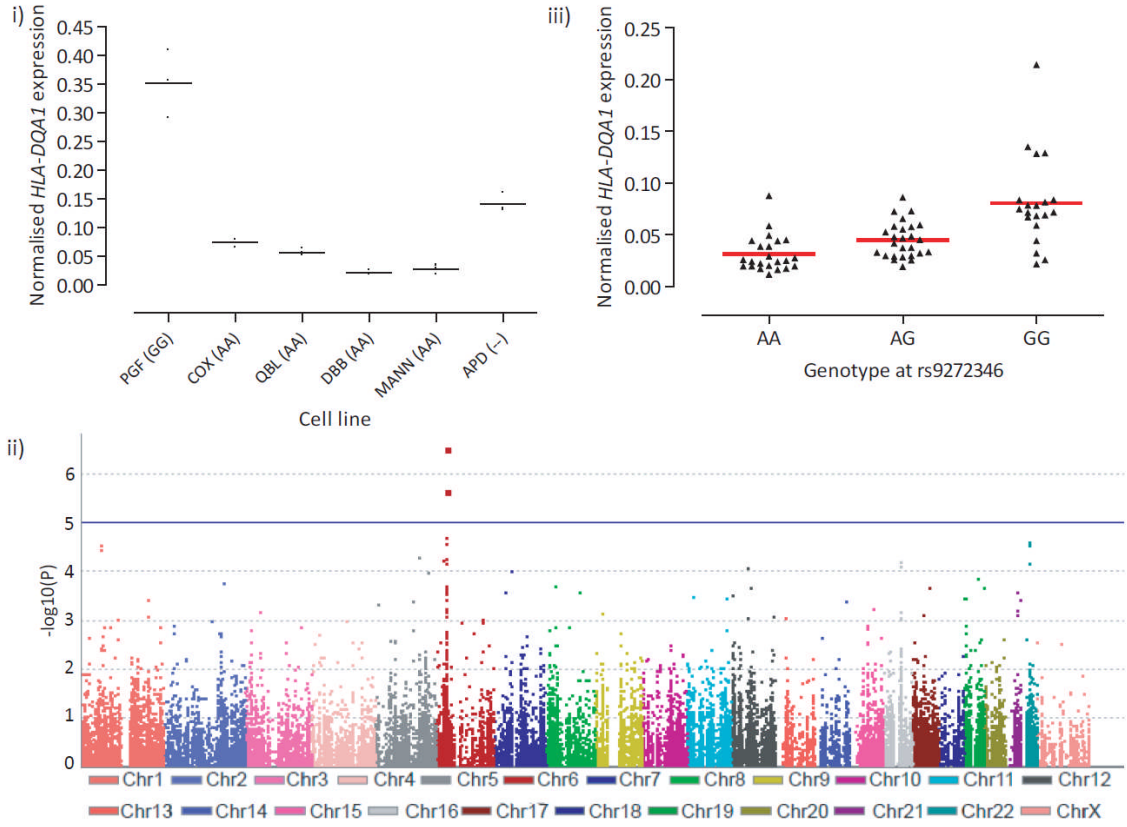


Figure 3.8: Genetic determinants of *HLA-DQA1* expression. (i) Expression of *HLA-DQA1* in six LCLs homozygous for the MHC determined by qPCR and normalised against *ACTB*, mean expression from six replicates is indicated (Ancestral allele A). Genotype for rs9272346 is shown in brackets after the name of the cell line; (ii) Genome-wide Manhattan Plot showing the association of SNPs with *HLA-DQA1* expression in 89 healthy volunteers. Suggestive association at genome-wide significance is indicated with a blue line ($p=1x10^{-5}$). (iii) *HLA-DQA1* expression by genotype for rs9272346, mean expression for each genotype is indicated by the red bar.

3.3.10 *HLA-DQB1*

The PGF cell line was shown to have much higher *HLA-DQB1* expression than either COX or QBL, and this was also seen in the results from the MHC array (Figure 3.9). The expression of *HLA-DQB1* was convincingly associated with genetic variation; several SNPs within the MHC were significantly associated with *HLA-DQB1* expression at a genome-wide level when analysed in the healthy volunteer cohort. The SNP rs1071630 predicts the level of *HLA-DQB1* expression with little or no expression being found in volunteers or LCLs that have the genotype rs1071630-GG. There was a significant difference in mean expression of *HLA-DQB1* between the three different genotypes in the volunteer cohort when analysed by Kruskal-Wallis test ($p<0.0001$) (Figure 3.9).

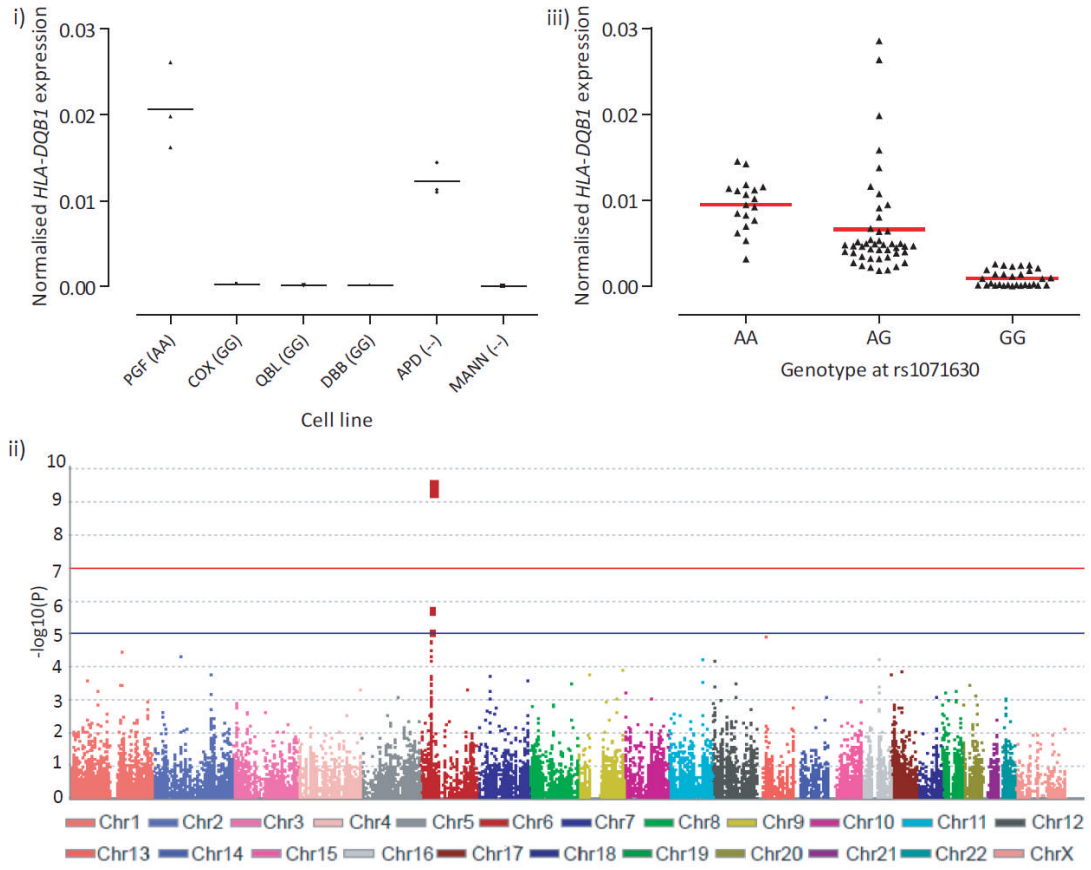


Figure 3.9: Genetic determinants of *HLA-DQB1* expression. (i) Expression of *HLA-DQB1* in six LCLs homozygous for the MHC determined by qPCR and normalised against *ACTB*. Genotype for rs1071630 is shown in brackets after the name of the cell line. Mean expression from the six replicates is indicated; (ii) Genome-wide Manhattan Plot showing the association of SNPs with *HLA-DQB1* expression in 94 healthy volunteers. Suggestive association at genome-wide significance is indicated with a blue line ($p=1x10^{-5}$), while genome-wide significance ($p=5x10^{-7}$) is indicated with a red line. (iii) *HLA-DQB1* expression by genotype for rs1071630, mean expression for each genotype is indicated by the red bar.

3.3.11 *HLA-DPA1*

As predicted by the results of the MHC array, the PGF cell line was shown to have higher *HLA-DPA1* expression than either COX or QBL (Figure 3.10). Study of the volunteer cohort shows SNPs most associated with expression of *HLA-DPA1* were found in *trans*, however there was some evidence of weak association of some SNPs on chromosome 6 with *HLA-DPA1* expression (shown in Figure 3.10). The most associated SNP to *HLA-DPA1* expression found within the MHC has no samples of the rarer allele (rs3130342-T) in a homozygous state in the volunteer cohort however this was expected due to the frequency of the other genotypes. Genotyping information for this SNP was not available for several of the LCLs. Despite this, there was a significant difference found between the mean expression of *HLA-DPA1* in the volunteers when analysed by Mann-Whitney test (p=0.0033).

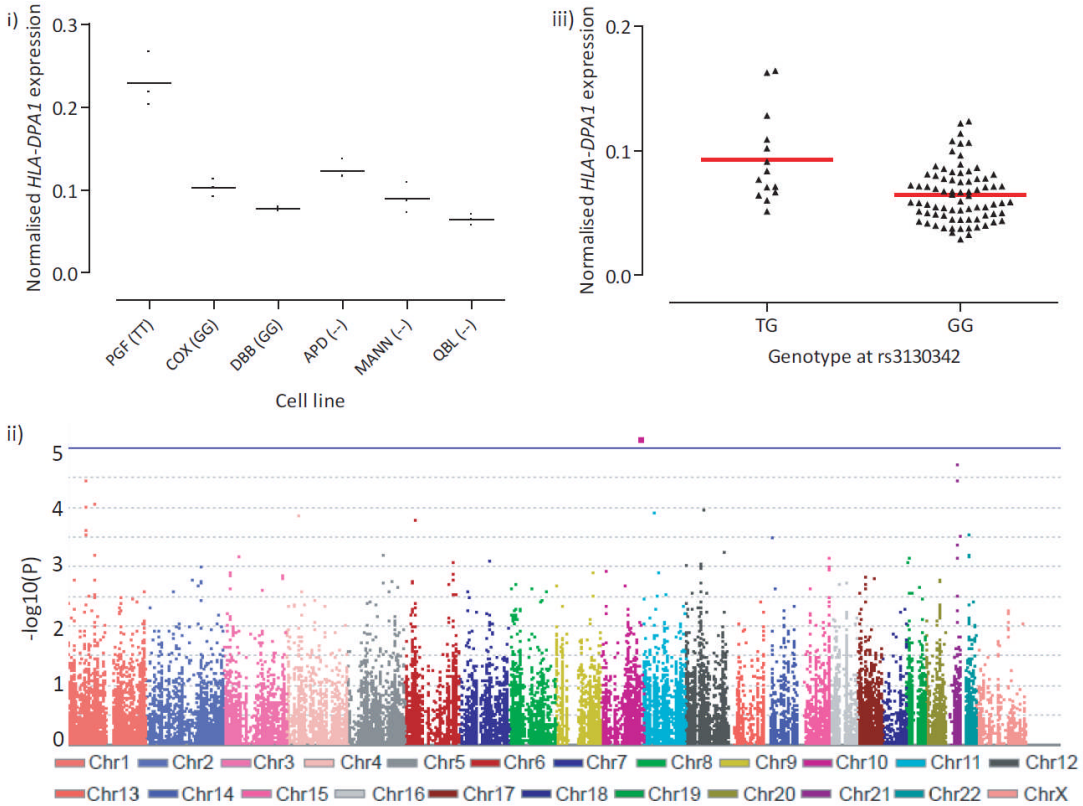


Figure 3.10: Genetic determinants of *HLA-DPA1* expression. (i) Expression of *HLA-DPA1* in six LCLs homozygous for the MHC determined by qPCR and normalised against *ACTB*. Mean expression from the six replicates is indicated. Genotype for rs9272346 is shown in brackets after the name of the cell line; (ii) Genome-wide Manhattan Plot showing the association of SNPs with *HLA-DPA1* expression in 89 healthy volunteers. Suggestive association at genome-wide significance is indicated with a blue line ($p=1x10^{-5}$). (iii) *HLA-DPA1* expression by genotype for rs3130342, mean expression for each genotype is indicated by the red bar.

3.3.12 *HLA-C*

As predicted by the MHC custom array, QBL had higher *HLA-C* expression than either COX or PGF (Figure 3.11). Analysis of the healthy volunteer cohort showed that several SNPs within the MHC were weakly associated with *HLA-C* expression although there were relatively few SNPs located close to *HLA-C* on the HumanCVD array. In view of this, the SNP rs9264942 was genotyped directly by Sanger sequencing in the LCLs and the volunteer cohort as it had been previously associated with HLA-C cell surface expression (Thomas 2009). Of the additional LCLs analysed for *HLA-C* expression, higher expression was seen in LCLs with the rs9264942-CC genotype, with a significant difference in mean expression seen between the two groups ($p<0.0001$ when determined with a T-test using Welch's correction). Differential *HLA-C* expression, summarised in Figure 3.11, was also seen according to genotype in the volunteer cohort when analysed using the Kruskal-Wallis test ($p=0.0481$).

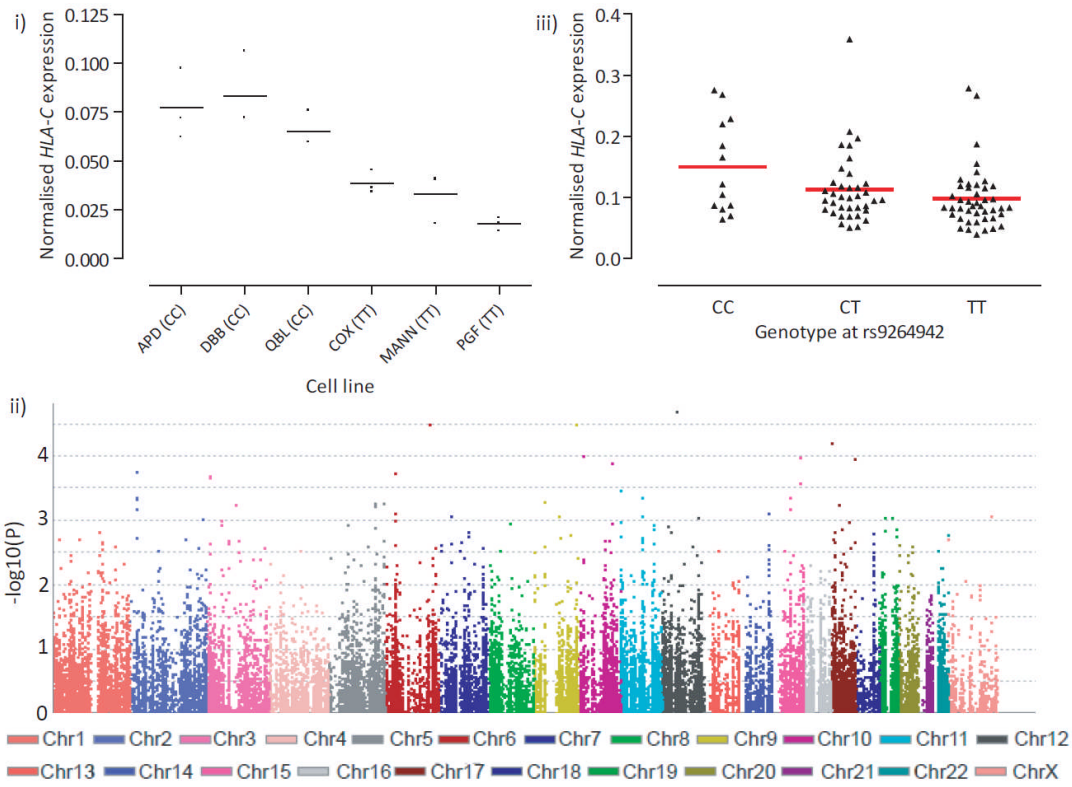


Figure 3.11: Genetic determinants of *HLA-C* expression. (i) Expression of *HLA-C* in six LCLs homozygous for the MHC determined by qPCR and normalised against *ACTB,* mean expression from the six replicates is indicated. Genotype for rs9272346 is shown in brackets after the name of the cell line (Ancestral allele C); (ii) Genome-wide Manhattan Plot showing the association of SNPs with *HLA-C* expression in 94 healthy volunteers. (iii) *HLA-C* expression by genotype for rs9264942, a SNP previously associated with HLA-C cell surface expression. Mean expression for each genotype is indicated by the red bar.

3.4 Discussion

3.4.1 Selection of a housekeeping gene

Differential expression was found in all genes investigated with qPCR. There was no discernable difference in the results of investigation of *ZFP57* expression when a different housekeeping gene was used for normalisation of expression. Cell line expression of *ZFP57* normalised to both *GAPDH* and *ACTB* showed the same pattern of expression, segregating by genotype at rs29228. This SNP was found to be the most associated SNP with *ZFP57* expression in the cohort of healthy volunteers regardless of the gene used in normalisation (with *GAPDH* $p=1.2x10^{-14}$, with *ACTB* $p=9.6x10^{-19}$). All other SNPs showing signs of association with *ZFP57* expression failed to reach significance at a genome-wide level when either housekeeping gene was used to normalise gene expression. Therefore one gene, *ACTB*, was selected for the purpose of normalising qPCR gene expression data. While differences in expression of both housekeeping genes have been described previously, these have generally been associated with alteration of the cell in some way, such as chemical stimulation (Chang 1998, Thellin 1999). In this case, as unstimulated cell lines and primary cells were analysed, it was appropriate to use the *ACTB* housekeeping gene.

3.4.2 Validating differential expression seen in nine MHC genes

The validation of differential expression using qPCR was successful for all five genes investigated from the custom array. All genes showed evidence of differential expression between LCLs and this was further investigated by eQTL analysis in primary PBMCs using the healthy volunteer cohort. Moreover, the expression-associated SNPs for these genes when analysed in the volunteer cohort were consistent with the observed expression differences in the three additional MHC homozygous LCLs APD, DBB and MANN studied, where genotyping information was available. The SNPs showing association with expression were generally found to be local, likely *cis-acting* variants, however this is not surprising as a majority of regulatory variants discovered are found to be in *cis* (Ge 2009). Although a *trans* association was seen in the analysis of expression of *TNF*, the genotype associated with increased expression had only two samples homozygous for the alternate allele (rs1150474-A).

The heterozygote samples did not differ significantly from the group homozygous for the reference (rs1150474-G) with a p value of 0.16 when analysed by t-test using Welch's Correction to correct for unequal variance. This may indicate that the association seen here is the result of three outliers showing higher expression of *TNF* rather than a genuine *trans* association. It is likely that in order to determine *trans* associations with expression a much larger cohort of samples would be needed (see Figure A.2) even though they are thought to be as numerous as *cis* associations (Cheung 2010). Further validation of the observed cis-associations in a larger independent cohort and fine mapping using a denser SNP set is also required to further interpret these results and this is explored further in Chapters 4 and 5.

There was good concordance between expression patterns seen in the MHC homozygous LCLs and the PBMC samples. All genes selected from the array results for validation of ASE were found to have variable gene expression levels in the PBMC samples when normalised using *ACTB* expression levels. The SNP most associated with these differences in gene expression was also indicative of relative gene levels in the MHC homozygous LCLs in many cases where a difference in genotype was found between any of the LCLs. Expression of *HLA-DQB2* could not be predicted by genotype in the MHC homozygous LCLs, however expression of *HLA-DQB2* was extremely low and so differences in expression may not have been as clearly determined as with other genes examined. Genotype information at rs9469220 was also not available for the APD cell line. The -35 SNP (rs9264942) previously associated with *HLA-C* expression (Thomas 2009) was found to be significantly associated with gene expression in the cohort of healthy volunteers (p=0.01455 at a genome-wide level, determined by PLINK analysis, p=0.034 determined by Kruskal-Wallis test on genotypes) and was indicative of a slight increase or decrease in gene expression according to genotype in the MHC homozygous LCLs.

3.4.3 Advantages of the MHC array

The MHC array avoids several confounding factors traditionally associated with commercial arrays by taking into account underlying sequence polymorphism. As the array includes a tiling probe set

across the whole MHC it allows for the discovery of new transcripts as well as splicing variants and changes in gene expression. The haplo-spliceo-transcriptome is a tantalising area for further research from both the perspective of ASE and alternative splicing (Graveley 2008) and the MHC custom array is a significantly improved resource with which to study it in the context of the MHC. Although RNA-seq would have higher resolution of all allelic differences it remains to be seen how successful mapping of unknown and highly polymorphic sequences in the MHC can be with RNA-seq (see Chapter 6) and it is still an extremely costly methodology. The MHC array gives a more cost-effective means for analysing gene expression and can be used to also study DNA methylation and chromatin accessibility based on DHSs while still taking into account the haplotypic variation (Vandiedonck 2011). The alignment of the transcription tracks and the DHS tracks over *ZFP57* show that there are clear differences in chromatin accessibility between COX, PGF and QBL DNA at the 5' end of *ZFP57* (Figure 3.2). This is likely to be due to more open chromatin conformation in the COX cell line allowing transcription of *ZFP57* to occur. The ability to assess differences in transcription with the difference in DHSs particularly allows new transcripts to be scrutinised for additional information that supports whether they are real results or artefacts. For example, high levels of intron bleeding seen across intron 1 of *ZFP57* is seen in all three LCLs analysed with the array despite very different expression levels of *ZFP57* between COX and the other two LCLs. When studied in conjunction with the DHS data, the difference seen in the COX line agrees with the up-regulation of *ZFP57* expression. As no change in found in the QBL and PGF lines the intron bleeding is more likely to be an artefact of the array. This problem may be resolved as more total RNA-sequencing is undertaken and the results of the two techniques are compared.

3.4.4 Selection of candidate genes for further analysis

In accordance with the results of the MHC array, the gene that showed most evidence of variable expression was Z*FP57*. The association of rs29228 to *ZFP57* expression was extremely strong for a relatively small sample size of 93 volunteers genotyped across only 45,237 SNPs. Its location just downstream of *ZFP57* also made it plausible as a marker of cis regulation.

ZFP57, *HLA-DQA1*, *HLA-DQA2*, *HLA-DQB1*, *HLA-DQB2* and *HLA-C* were all selected for further analysis after showing varying degrees of genetic association with gene expression within the MHC region. Although *HLA-C* did not have particularly significant p values for association of MHC SNPs with expression the data does replicate a previously associated SNP rs9264942 (Thomas 2009). Previously, individuals homozygous at rs9264942, showed significantly different cell surface expression of HLA-C detected by FACS (p=0.0003). While we found the difference between the homozygous samples in our study was less significant (p=0.0281, determined by Mann-Witney two-tailed test) this was based on the RNA level rather than cell surface protein; so additional levels of regulation could account for the relative difference in significance observed. Additionally, the density of the genotyping over the *HLA-C* region was extremely low on the genotyping array used (humanCVD bead array), thus further analysis of the expression of *HLA-C* when denser genotyping information is available could be informative. Apart from *ZFP57,* all these genes encode proteins that are part of the antigen presentation pathway and all have been implicated in human disease. The combination of this previous knowledge and the associations seen with differential gene expression make them the most suitable choice for further study. SNPs shown to be associated with the expression of these genes may be causative, for example by modulating the binding of polymerase or an enhancer element, but are more likely to be in LD with a causative SNP. By further investigating gene expression in a larger more densely genotyped cohort, there is a higher chance that functional variants can be identified. However, functional assays will still be necessary to determine true causative SNPs.

3.4.5 Conclusion

Variable gene expression was detected in LCLs and primary tissue for all genes investigated. Strong association of genetic variation with expression was seen for the gene *ZFP57*, and for several other MHC genes that were investigated either because of their suspected haplotype specific differential expression identified by the MHC array, or due to heritable SNP associations with expression of these immune related genes. This proved the value of the MHC array and showed that consistent

results could be demonstrated that were reproducible with qPCR. Promising candidate genes were selected to further define the *cis* associations seen in the primary cells and to try to determine any *trans* associations in a larger sample.

Chapter 4 - Functional genomics of ZFP57, a variably expressed transcription factor encoded in the MHC class I region

4.1 Introduction

4.1.1 ZFP57 in health and disease: potential roles in transient neonatal diabetes and imprinting

ZFP57 is a zinc finger protein comprising of seven zinc finger domains and one Kruppel-associated box (KRAB) domain that has properties of a transcriptional regulator (Okazaki 1994, Alonso 2004). ZFP57 has been shown to be important in developmental settings for maintenance of maternal and paternal imprinting (Li 2008, Mackay 2008) as well as playing a role in transient neonatal diabetes (TND) (Mackay 2008) when a coding mutation is present. It has been identified as a gene that is preferentially expressed in early development, particularly in undifferentiated embryonic stem cells (Li and Leder 2007) and as a downstream molecule in STAT3 signalling (Akagi 2005). Although the establishment and maintenance of imprinting has been associated with epigenetic modifications (Looman 2002, Lewis and Reik 2006), many aspects of it remain mechanistically unclear. It has been noted that many TND patients have aberrant methylation patterns at the genomic loci 6q24 (Temple and Shield 2002) and often have hypomethylation across other areas of the genome (Mackay 2006). Mutations in *ZFP57* were found in TND patients with this hypomethylated mosaic pattern but not in TND patients with only aberrant methylation around 6q24, leading to the prediction that ZFP57 has a genome-wide maintenance role in imprinting. Its absence was shown to cause a lethal phenotype before the age of weaning in mice when ZFP57 was absent from both oocyte as either mRNA or protein product and from the embryo (Li 2008). This has led to the hypothesis that aberrant embryonic maintenance of methylation mediated by *ZFP57* may be a possible mechanism for manifestation of TND (Hirasawa and Feil 2008).

Comparatively little work has been carried out on *ZFP57* in humans, beyond its putative role in development. Further studies attempting to characterise *ZFP57* mutations in patients, or mothers to investigate a maternal effect, with hypomethylation at imprinted loci in disorders such as Beckwith-

Wiedemann syndrome (OMIM 130650) have proved unsuccessful so a generalised function for ZFP57 has not yet been conclusively proved (Boonen 2011).

4.1.2 Differential expression of *ZFP57*

Analysis of haplotype-specific gene expression in three LCLs possessing HLA homozygous haplotypes, associated with commonly occurring autoimmune diseases, using the MHC array demonstrated evidence that there was differential expression of *ZFP57* in COX cell line carrying HLA-A1-B8-Cw7-DR3 (Chapter 3 and Figure 4.1).

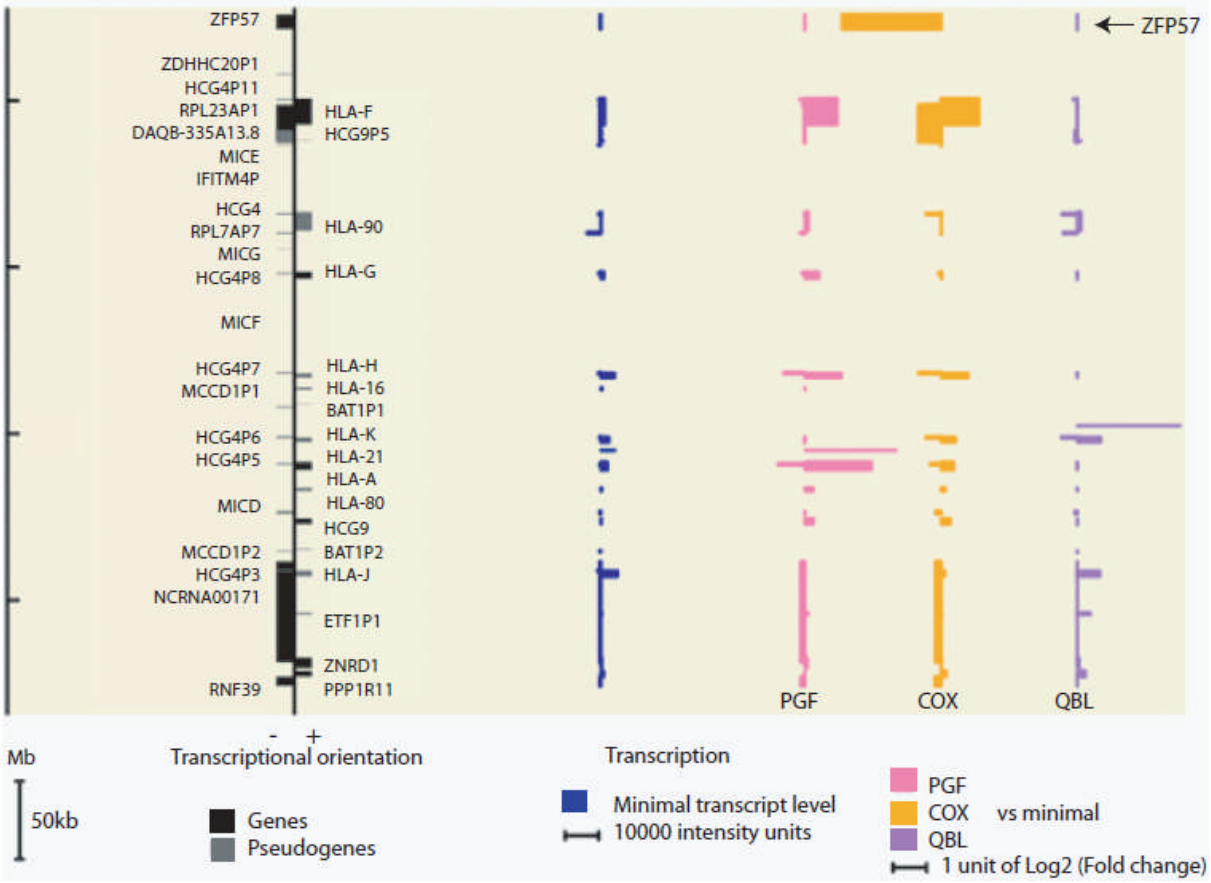


Figure 4.1: *ZFP57* expression in three LCLs determined by custom MHC array. Reproduced from (Vandiedonck 2011). Differential expression of *ZFP57* showing much higher expression in the COX cell line compared with the PGF and QBL cell lines.

4.2 Aims

This chapter seeks to characterise *ZFP57* expression and validate the previous result showing differential expression, both in LCLs and PBMCs. Specific aims of the chapter are:

1. To validate the observed association of *ZFP57* expression assayed by microarray in HLA-homozygous LCLs.

2. To investigate the relationship between *ZFP57* expression and genetic variation by expression quantitative trait mapping.

3. To characterise *ZFP57* at the protein level, expression in diverse biological cell types and tissues and co-expressed genes in order to better understand the functional significance of *ZFP57.*

4. To investigate the relationship between *ZFP57*, genetic variation and disease.

4.3 Results

4.3.1 Validation of differential gene expression observed by microarray

ZFP57 is differentially expressed in the COX cell line

Differential *ZFP57* expression was previously shown in Figure 3.3; however this looked only at unstimulated cells. Expression of *ZFP57* was analysed to determine the effect of PMA/ionomycin stimulation on expression of the gene, to see if this would cause up-regulation of expression. Cells were stimulated for six hours and were compared with gene expression in unstimulated cells harvested at the time of stimulation (Figure 4.2).

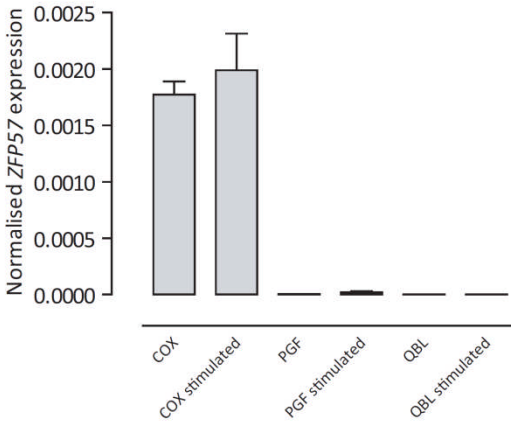


Figure 4.2: Differential expression of *ZFP57* in three LCLs homozygous over the MHC. Mean gene expression in unstimulated and stimulated, 200nM PMA and 125nM Ionomycin, is shown. Expression was determined using qPCR relative to *ACTB*, and standard deviation between the three replicates is shown. Stimulation of the LCLs has no effect on expression of *ZFP57*, and expression is virtually undetectable in both the PGF and QBL lines.

Potential confounding effect of *MOG*

ZFP57 is located at the boundary of the classical MHC and was the final gene included in the tiling probe design such that no information was available for expression of the immediately telomeric region. This contained another gene *MOG* (encoding myelin oligodendrocyte glycoprotein) which may be confounding the observed expression differences. Genetic variation in *MOG* has been associated with autoimmune disease susceptibility, for example in MS (D'Alfonso 2008).

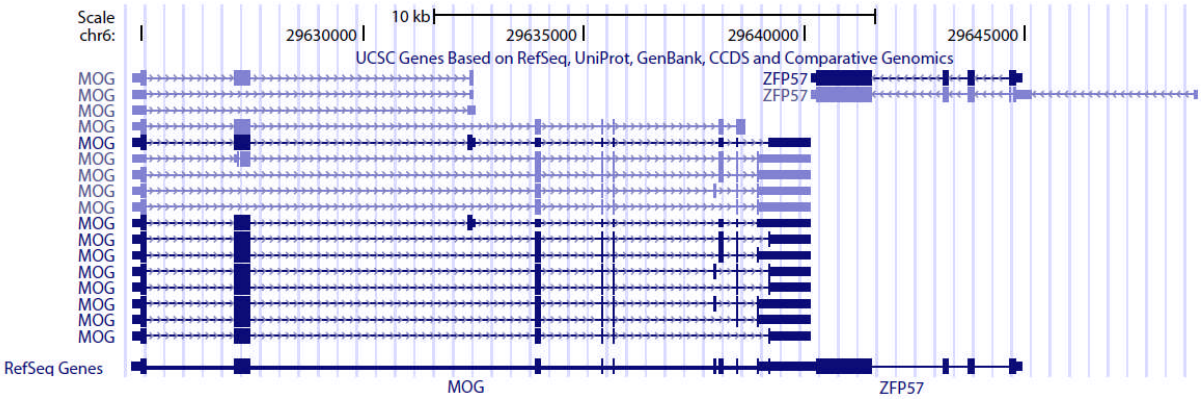


Figure 4.3: Genomic location of *ZFP57* and *MOG* on chromosome 6. A plot from the UCSC browser (Kent 2002) shows the genomic loci of the 2 genes. Multiple isoforms are shown for *MOG* while *ZFP57* is annotated as having only 2 isoforms. RefSeq Gene information at the bottom of the figure shows the extremely close proximity of the 2 genes to one another.

Despite the location of *MOG* on the opposite strand of DNA to *ZFP57* it was interrogated in the homozygous LCLs using qPCR against cDNA to ensure the differential expression seen in *ZFP57* was

not as a result of *MOG* differential expression. As can be seen in Figure 4.3, there is a documented possibility of run-on between the two genes shown in the RefSeq gene track. Transcription of *MOG* occurs in the opposite direction to *ZFP57* and it is possible that run on of Pol II from *MOG* expression could lead to confounding expression of *ZFP57* particularly at the 3' end of the gene. As shown in Figure 3.2, the array showed strongest expression of *ZFP57* towards the 3' end of the gene, thus investigation of *MOG* expression was carried out using SYBRgreen real-time qPCR with primer sets designed at two exon junctions of the gene due to the documented different isoforms (Figure 4.4). RNA derived from brain tissue was used as a positive control due to the previously noted high expression of *MOG* in the brain (Allamargot and Gardinier 2007).

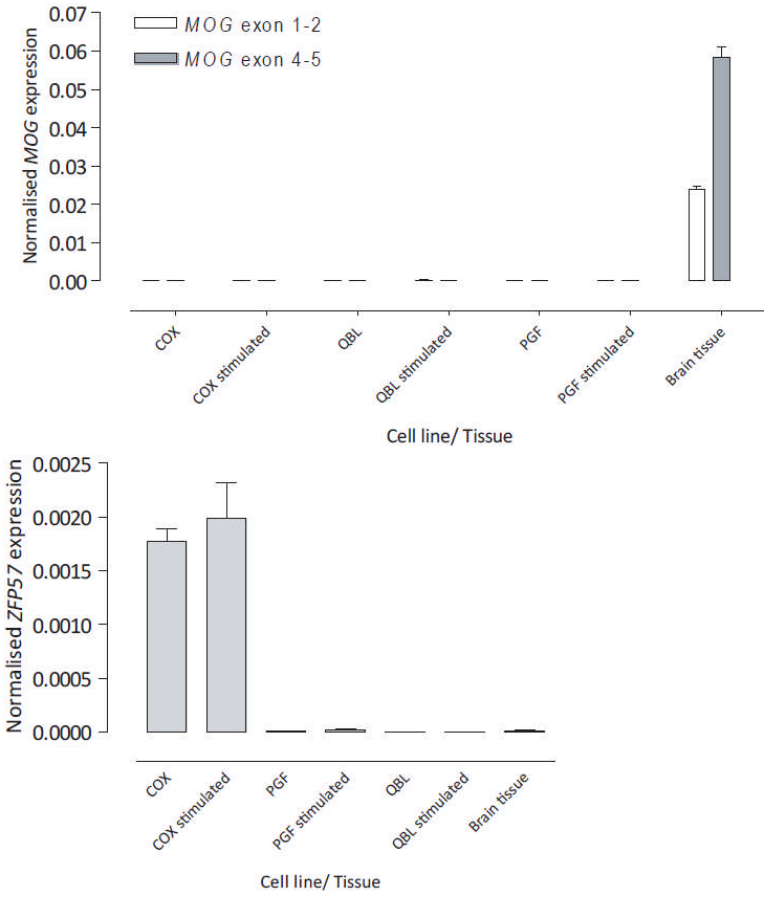


Figure 4.4: Expression of i) *MOG* and *ii) ZFP57* in LCLs and Brain tissue. *MOG* expression was analysed over two separate exon junctions (i), and *ZFP57* expression was analysed over exon junction 5-6 (ii) to determine if the same expression patterns of *ZFP57* seen in the three homozygous cell lines PGF, QBL and COX (both unstimulated and stimulated samples) were also seen for *MOG*. To ensure expression was determinable by qPCR for *MOG*, a positive control of brain tissue cDNA was also analysed. All qPCR reactions were carried out in triplicate and the mean and standard deviation is shown. All qPCR analysis was carried out relative to *ACTB*.

4.3.2 Assessment of *ZFP57* isoforms

Further expression analysis of *ZFP57* was carried out using exon spanning primers (see Figure 4.5) for all exons to determine if there was evidence of additional isoforms beyond the two reported by UCSC and VEGA genes in the COX cell line (as shown in Figure 4.3). The COX cell line was chosen as it had repeatedly shown the highest expression of *ZFP57* at the mRNA level, and showed increased expression of *ZFP57* towards the 3' end of the gene in the MHC array analysis (Figure 3.2).

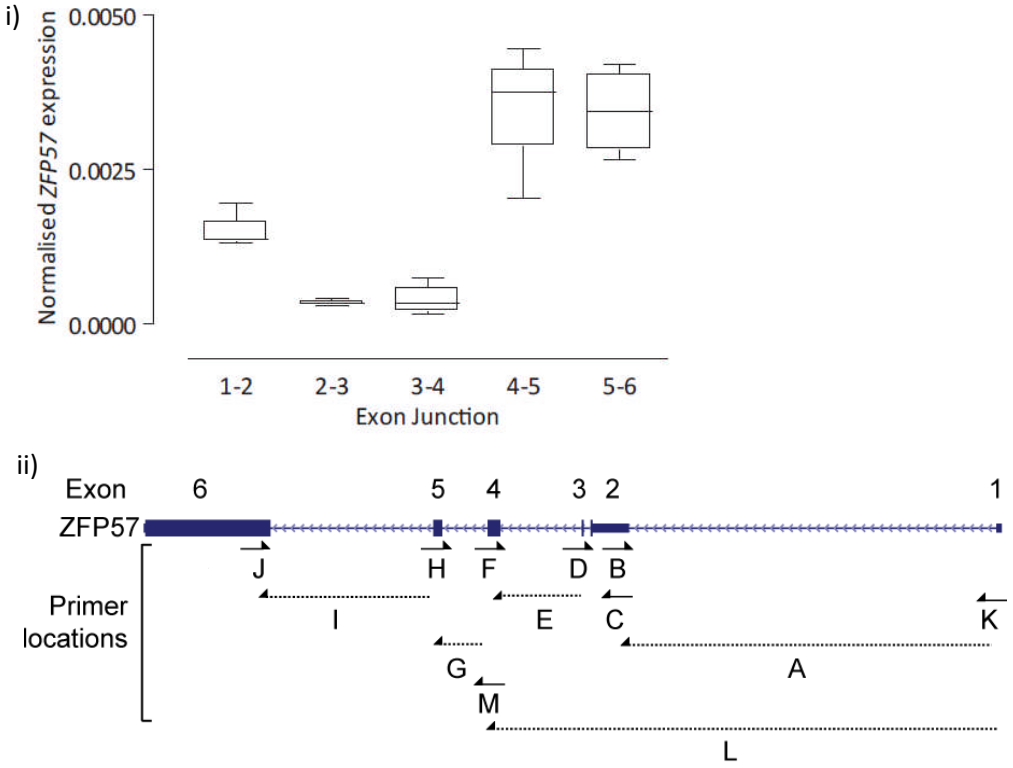


Figure 4.5: Expression of *ZFP57* across all exon junctions. i) *ZFP57* expression was analysed using four replicates of qPCR in the COX cell line with exon junction specific primers (see Table A.1 for sequences). Expression values were normalised against *ACTB*, mean expression and standard deviation are shown. ii) Key to the primer locations in *ZFP57.*

Expression was found to be significantly lower in the 2nd and 3rd exons of *ZFP57*, with highest expression seen towards the 3' end of the gene. Following this discovery, Rapid Amplification of cDNA ends (RACE) PCR was carried out to characterise the isoforms of *ZFP57* present in the COX cell line. Again, COX was used as it showed reproducible *ZFP57* expression and had shown evidence of variation in expression of different *ZFP57* exons. RACE-PCR was performed in both directions using

both 3' and 5' adapted cDNA for RACE. *ZFP57* specific sequences found by RACE-PCR are shown in Figure 4.6 and showed that in addition to the expected isoform (detected with 3' RACE-PCR) two additional isoforms could be seen:

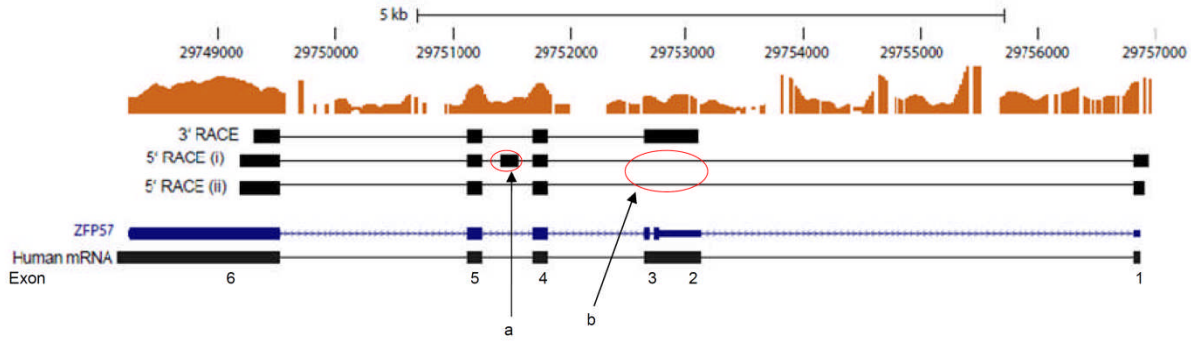


Figure 4.6: RACE derived isoforms of *ZFP57* aligned to the transcript profile from the MHC array for the COX LCL. The predicted human mRNA for *ZFP57* is shown at the bottom of the schematic, with the RefSeq gene for COX above this. RACE-PCR result sequences are shown above this, with MHC custom array expression of *ZFP57* in COX at the top. RACE-PCR sequences were confirmed with Sanger sequencing (two replicates) following gel extraction of the RACE-PCR products. An additional exon seen with RACE-PCR

Consistent with the expression array data, exons 4, 5 and 6 are seen in all three sequences determined by RACE-PCR, shown in Figure 4.6. There is also expression seen on the array between exons 4 and 5 that could be evidence of an alternative exon shown as "a" in Figure 4.6 by RACE-PCR, although it is likely to be much less expressed than other exons. It is around the same level as the tracks over exon 2-3 (which is lower than the expression seen for exons 1, 4, 5 and 6) and does not show the clear enrichment in expression seen over exons 4 and 5 which flank it. Additionally both isoforms detected with 5' RACE show that alternative splicing is occurring to produce a transcript which does not contain exons 2 or 3, shown as "b" in Figure 4.6.

The signal on the array within intron 1 of *ZFP57* was postulated to be a result of an artefact as it was seen in all three LCLs analysed on the array (COX, PGF and QBL). RACE-PCR showed that there was no detectable run-through from the first exon in the COX cDNA and so the hypothesis of an array artefact was upheld. Possible isoforms for *ZFP57* from the evidence of RACE-PCR in COX are shown in Figure 4.7.

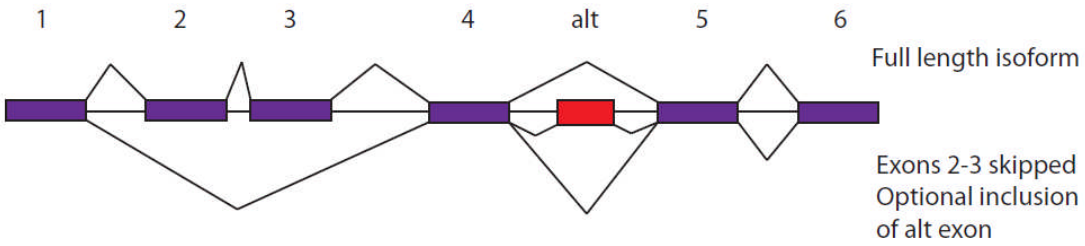


Figure 4.7: Schematic showing the different isoforms of *ZFP57* determined by RACE-PCR. The known annotated exons of *ZFP57* are shown in purple while the newly discovered alternative exon is coloured red. Splicing is indicated between the exons for the 3 different isoforms.

Primers were designed to test the RACE-PCR results, specific to exon 1, exon junction 1-4 and exon 4-alt. All resulted in single bands following PCR amplification and when visualised on an agarose gel (Data not shown). The new primer in exon 1 was also used with an existing primer in exon 4 to show the two alternatively spliced isoforms can be amplified in the same PCR reaction (Figure 4.8).

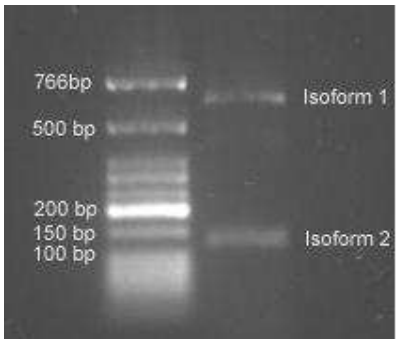


Figure 4.8: PCR showing the 2 isoforms of *ZFP57*. PCR is carried out on COX cDNA using primers located within exons 1 and 4 of the *ZFP57* gene and the products are visualised on a 1.5% agarose gel. Isoform 1 is the full length transcript including exons 2 and 3, and Isoform 2 is the "skipped" isoform where exons 2 and 3 have been excluded.

4.3.3 Expression quantitative trait mapping of *ZFP57*

As shown in Chapter 3, eQTL mapping using a cohort of 96 healthy volunteers showed a possible association of a variant, rs29228, with *ZFP57* expression. This was validated and characterised in a second cohort of healthy volunteers. This involved recruitment of 288 individuals of self-reported European ancestry over a four month period, taking a 50ml blood sample from each and purifying a mixed PBMC population as well as separated B cells and monocytes. Volunteers were genotyped

using the Illumina HumanOmniExpress-12v1.0 Beadchips for 733,202 genetic variants. After standard QC, a total of 651,210 markers were available for analysis (SNP call rate >96%, MAF > 1%). Possible underlying genetic stratification in the population was assessed by multi-dimensional scaling using data from the International HapMap Project (CEU, YRI and CHB samples) combined with IBS cluster analysis (complete linkage agglomerative clustering based on pairwise identity-by-state (IBS) distance). Four individuals demonstrating potential admixture were identified and removed from the analysis, together with one individual with low genotyping call rates and one individual with a familial relationship to another volunteer in the cohort, resulting in a final analysis of 282 individuals (122 males and 161 females). Imputation over the region surrounding *ZFP57* using the data from the 1000 genome project (Pennisi 2010) led to identification of 19,129 additional SNPs that were included in the PLINK linear analysis. The program IMPUTE2 was used to predict the genotypes at additional loci around the *ZFP57* loci, using the co-ordinates Chr6: 26400239-32199517 (Hg19), taken from the UCSC genome browser.

Gene expression for *ZFP57* was once again determined in PBMCs using qPCR and expression values were normalised to *ACTB* expression levels. Having seen alternative isoforms of *ZFP57* in the COX LCL, the primers against the exon boundary of exons 5 and 6 were again used for expression analysis. eQTL analysis was again performed as described in Section 3.3 using PLINK linear analysis and found 585 SNPs that were associated with *ZFP57* expression at GWAS significance (taken to be $p \leq 10^{-7}$). Genetic variants said to be in *cis* were here defined as those within a 250kb window flanking the gene, complying with a previous *cis* regulation study (Grundberg 2011). All SNPs available except one defined as significantly associated with *ZFP57* expression at a genome-wide level were found on Chromosome 6, with 154 typed variants and 683 imputed variants said to be in *cis* with *ZFP57* (837 in total). SNPs with the highest probability of being associated with *ZFP57* expression were found to be covering the gene itself particularly within the first intron and second exon.

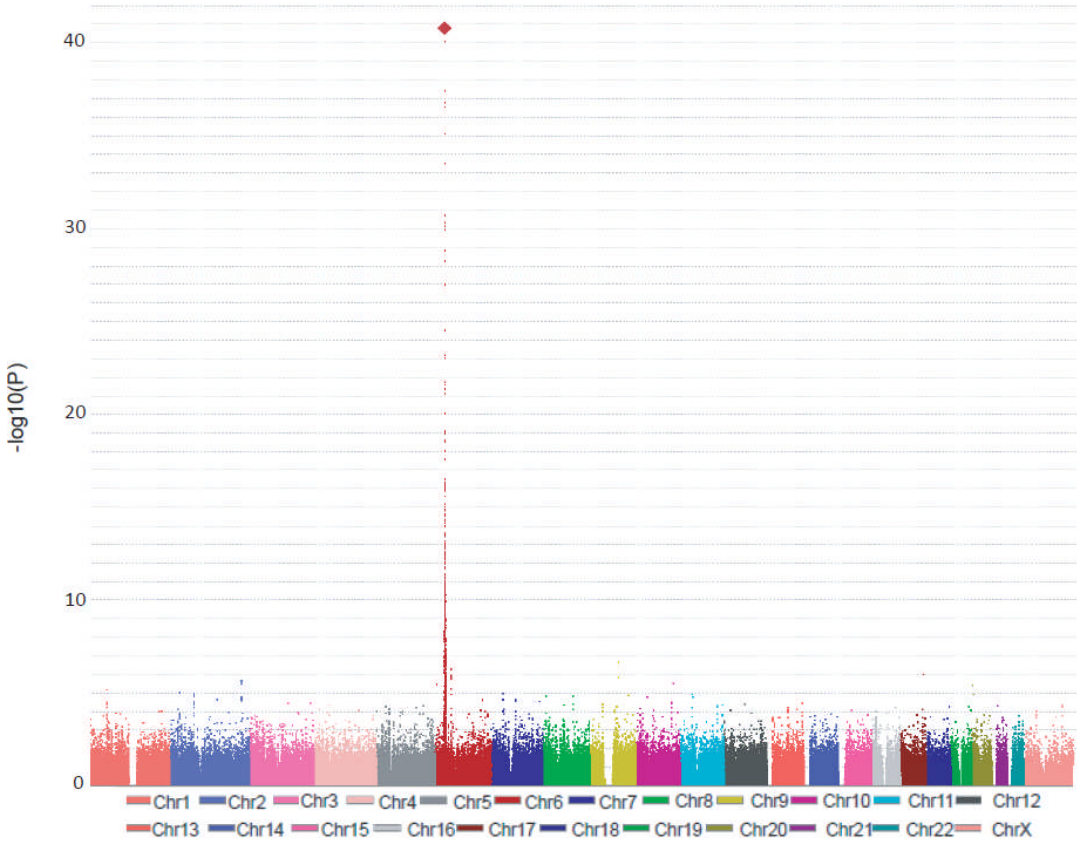


Figure 4.9: Manhattan Plots showing typed SNPs and their association with *ZFP57* expression over the whole genome. *ZFP57* expression was determined in 280 healthy volunteer samples using qPCR and normalised to *ACTB* expression. Healthy volunteer gDNA was genotyped using Illumina Infinium high-density genotyping bead arrays and genotype is associated to expression phenotype using PLINK. A peak of SNPs showing strong association with *ZFP57* is found around the *ZFP57* locus in the MHC. The top SNP associated with *ZFP57* expression (rs2535238) is indicated by the larger red diamond ($p= 1.08X10^{-41}$).

All genotyped and imputed SNPs were plotted according to p value of association in the Manhattan plot shown in Figure 4.9. The region isolated by the Manhattan plot as showing most association with *ZFP57* expression was investigated further using a recombination plot of the *ZFP57* loci and surrounding area (Figure 4.10). Recombination rates plotted in the area showed that association to *ZFP57* expression decreased after peaks of recombination as expected.

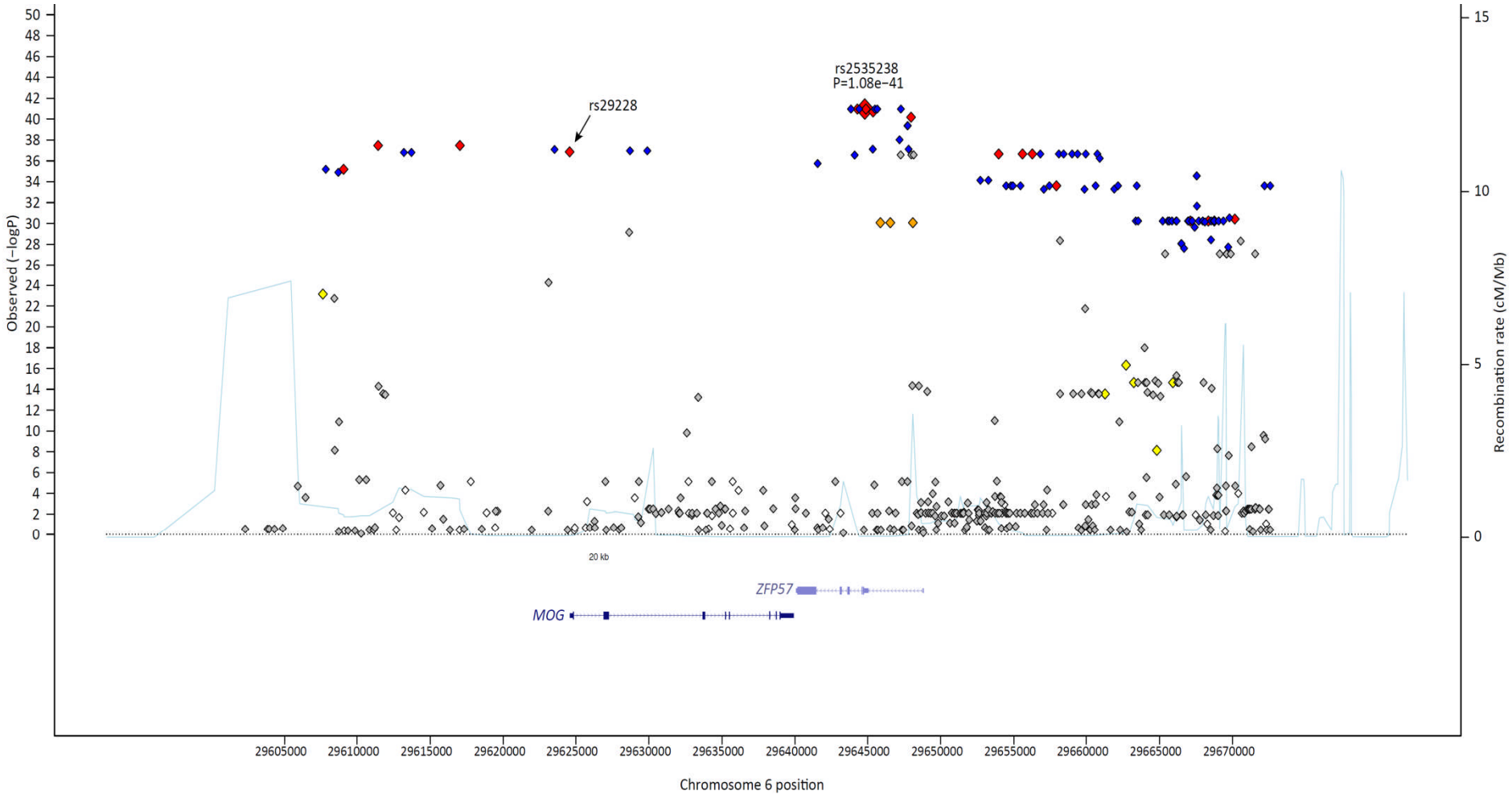


Figure 4.10: Recombination plot showing the *ZFP57* locus and surrounding recombination rates. rs2535238, the most associated SNP with *ZFP57* expression is shown in blue. Typed SNPs are shown in red (LD 0.8≤1.0 to rs2535238), orange (LD 0.5≤0.8), yellow (LD 0.2≤0.5) and white (LD 0.0≤0.2). Imputed SNPs are shown in blue (LD 0.8≤1.0) and grey (LD<0.8).

The volunteers were separated into three groups according to genotype at rs2535238, one of the most associated typed SNPs to *ZFP57* expression. Expression of *ZFP57* was then compared between these three groups, and against the three genotype groups for rs29228 (the previously most associated SNP to *ZFP57* expression) in Figure 4.11:

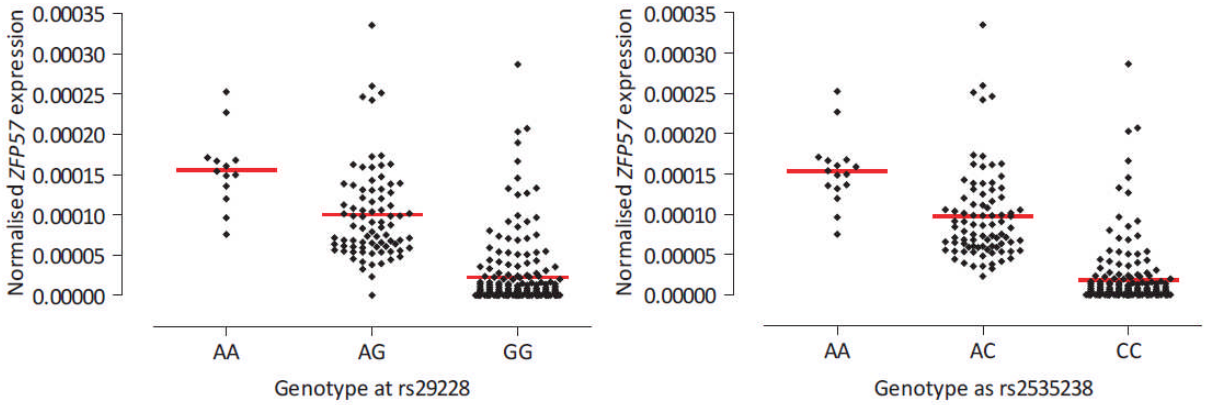


Figure 4.11: *ZFP57* expression according to genotype at i) rs29228 and ii) rs2535238. *ZFP57* expression was determined using qPCR and was normalised to *ACTB*. Expression according to genotype is shown, with mean expression indicated by the red bar. Genotyping of the volunteers DNA samples was performed using the Illumina Infinium high-density genotyping bead array. Both SNPs were found to show significant differences between the different genotype groups ($p<0.001$, Kruskal-Wallis test). R^2 between the SNPs is 0.88.

The different genotype groups for each SNP (rs22928 and rs2535238) were shown to have significantly different *ZFP57* expression, with the minor allele indicating higher *ZFP57* expression in both cases.

4.3.4 HapMap LCLs and expression of *ZFP57*

Variants in the six MHC homozygous LCLs analysed in Chapter 3 could in general be used to predict gene expression level. With this in mind, a selection of CEU and YRI cell line cDNA was taken and analysed for *ZFP57* expression. SNPs highlighted as associated to *ZFP57* expression in the volunteer cohort were used to group volunteers to assess if they could be used as predictors of *ZFP57* expression (Figure 4.12), even where the population differed from the population where the initial observation was made.

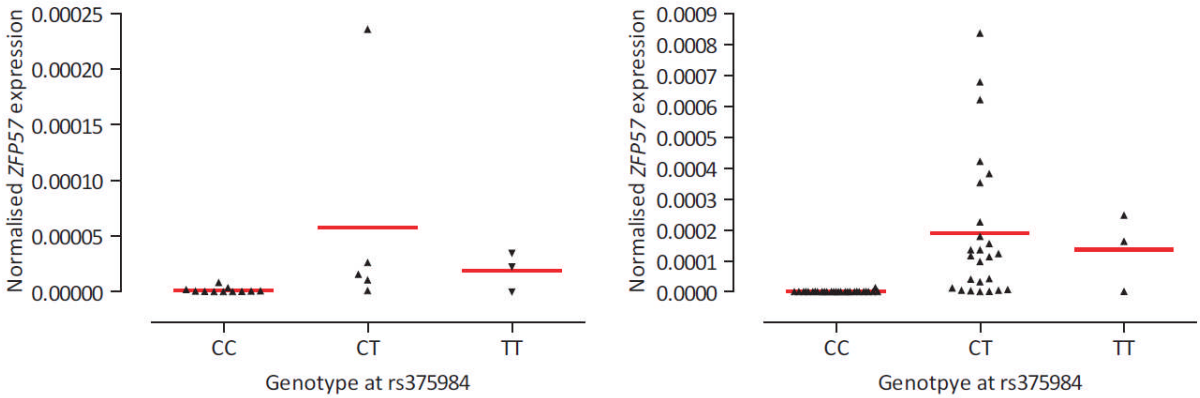


Figure 4.12: Expression of *ZFP57* in i) CEU cell lines and ii) YRI cell lines according to genotype at rs375984. *ZFP57* expression was detected using qPCR and results were normalised against *ACTB* expression. Expression according to genotype is shown, with mean expression indicated by the red bar. There was a significant difference between the different genotypes in terms of mean expression for each population (CEU: p=0.0289, YRI: p<0.0001, Kruskal-Wallis test).

Increase in *ZFP57* expression was seen in several of the LCLs that possessed the genotype showing increased expression in the volunteer cohort (rs375984-T), a SNP in complete LD with rs2535238, although only a small number of CEU LCLs were available for analysis meaning the result may not be conclusive. When genotype at rs29228, the SNP found to be most associated with *ZFP57* expression in the first volunteer cohort, was used to group the LCLs it did not predict an increase in gene expression in either the CEU or the YRI lines. This highlights the importance of fine-mapping associations to be able to use them outside of the population of original study.

4.3.5 Haplotypic analysis of *ZFP57* expression

Haploview was used to analyse the LD and association of haplotype to *ZFP57* expression further (Barrett 2005). The regional LD is shown in Figure 4.13 and the block most associated with *ZFP57* expression is indicated.

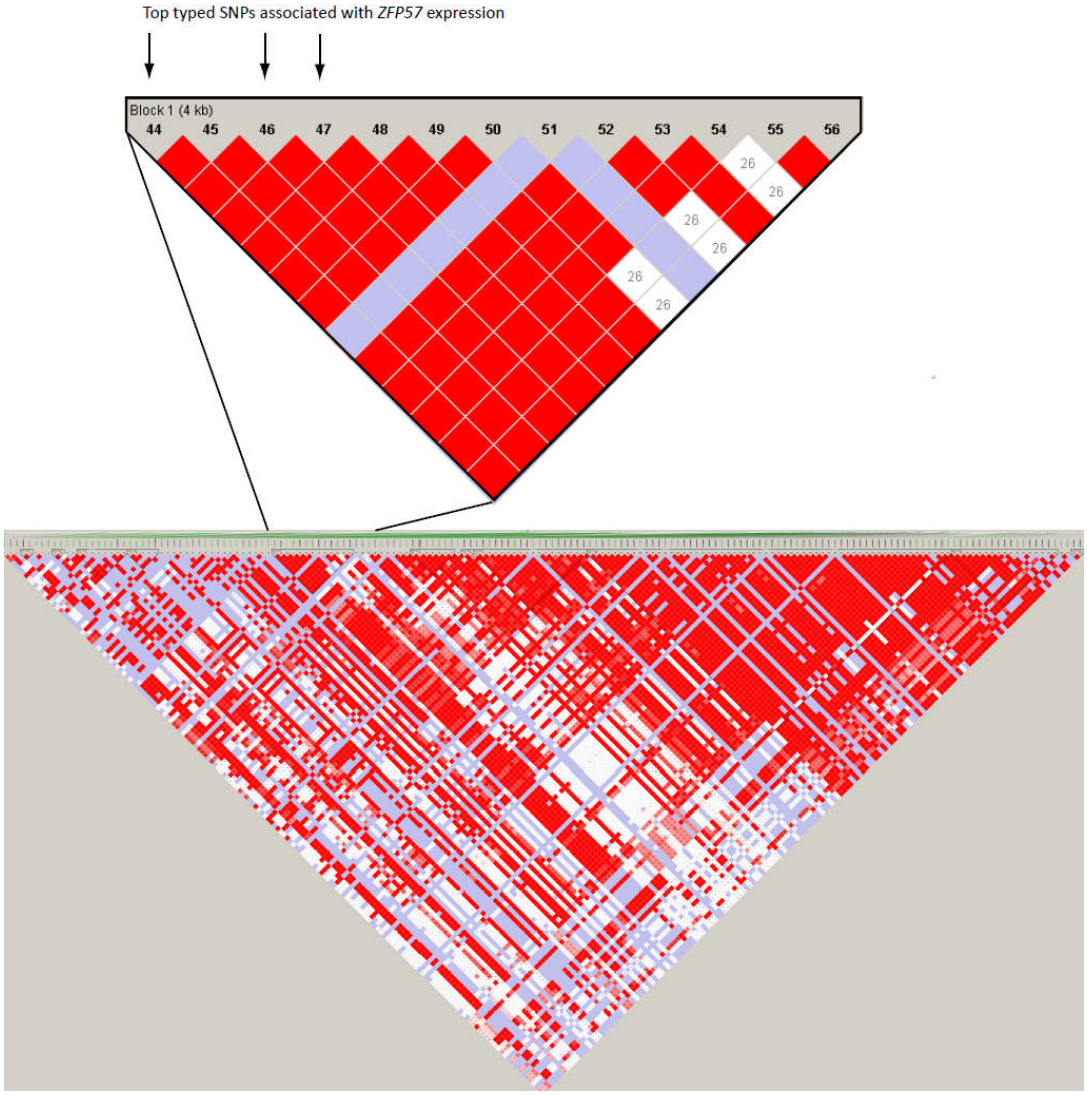
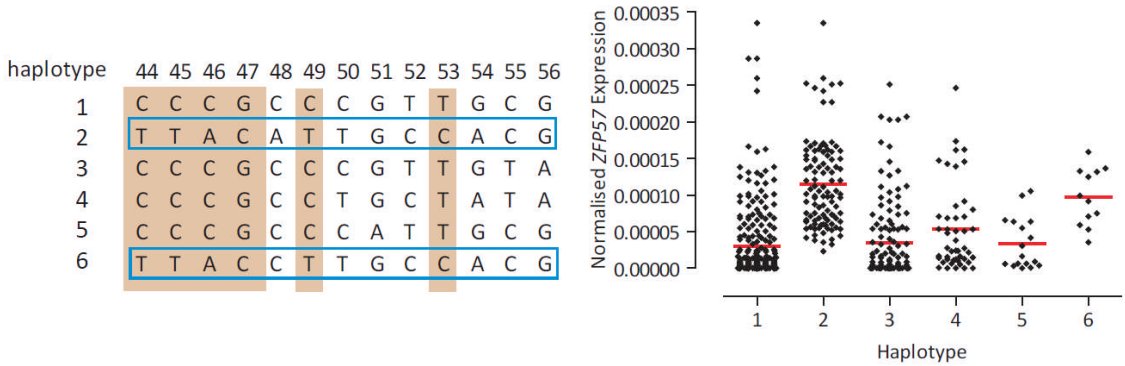


Figure 4.13: Haploview analysis of LD and haplotype structure in the *ZFP57* region. Genotyped SNPs from the volunteer study and HapMap genotyped SNPs of greater than 1% minor allele frequency were used to infer haplotypic structure with LD blocks predicted by the *Confidence Intervals* algorithm (Gabriel 2002b) using HaploView software (Barrett 2005). Red diamonds show allelic association (measured by the D' statistic) between two SNPs with an LOD>2; the darker the shade of red, the higher the value of D', where 1 is the highest value. White squares indicate no statistically significant evidence of LD but have an LOD>2. Blue squares show no LD and have an LOD<2 between SNPs. Typed SNPs rs375984, rs2535238 and rs2747421 that show most association with *ZFP57* expression are indicated.

The block most associated with *ZFP57* expression has six inferred haplotypes. The "A" genotype at rs2535238 is associated with increased expression, and is found in two of the possible haplotypes. These haplotypes differ by only one variant. This SNP (rs3129063, SNP 48 in Figure 4.14) was not typed directly on the Illumina Infinium high-density genotyping bead array; it was however imputed in the subsequent imputation analysis over the *ZFP57* region.

The six possible haplotypes for each region were defined in the volunteer samples and the mean *ZFP57* expression for each haplotype and its frequency, compared to the expected frequencies for each haplotype, is shown in Figure 4.14



Haplotype	Predicted frequency	Actual frequency	Mean Expression *ZFP57*
1	0.425	0.451613	$2.96x10^{-5}$
2	0.25	0.184588	$1.15x10^{-4}$
3	0.2	0.227599	$3.51x10^{-5}$
4	0.067	0.084229	$5.35x10^{-5}$
5	0.033	0.030466	$3.39x10^{-5}$
6	0.025	0.021505	$9.74x10^{-5}$

Figure 4.14: Haplotype block analysis and *ZFP57* expression. The six possible haplotypes over the region are shown with those associated with higher ZFP57 expression by their possession of an "A" at rs2535238 (46) outlined in blue. Shaded SNPs are indicative of those that are informative of the higher expression phenotype and indicated possible functional SNPs. Average expression of *ZFP57* by haplotype is shown normalised to *ACTB* expression, and the frequencies of each haplotype are indicated as both the predicted proportion and the actual proportion of the haplotypes seen in the volunteer cohort.

Haplotype analysis did not further define the functional SNP as no significant difference in *ZFP57* levels was seen between haplotypes 2 and 6 (Mann-Witney test p=0.41), which showed higher expression. A significant difference (p<0.0001) was seen performing a Mann-Witney test between the two groups (group 1 comprising haplotypes 1, 3, 4 and 5; group 2 comprising haplotypes 2 and 6).

4.3.6 Imputed HLA type and *ZFP57* expression

The 2-digit HLA type generally refers to the specific serological antigen of the allele. The serotype of an allele is usually determined by the antibody binding specificity. Traditionally, HLA types have been

identified experimentally by serological testing; however more recently DNA sequence based analysis has been used. To accomplish this, variants in exons 2 and 3 for class I molecules and exon 2 for class II molecules are determined which in turn predict the HLA type. These correspond to the binding cleft of the peptides and thus show structural variation that could be affecting the function of the molecule. All identified HLA alleles can be found at the HLA nomenclature website (http://hla.alleles.org/nomenclature/naming.html).

Using the genotype information for the cohort of 288 volunteers, the 2 and 4-digit HLA types at six regions (HLA-A, HLA-B, HLA-C, HLA-DQA, HLA-DQB and HLA-DR) were imputed with the assistance of Stephen Leslie and Alexander Dilthey (Leslie 2008). Briefly, an informative panel of SNPs typed in the healthy volunteers was selected based on their ability to be used to inform HLA type. Data was available for SNP genotyping and HLA type in a training panel, and this was used to select suitable SNPs from the genotyping data for the healthy volunteers. The known data was split so that a third could be HLA typed using the SNP panel selected to determine the accuracy of the prediction. Phasing of the imputed HLA types was carried out using phasing algorithms. These carry the possibility of a switch error whereby local phasing is accurate but over a larger scale the alleles have "switched" chromosomes from their original placement, but this is thought to affect a minority of samples due to extensive LD in the MHC. Following QC, the 2-digit types were deemed to be accurate over all predicted HLA types (see Section A.3.3), whereas several of the 4-digit types were predicted with low confidence. Therefore *ZFP57* expression determined by qPCR in PBMCs was plotted in Figure 4.15 according to 2-digit type over the six classical HLA gene regions to assess if any particular MHC alleles were associated with increased expression.

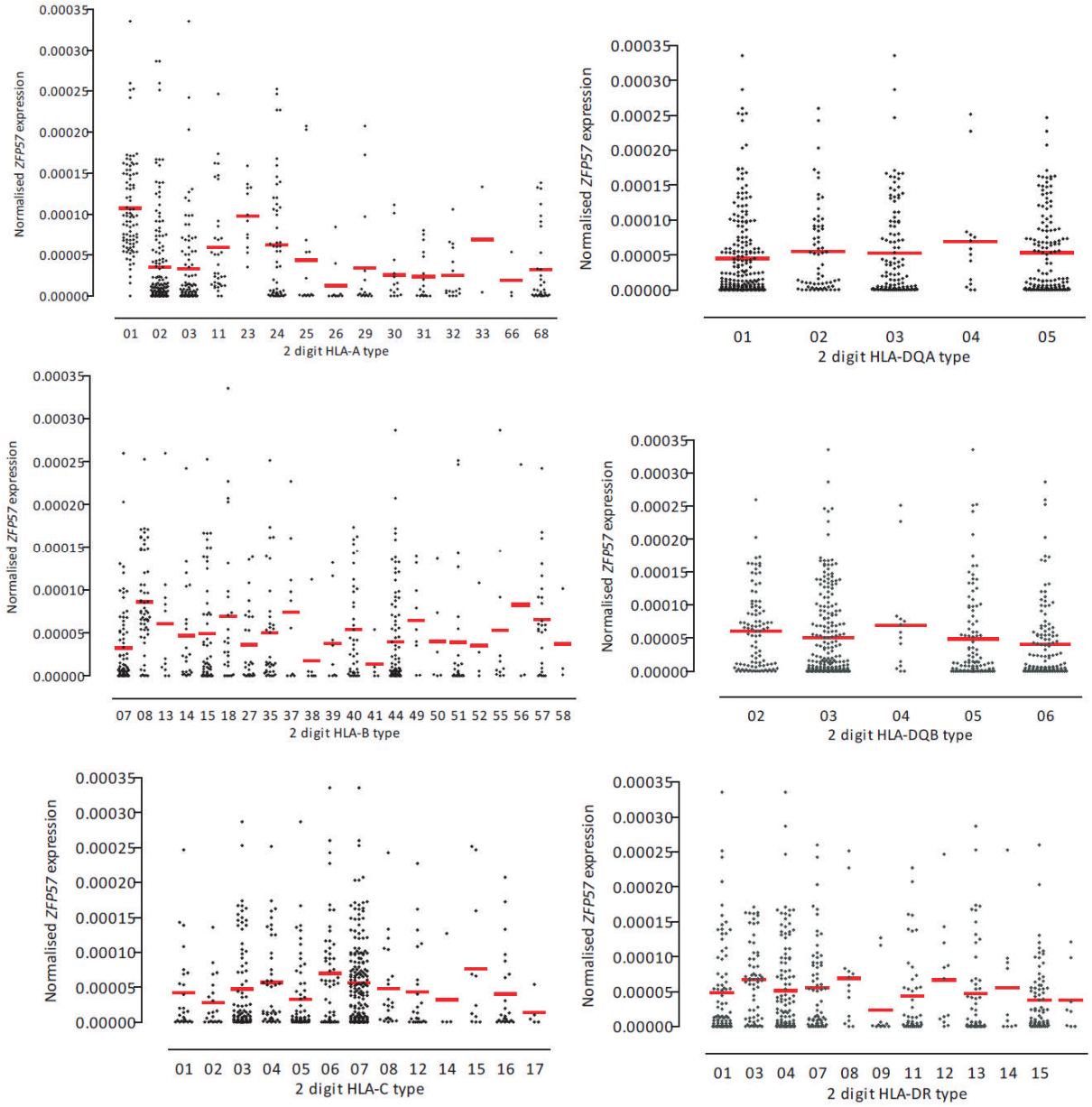


Figure 4.15: Expression of *ZFP57* according to 2-digit haplotype at i) HLA-A, ii) HLA-B, iii) HLA-C, iv) HLA-DQA, v) HLA-DQB and vi) HLA-DR. Haplotype imputation in the volunteer cohort was performed for 283 volunteer samples. Volunteers were separated by 2-digit haplotype at each locus to determine if there were any association with *ZFP57* expression. Mean expression of *ZFP57* in volunteers with each allele is indicated.

Gene expression was most variable between the HLA-A alleles, with HLA-A*01 and HLA-A*23 showing higher expression ($p<0.0001$ when analysed using a Mann-Whitney test). To further define this, expression was plotted for the different haplotypes of HLA-A, using the genotype at rs2535238 as a covariate:

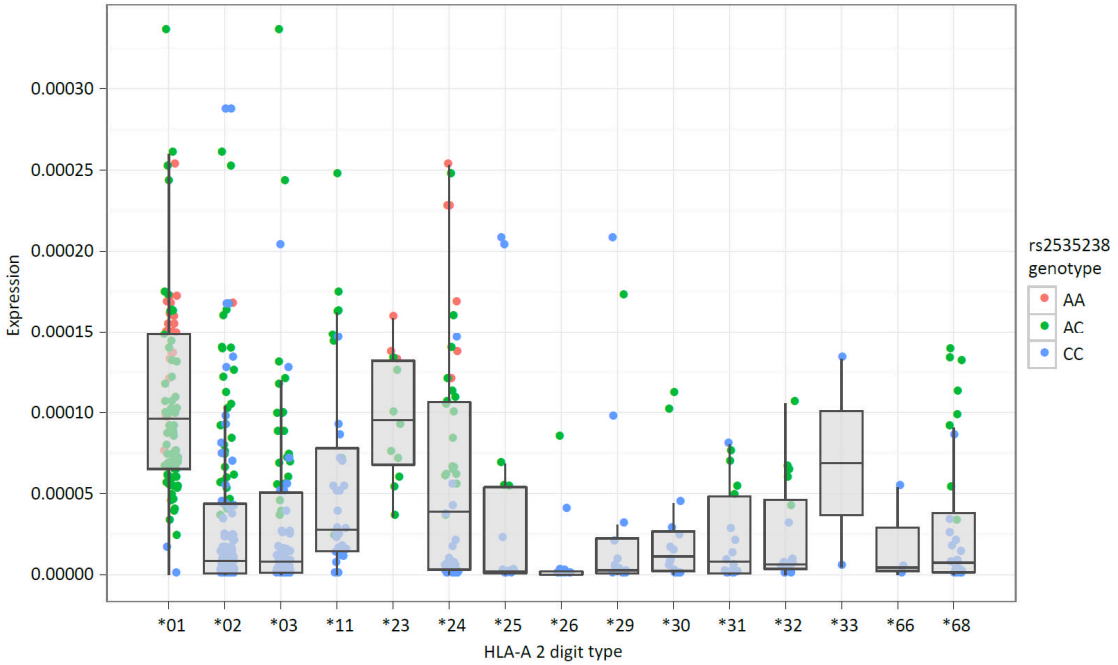


Figure 4.16: Expression of *ZFP57* according to 2-digit type at HLA-A. Expression of *ZFP57* as determined using qPCR was plotted against HLA-A 2-digit haplotype determined by imputation for 283 healthy volunteers. Genotype at rs2535238 was used to colour the individual expression values; AA (associated with highest *ZFP57* expression) red, AC green and CC blue.

Figure 4.16 shows that the HLA types HLA-A*01, *23 and *24 were most likely to contain the SNP associated with high expression. As HLA-A*23 and HLA-A*24 are sister types of the broader HLA-A*09 type (Fussell 1996) their shared possession of a particular SNP is unsurprising. To analyse if there was a dose dependent effect of these HLA types, expression of *ZFP57* was plotted in Figure 4.17, where the volunteers were grouped by number of copies of HLA-A*01, *23 and *24.

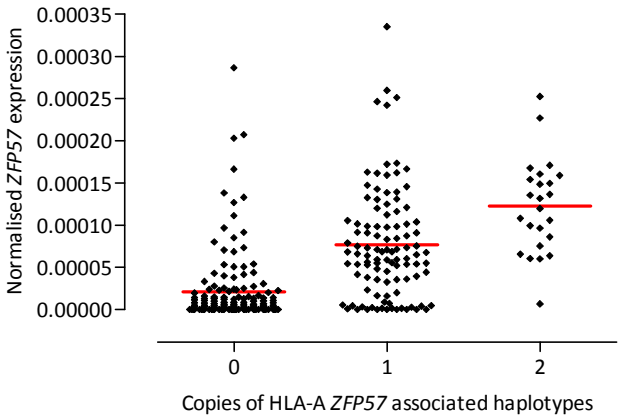


Figure 4.17: Expression of *ZFP57* according to copy number of HLA-A *01, *23 and *24. Healthy volunteers were grouped according to copy number of *ZFP57* expression associated HLA-A alleles. Expression of *ZFP57* as determined using qPCR was plotted and a significant difference in mean expression (indicated by the red bar) was observed between the three groups (Kruskal-Wallis test $p<0.0001$).

As the group of volunteers possessing a copy of the HLA-B*08 allele also appeared to have higher *ZFP57* expression than other groups, expression of *ZFP57* was also analysed in the volunteers grouped by copy number of HLA-A*01, *23 and *24 in the presence or absence of one or more copies of HLA-B*08 (Figure 4.18).

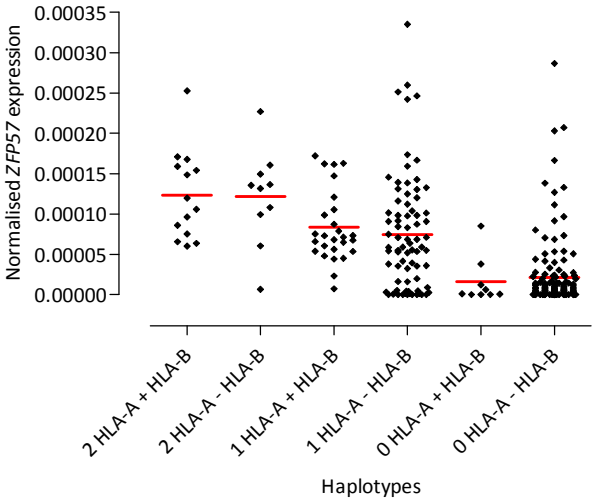


Figure 4.18: Expression of *ZFP57* according to copy number of HLA-A* 01, *23 and *24 and presence or absence of HLA-B*08. Healthy volunteers were grouped according to copy number of *ZFP57* expression associated HLA-A haplotypes and whether or not these individuals also possessed a copy of the HLA-B*08 haplotype. Mean expression is indicated by the red bar. No significant difference in expression was seen between the means of each 2 HLA-A groups (with or without a copy of HLA-B*08) using a Mann Whitney test; 2 copies HLA-A*01, *23, *24 (p=0.98), 1 copy HLA-A*01, *23, *24 (p=0.14), 0 copies HLA-A*01, *23, *24 (p=0.53).

Expression can be seen to clearly vary in response to copy number of HLA-A*01, *23, or *24, but there is no difference between the mean expression values of the groups when segregated by HLA-B*08 presence. Haplotype association with *ZFP57* expression is therefore confined to the HLA-A locus.

The COX ancestral allele contains the HLA-A*01 allele that has been shown to be associated with increased *ZFP57* expression. Healthy volunteers who possessed a copy of one of the extended conserved haplotypes of the three MHC homozygous LCLs (COX HLA-A1-B8-Cw7-DR3; PGF HLA-A3-B7-Cw7-DR15; and QBL HLA-A26-B18-Cw5-DR3-DQ2) (Horton et al.2008) were identified. In the cohort, 19 individuals were found to have at least one copy of the COX haplotype and 12 individuals

were found who had a copy of the PGF haplotype at 2-digit resolution. No individuals with the QBL haplotype were found in the volunteer cohort, reflecting its far less common presence in the population. When comparing the expression of *ZFP57* between individuals who possessed a copy of the full COX haplotype to those with a copy of the full PGF haplotype in Figure 4.19, a clear difference in expression levels was seen:

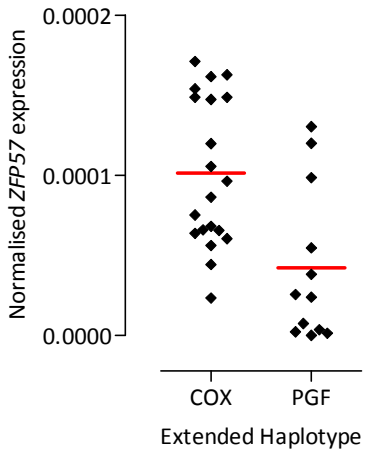


Figure 4.19: Expression of *ZFP57* in individuals with the COX or PGF extended haplotype. Volunteers who possessed either a copy of the COX extended haplotype, or the PGF extended haplotype were compared to determine if a significant difference in expression of *ZFP57* could be seen. A Mann-Whitney test between mean expression of each group found them to be significantly different ($p<0.0001$).

4.3.7 Functional consequences of SNPs associated with differential *ZFP57* expression

SNPs showing the strongest association with *ZFP57* expression were investigated using the protein binding prediction software JASPAR (Bryne 2008) to determine if the allele showing association with expression destroyed or created a new binding site for a TF.

SNP	Predicted TF Binding site	Status	95% threshold	ENCODE data?	Overlap
rs374317	HIFA::ARNT	Destroyed	No	No	NA
	TFAP2A	Destroyed	No	No	NA
	MZF1_5-13	Created	No	No	NA
	GATA2	Created	**Yes**	**HUVEC**	**No**
	ELK4	Created	No	No	NA
rs375984	ETS1	Destroyed	No	**GM12878**	**No**
	Myf	Destroyed	No	No	NA
	MZF1_5-13	Created	No	No	NA
rs376319	ZNF354C	Destroyed	No	No	NA
	YY1	Created	No	**GM12878**	**No**
	GATA3	Created	No	No	NA
	GATA2	Created	No	**HUVEC**	**No**
rs387640	HOXA5	Destroyed	No	No	NA
	INSM1	Created	No	No	NA
	BRCA1	Created	No	**GM12878**	**No**
rs448489	None	NA	NA	NA	NA
rs2535238	SOX 10	Created	No	No	NA
rs2747421	SOX 10	Created	No	No	NA
	FOXA1	Created	No	**HepG2, T-47B, ECC-1**	**No**
rs3129062	ZNF354C	Destroyed	No	No	NA
	NFIC	Created	**Yes**	No	NA

Table 4.1: TF binding sites affected by *ZFP57* expression associated SNPs. SNPs most associated with *ZFP57* expression by p value ($p=1.08x10^{-41}$) were interrogated using JASPAR to assess their impact on potential TF binding sites. Binding sites affected by sequence variation of the relevant SNPs are shown using the default significance threshold of 80% and detailing if they are created or destroyed by the non-reference allele. Additionally, whether they remain likely to affect binding using a significance threshold of 95% was tested to define the strength of the potential binding site. ENCODE TF binding data was used to search for all TFs identified by JASPAR and cell lines where there was whole genome binding data available are indicated. If ENCODE data was available this was analysed over the *ZFP57* region to determine if known binding data overlapped with the JASPAR predictions.

Table 4.1 shows the binding sites created and destroyed at 80% relative score threshold, and whether they remain significant if the threshold is increased to 95%. Prediction of TF binding sites *in silico* is known to have significant selectivity and specificity problems hence the use of two thresholds for analysis. While binding sites identified by JASPAR as being disrupted by SNPs may show effects if tested *in vitro,* only a minority of predicted sites also identify a true *in vivo* interaction. The recently released in vivo TF binding data from ChIP-seq as part of the ENCODE Project (Myers 2011) was used to see if any of the *in silico* identified binding sites had any *in vivo* evidence for their presence. None

of the binding sites predicted to be destroyed were detected by ENCODE ChIP-seq (although relatively few of the predicted TFs had been investigated by ENCODE). There was also no evidence of TF binding of any of the factors predicted to have a binding site created by the alternative genotype. As no convincing candidates for TF binding disruption by SNPs were found this possibility was not investigated further by Electrophoretic Mobility Shift Assay or ChIP.

4.3.8 Chromatin accessibility, modifications and TF binding at *ZFP57*

ENCODE data was used to analyse DHSs, TF binding and histone marks over the *ZFP57* gene to determine the transcriptional landscape over this region and interpret this in the context of observed eQTL SNPs. DHSs can predict regions of actively transcribed or regulatory DNA as the chromatin structure has been relaxed to allow protein access to the region. Likewise, histone marks are also useful in predicting areas of the genome that are important in regulation as particular marks such as methylation and acetylation are associated with regulation, transcription and gene silencing (Heintzman 2009). Identification of TFs that bind in a region may give insights to specific pathways the gene is involved in.

Despite finding no overlap between binding sites predicted by JASPAR over the most associated SNPs, some TF binding sites have been annotated over the region of *ZFP57* most associated with its expression. DHSs overlap with many of these TF binding sites, implying they are functional and chromatin structure is altered to allow access to the DNA. DHS sites annotated by ENCODE are also found to overlap with DHS sites predicted by the MHC array (see Figure 3.2).

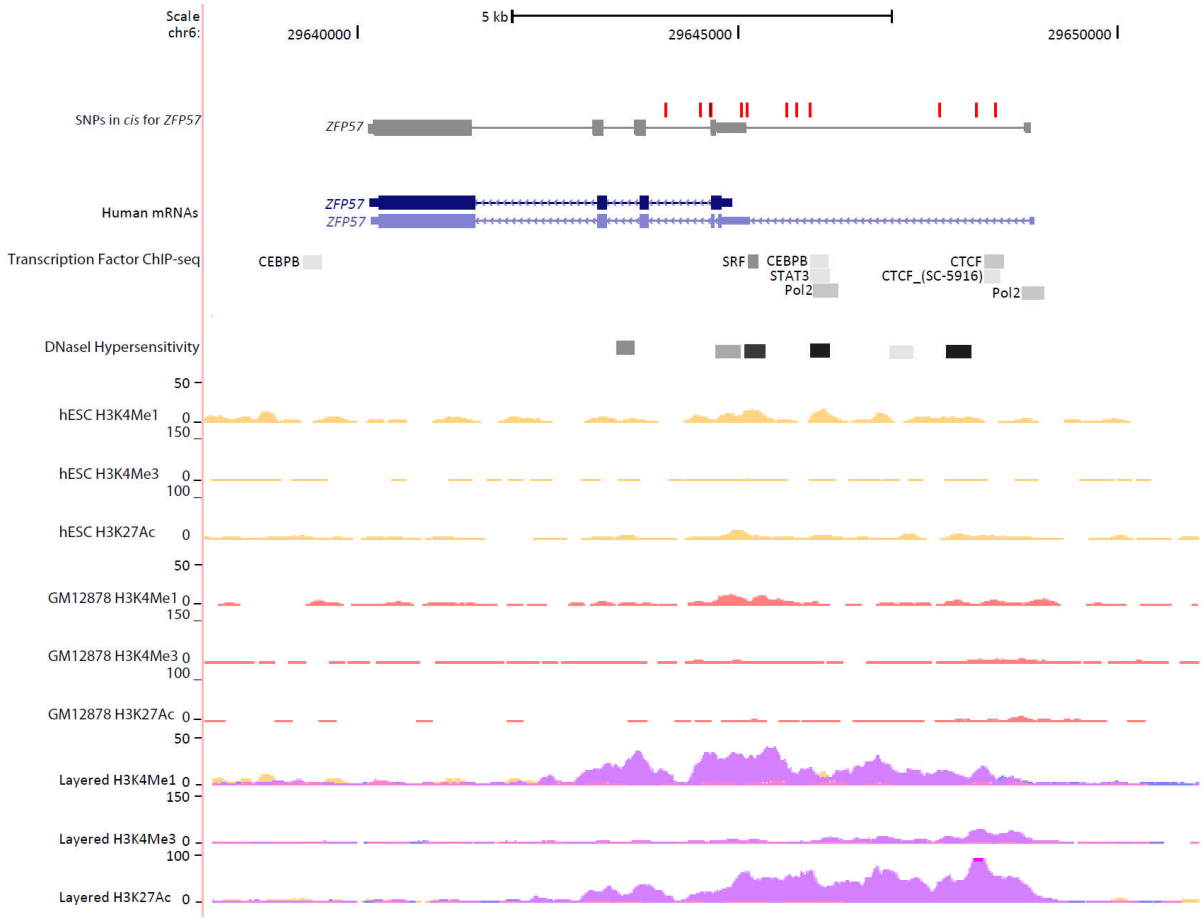


Figure 4.20: ENCODE data regulatory events in *ZFP57*. Two isoforms of *ZFP57* identified by Refseq are shown with the positions of the most significantly associated SNPs to expression (typed and imputed). DHS clusters, annotated by ENCODE, are seen over the first intron and second and third exons. Several TF binding sites are identified in the region, including a Serum Response Factor (SRF) binding site near the most associated SNP rs2535238. ChIP-seq data for three histone marks is shown; H3K4Me1 (often indicating regulatory elements), H3K4Me3 (often indicating a promoter region), and H3K27Ac (often indicating active regulatory elements) in H1-hESC cell line (yellow), GM12878 (red) and layered in all seven cell lines investigated by ENCODE.

In the NHEK cell line (Normal Human Epidermal Keratinocytes - Adult) shown in pink in Figure 4.20, evidence is seen of regulatory marks (indicated by H3K4Me1 and H3K27Ac), overlapping the SNPs most associated with *ZFP57* expression, as well as the TF binding sites and the DHS sites. The mark H3K4Me3 generally found near promoter regions is less obvious, however again in the NHEK cell line there is evidence of an increase in the mark around the 5' end of the gene in the first intron. In the two cell lines selected to be investigated individually (hESC as it is an embryonic stem cell line) and GM12878 (as it is a CEU derived LCL) the histone modification marks are not so striking but there is

evidence of regulatory marks associated with the area overlapping the SNPs most associated with *ZFP57* expression.

4.3.9 Confirmation of ZFP57 protein translation in cell lines

The translation of ZFP57 protein was confirmed by western blot in three LCLs homozygous across the *ZFP57* genomic region (Figure 4.21). LCLs COX and APD were used as examples of cells with "high" *ZFP57* expression, which would be expected to have higher protein expression, while PGF was used as a "low" expresser of *ZFP57*. Jurkat nuclear lysate was also run on the western as this was used as a positive control by Abcam the manufacturers of the α-ZFP57 antibody.

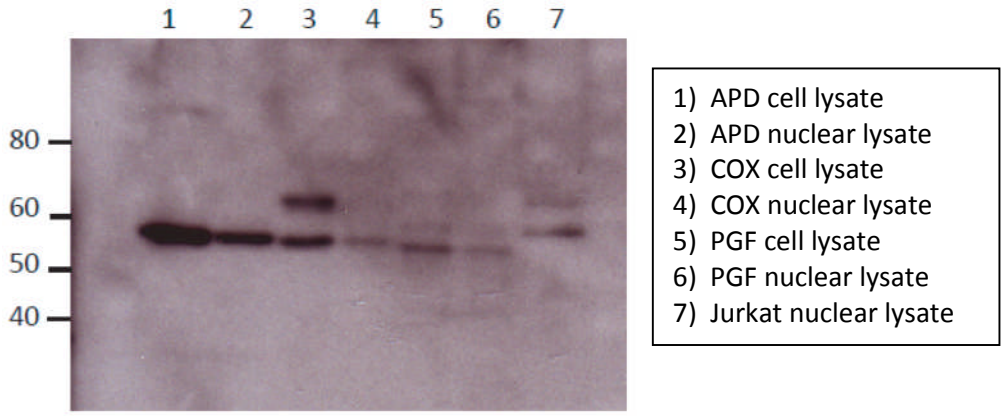


Figure 4.21: ZFP57 expression in LCL cell and nuclear lysate confirmed by Western Blot. Expression of ZFP57 is detected using α-ZFP57 antibody and detected using α-Rabbit secondary antibody conjugated to HRP. Concentration of the lysate for all cell lines was determined by BCA assay and 10µg was loaded in each lane. Size of protein in kDa is indicated on the left hand side of the blot, determined by comparison with known protein standards. Expression is seen in all cell lines analysed in both cell and nuclear lysate around 55kDa. A loading control (KAP1) of the same gel can be seen in Figure 7.3.

ZFP57 protein was detectable in all three LCLs at the 55kDa mark, the correct molecular weight for ZFP57 protein. Notably there was less ZFP57 protein in the PGF nuclear and cell lysates compared to the higher transcript expressing cell lines APD and COX. Two bands were seen in the COX cell line in the cell lysate, perhaps indicating different isoforms of ZFP57 are present caused by alternative splicing or post-translational modification, and a second very faint band was observable in the PGF cell line. The ZFP57 antibody (ab50944) was raised against an N-terminal portion of the protein (from within amino acids 20-50) so it is unclear how many different isoforms of ZFP57 could be

detected using this antibody. The "skipped" isoform identified in Section 4.3.2 would be undetectable using this antibody as the start site for translation would not be included so the protein would be curtailed. This could account for the comparatively faint ZFP57 band in the COX nuclear lysate sample. Two bands were also detected in the Jurkat nuclear lysate; this was comparable to the result seen by Abcam in their positive control.

4.3.10 Protein prediction of function of ZFP57

ZFP57 has been previously annotated as a KRAB-ZNF protein. Due to its localisation and the affect of mutations in the gene, its function is assumed to be a DNA binding regulatory protein (Okazaki 1994, Hirasawa and Feil 2008). However, the functional impact of different isoforms of the gene has not been assessed.

Prediction of functionally important sites in the ZFP57 protein using a conserved domain search on the amino acid sequence shows that the KRAB domain presumed functional for DNA binding is close to the start of the protein (Figure 4.22). When exons 2 and 3 have been spliced out, the start site for transcription (the coding position for the initial methionine residue is found in exon 2) is missing. The next possible start site would mean that the KRAB domain of the protein is curtailed.

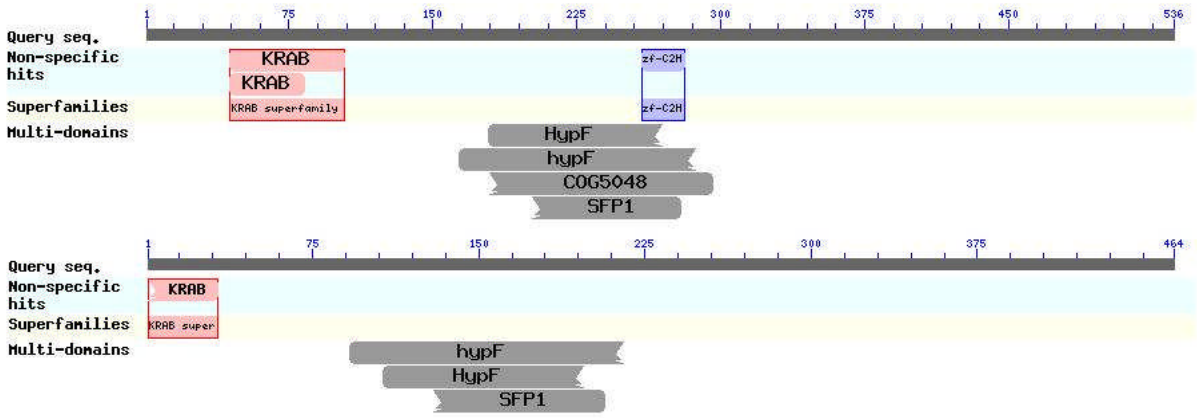


Figure 4.22: Potential changes in predicted structural elements of ZFP57. Conserved protein domains of ZFP57 are shown following analysis of the full-length sequence (top) and the shortened sequence (below) (from the first Methionine residue in exon 4) generated using the conserved domain software from NCBI (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml). The KRAB domain is shortened extensively in the shorter isoform.

4.3.11 Expression of *ZFP57* in different tissues

ZFP57 expression was extremely low in the brain tissue sample analysed when compared to *MOG* expression, however its expression in other tissues after development was still unknown. A panel of mixed RNAs from different individuals across several different tissue types was analysed for *ZFP57* expression in Figure 4.23 to determine how expression varies across tissues.

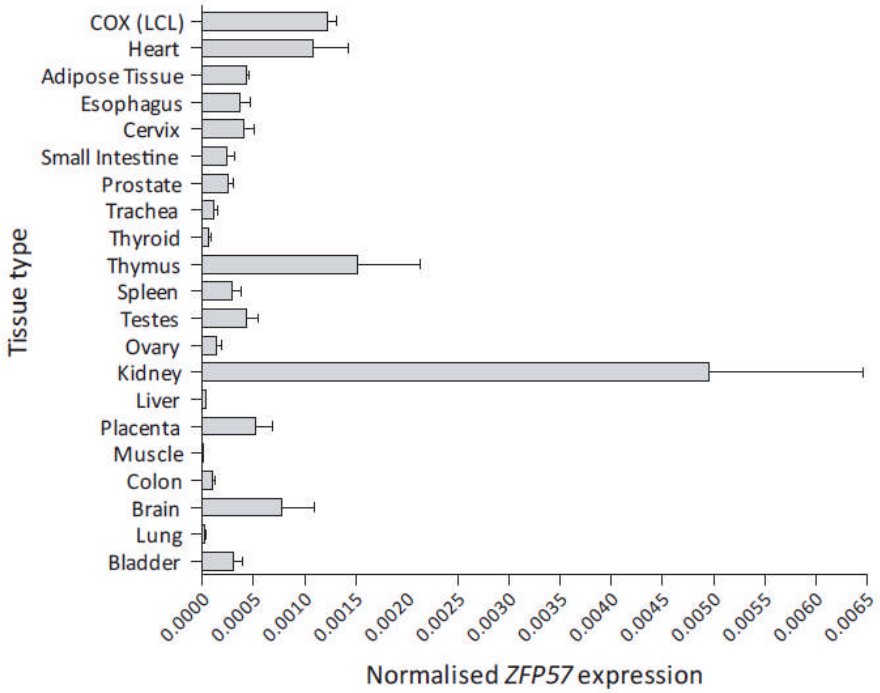


Figure 4.23: *ZFP57* expression in different tissue types. cDNA was made from RNA samples of pooled adult individuals' tissue samples. *ZFP57* expression was determined using qPCR in three replicates against exon junction 5-6 and was normalised to *ACTB*. Mean expression and standard deviation is shown. The COX LCL expression previously determined is included for comparative reference.

The heart and thymus were found to have similar expression of *ZFP57* to the COX cell line, with much higher expression being found in the kidney. Genetic information was not available for the tissue panel; however the wide range of expression across the different tissues suggests that there may be an unreported function for ZFP57 beyond development.

4.3.12 Expression of ZFP57 protein in primary B cells and monocytes

To confirm ZFP57 translation in primary cells in addition to LCLs, FACS was used to detect ZFP57 expression in both PBMCs labelled with fluorescent antibodies for cell type markers and in purified

cell populations and the results are shown in Figure 4.24. B cells and monocytes were separated using the Miltenyi MACS separation system from PBMCs from a healthy volunteer. ZFP57 was detectable in all samples using the α-ZFP57 primary antibody raised in rabbit and a fluorescently labelled secondary α-rabbit antibody. However the separated B cells and monocytes gave clearer results and showed a larger difference between the secondary antibody only control and the detected ZFP57. ZFP57 was seen to have higher protein expression in the monocytes compared to the B cells both in the PBMC sample and in the purified cell populations.

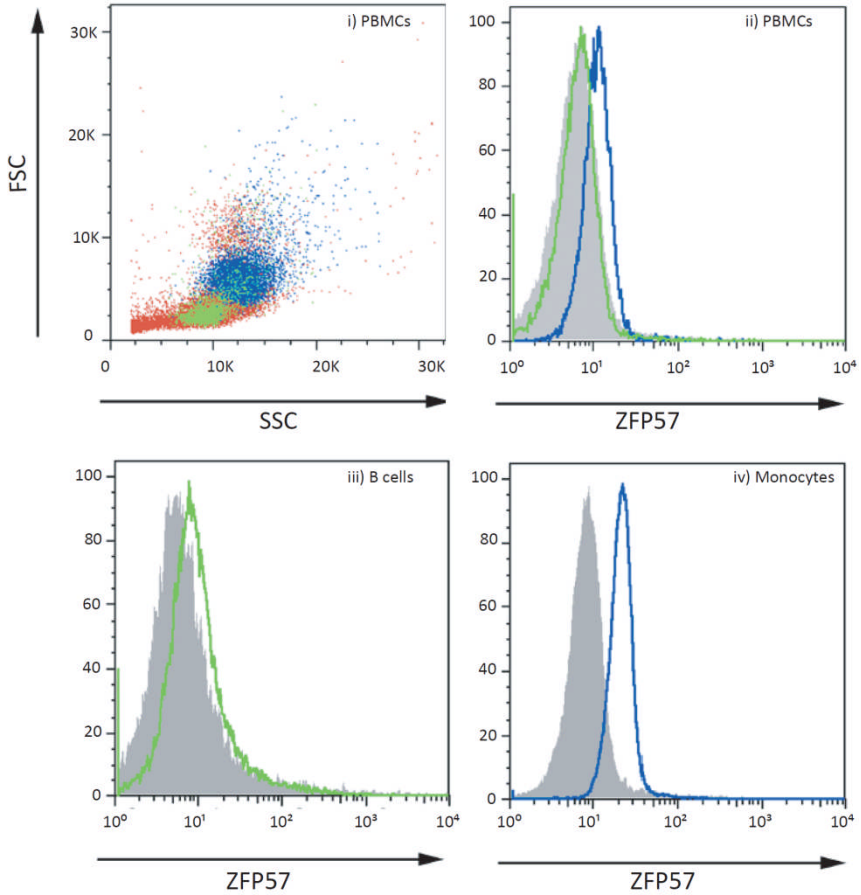


Figure 4.24: FACS analysis PBMCs (i and ii), B cells (iii) and monocytes (iv) for intracellular ZFP57 expression

i) Scatter pattern of the PBMC mixture. Cell surface markers are targeted with fluorescently labelled antibodies to determine B cells (colour) and monocytes (colour) from the remaining PBMCs. ii) PBMCs labelled for cell type and detecting ZFP57 versus a no primary antibody control. iii) Separated B cells stained for ZFP57 versus a no primary antibody control. iv) Separated monocytes stained for ZFP57 versus a no primary antibody control.

4.3.13 Differentially expressed genes associated with *ZFP57*

A large genome-wide gene expression dataset was generated by Fairfax and colleagues in the Knight

lab for separated B cells and monocytes using the Illumina HumanHT-12 v4 BeadChip platform with 48,804 probes to detect gene expression. Using the Partek Genomics Suite, *ZFP57* gene expression determined by qPCR was designated a variable phenotype. Genes that had variable expression associated with *ZFP57* expression were potential candidates for genes controlled by or controlling *ZFP57*. The results of this analysis are shown in the Table 4.2.

Monocytes		**B cells**	
Gene ID	**p value association**	**Gene ID**	**p value association**
ZFP57	9.38×10^{-12}	*HLA-A29.1*	1.01×10^{-8}
HLA-A29.1	1.68×10^{-10}	*HLA-H*	4.22×10^{-8}
HLA-H	6.50×10^{-7}	*ZFP57*	5.31×10^{-8}
HSPB7	8.56×10^{-7}	*HS.224794*	2.55×10^{-6}
HLA-A	1.27×10^{-6}	***BTN3A2***	$\mathbf{3.78 \times 10^{-6}}$
CEMP1	7.17×10^{-6}	*HLA-G*	5.49×10^{-6}
BTN3A2	$\mathbf{1.08 \times 10^{-5}}$	*UBE2Q2*	1.12×10^{-5}
BTN3A2	$\mathbf{2.94 \times 10^{-5}}$	*HCG4*	2.62×10^{-5}
LCA5	4.53×10^{-5}	*RPS3A*	3.03×10^{-5}
LIG3	4.84×10^{-5}	*LOC645231*	3.43×10^{-5}
HLA-G	5.83×10^{-5}	*LOC389223*	3.79×10^{-5}
HS.512285	8.40×10^{-5}	*PAPSS1*	4.46×10^{-5}
LCE5A	8.78×10^{-5}	*PMS2L4*	6.21×10^{-5}
ZBTB8OS	9.46×10^{-5}	*C5ORF38*	7.23×10^{-5}
KCNC4	1.02×10^{-4}	*AGPAT2*	7.27×10^{-5}
SNCG	1.30×10^{-4}	*LOC400446*	8.61×10^{-5}
HS.544222	1.61×10^{-4}	*MAP2K6*	8.83×10^{-5}
LOC100128474	1.64×10^{-4}	*TLE6*	9.43×10^{-5}
LOC729446	1.69×10^{-4}	*ATP1B3*	1.02×10^{-4}

Table 4.2: Genes associated with *ZFP57* expression determined from genome-wide expression profiles of B cells and Monocytes. Genome-wide gene expression for B cells and monocytes were correlated with *ZFP57* expression in PBMCs determined by qPCR. Probes that are associated with variance in *ZFP57* expression in PBMCs are shown in the table with their p value for association in each cell type.

Some probes show significant association to *ZFP57* expression variance however few are shared in both B cells and monocytes. As ZFP57 protein has been detected at very low gene expression levels, for example in the PGF LCL (see Figure 4.21), and both B cells and monocytes showed evidence of ZFP57 protein expression with FACS analysis, only associations found in both cell types were considered for further functional analysis. It was likely that associations with HLA genes were the result of shared haplotypic association, so were not investigated further.

4.3.14 Validation of *ZFP57* association with *BTN3A2* expression

BTN3A2 (Butyrophilin subfamily 3 member A2 isoform a precursor) was identified as being associated with *ZFP57* expression in both B cells and monocytes. While not the most significant association it was seen in both B cells and monocytes and with two separate array probes in the monocytes, which were thought to have higher *ZFP57* expression (see Figure 4.24). Expression of *BTN3A2* was investigated in the mixed PBMC samples from the healthy volunteer cohort using qPCR to confirm that expression varied in association with *ZFP57* and was not a result of array artefacts. An eQTL analysis was carried out, the results of which are presented in the Manhattan plot in Figure 4.25:

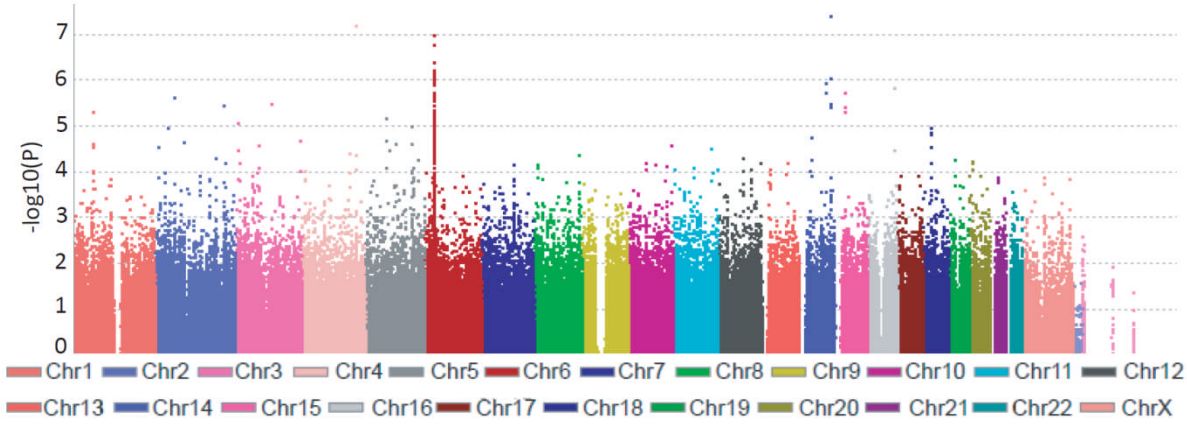


Figure 4.25: Manhattan plot showing genome-wide association of SNPs with *BTN3A2* expression. *BTN3A2* expression was determined in 285 healthy volunteers using qPCR and normalised to *ACTB* expression. Association was investigated for 651,210 SNPs using PLINK. A spike of SNPs showing strong association with *BTN3A2* expression is found around the *ZFP57* locus in the MHC.

Several of the most associated SNPs to *BTN3A2* expression are also shown to affect *ZFP57* expression (see Figure 4.25 and Table A.7). The SNP rs2517911, most associated with *BTN3A2* expression, is located upstream of *ZFP57* towards *HLA-F* and has an R^2 score of 0.692176 in linkage to rs2535238. This shows linkage of rs2517911 with the SNP most associated with *ZFP57* expression. It may be possible that ZFP57 as a TF is controlling expression of *BTN3A2*, thus their gene expression levels are directly correlated. When looking at *BTN3A2* expression in a haplotype specific manner it was clear that like *ZFP57* expression levels varied by haplotype at HLA-A*01 however no variation in expression was seen for haplotypes HLA-A*23 or *24 (see Figure 4.26).

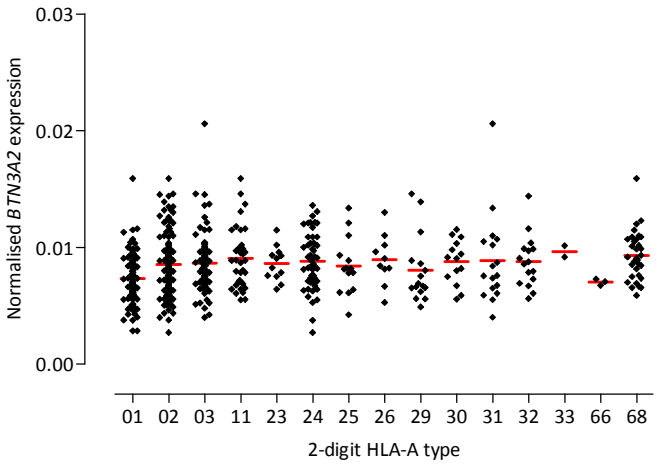


Figure 4.26: Expression of *BTN3A2* according to 2-digit haplotype at HLA-A. Expression of *BTN3A2* as determined using qPCR was plotted against HLA-A 2-digit haplotype determined by imputation for 283 healthy volunteers. Mean expression is indicated by the red bar. Lower expression was seen in the group possessing a copy of HLA-A*01 compared with all other haplotypes ($p<0.0001$). Lower expression seen in the group possessing a copy of HLA-A*66 was discounted due to the small sample number (n=3).

Lower expression of *BTN3A2* seen in volunteers possessing a copy of HLA-A*01 was confirmed by analysing the effect of copy number of HLA-A*01 in Figure 4.27:

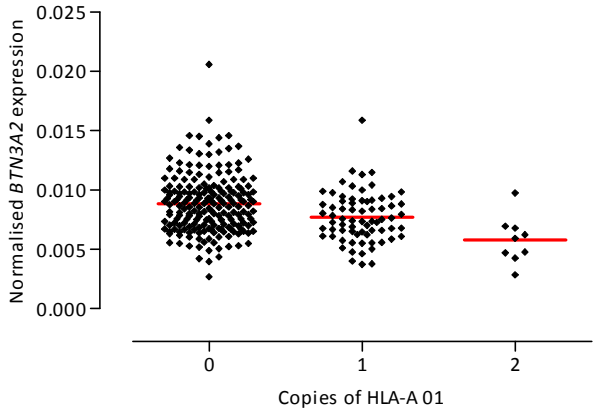


Figure 4.27: Expression of *BTN3A2* according to copy number of HLA-A*01. Expression of *BTN3A2* as determined using qPCR in the volunteer PBMC samples was plotted according to copy number of HLA-A*01. Mean expression in each group is indicated with the red bar. A statistically significant difference was seen between mean expression for all 3 groups determined using a Kruskal-Wallis test ($p<0.0001$).

This analysis confirms the impact of haplotype on gene expression for both *ZFP57* and *BTN3A2* and implies that the HLA-A*01 haplotype is most associated with increase in *ZFP57* gene expression and decrease in *BTN3A2* gene expression.

4.3.15 *ZFP57* and disease

Evidence has been presented in this chapter that there are highly significant genetic determinants of *ZFP57* expression and that this TF is expressed in a range of tissue types including monocytes. Previous reports have provided evidence of a role in TND and imprinting (Li 2008, Mackay 2008). Here we sought to investigate the identified expression-associated SNPs in the context of the rapidly growing GWAS datasets for common disease.

Using the catalogue of published GWAS data (www.genome.gov/gwastudies, Accessed September 2011), SNPs associated with *ZFP57* expression were searched in terms of SNPs with reported disease associations at genome-wide significance ($p<1x10^{-8}$). SNPs showing a cis eQTL association with *ZFP57* expression in PBMCs from the healthy volunteer cohort were compared to GWAS SNPs, and those SNPs in direct linkage with known disease associated SNPs that had been typed for the cohort (proxy SNPs). Proxy SNPs were defined using 1000 genomes CEU pilot data and had an R^2 of 0.8 or greater calculated using the SNAP website (http://www.broadinstitute.org/mpg/snap/) (Johnson 2008). Table 4.3 shows which disease associations had SNPs shared with *ZFP57* expression associated SNPs, and Table 4.4 details the mean expression of *ZFP57* associated with the different disease associated genotypes.

Disease Trait	SNP	*ZFP57* eQTL p value	Disease association p value	Reference
Nasopharyngeal carcinoma	rs3129055	**$7.15x10^{-31}$**	$7.00x10^{-11}$	Tse, 2009
	rs2517713	**$1.72x10^{-5}$**	$4.00x10^{-20}$	Tse, 2009
	rs2860580	**$1.25x10^{-4}$**	$5.00x10^{-7}$	Bei, 2010
AIDS progression	rs8321	**$2.75x10^{-11}$**	$5.00x10^{-7}$	Limou, 2009
HIV-1 control	rs259919	**$5.20x10^{-11}$**	$3.00x10^{-7}$	Fellay, 2009
	rs3131018	**$2.58x10^{-3}$**	$4.00x10^{-16}$	Pereyra, 2010
	rs7758512	**$5.57x10^{-3}$**	$2.00x10^{-8}$	Fellay, 2009
Lung adenocarcinoma	rs4324798	**$2.91x10^{-10}$**	$2.00x10^{-8}$	Landi, 2009
	rs3117582	**$1.99x10^{-4}$**	$5.00x10^{-12}$	Landi, 2009
Lung cancer	rs3117582	**$1.99x10^{-4}$**	$5.00x10^{-12}$	Broderick, 2009
	rs3117582	**$1.99x10^{-4}$**	$5.00x10^{-12}$	Wang, 2008
Nevirapine-induced rash	rs1265112	**$2.86x10^{-5}$**	$1.00x10^{-8}$	Chantarangsu, 2011
Systemic sclerosis	rs3130573	**$4.08x10^{-5}$**	$6.00x10^{-10}$	Allanore, 2011
Neonatal lupus	rs3099844	**$6.19x10^{-5}$**	$5.00x10^{-10}$	Clancy, 2010
Systemic lupus erythematosus	rs3131379	**$3.52x10^{-4}$**	$2.00x10^{-52}$	Harley, 2008
Vitiligo	rs6904029	**$6.86x10^{-5}$**	$1.00x10^{-21}$	Jin, 2010
Schizophrenia	rs911186	**$3.50x10^{-4}$**	$1.00x10^{-8}$	Purcell, 2009
	rs911186	**$3.50x10^{-4}$**	$1.00x10^{-8}$	Shi, 2009
	rs911186	**$3.50x10^{-4}$**	$1.00x10^{-8}$	Stefansson, 2009
	rs3800316	**$5.07x10^{-4}$**	$1.00x10^{-12}$	Stefansson, 2009
	rs7745603	**$6.37x10^{-4}$**	$1.00x10^{-8}$	Purcell, 2009
	rs7745603	**$6.37x10^{-4}$**	$1.00x10^{-8}$	Shi, 2009
	rs7745603	**$6.37x10^{-4}$**	$1.00x10^{-8}$	Stefansson, 2009
	rs6932590	**$9.82x10^{-4}$**	$1.00x10^{-12}$	Stefansson, 2009
	rs3131296	**$1.27x10^{-3}$**	$2.00x10^{-10}$	Stefansson, 2009
	rs13194053	**$5.35x10^{-3}$**	$1.00x10^{-8}$	Purcell, 2009
	rs13194053	**$5.35x10^{-3}$**	$1.00x10^{-8}$	Shi, 2009
Type 1 diabetes	rs2647044	**$4.24x10^{-3}$**	$1.00x10^{-16}$	Hakonarson, 2007

Table 4.3: *ZFP57* associated *cis* eQTL SNPs and their disease associations. Disease traits that have associated SNPs determined by the NHGRI GWAS data catalogue or their proxies from the genotyping data in the healthy volunteer cohort were interrogated for their association with *ZFP57* expression. Disease trait associated SNPs and the p value for association with *ZFP57* expression are shown, as well publications that reference the disease association.

Some SNPs that are strongly associated with *ZFP57* expression have been seen in GWAS for a variety of complex conditions. HIV-1 control and some forms of cancer seem to be the most likely disease associations to have a direct relationship with *ZFP57* expression however there is also some weaker evidence of association with various autoimmune conditions.

Disease trait	Top SNP	Genotype	AA	AB	BB	
Nasopharyngeal carcinoma	rs3129055	**G**/A	**0.07857**	**0.3357**	0.5857	Frequency
			0.00013	**$8.73x10^{-5}$**	$1.85x10^{-5}$	Mean Expression
			$5.44x10^{-5}$	**$6.14x10^{-5}$**	$4.04x10^{-5}$	SD
AIDS progression	rs8321	**C**/A	**0.02143**	**0.1321**	0.8464	
			0.000121	**0.000107**	$3.97x10^{-5}$	
			$3.99x10^{-5}$	**$6.46x10^{-5}$**	$5.75x10^{-5}$	
HIV-1 Control	rs259919	A/G	0.1143	0.4321	0.4536	
			$9.45x10^{-5}$	$6.42x10^{-5}$	$2.59x10^{-5}$	
			$5.74x10^{-5}$	$6.83x10^{-5}$	$4.79x10^{-5}$	
Lung adenocarcinoma	rs4324798	**A**/G	**0.01434**	**0.1254**	0.8602	
			0.000136	**0.000104**	$4.10x10^{-5}$	
			$4.18x10^{-5}$	**$6.70x10^{-5}$**	$5.79x10^{-5}$	
Nevirapine-induced rash	rs1265112	**C**/T	**0.1071**	**0.375**	0.5179	
			$8.56x10^{-5}$	**$5.83x10^{-5}$**	$3.73x10^{-5}$	
			$6.15x10^{-5}$	**$6.29x10^{-5}$**	$6.02x10^{-5}$	
Systemic Sclerosis	rs3130573	**G**/A	**0.08571**	**0.4429**	0.4714	
			$9.73x10^{-5}$	**$5.43x10^{-5}$**	$3.81x10^{-5}$	
			$5.92x10^{-5}$	**$6.08x10^{-5}$**	$6.18x10^{-5}$	
Neonatal Lupus Erythematosus	rs3099844	**A**/C	**0.01071**	**0.1821**	0.8071	
			0.000135	**$7.55x10^{-5}$**	$4.35x10^{-5}$	
			$4.40x10^{-5}$	**$5.71x10^{-5}$**	$6.25x10^{-5}$	
Systemic Lupus Erythematosus	rs3131379	**A**/G	**0.01083**	**0.1697**	0.8195	
			0.000161	**$6.95x10^{-5}$**	$4.48x10^{-5}$	
			$1.13x10^{-5}$	**$5.12x10^{-5}$**	$6.40x10^{-5}$	
Vitiligo	rs6904029	**A**/G	**0.1004**	**0.3763**	0.5233	
			$3.51x10^{-5}$	**$3.22x10^{-5}$**	$6.64x10^{-5}$	
			$6.73x10^{-5}$	**$5.15x10^{-5}$**	$6.61x10^{-5}$	
Schizophrenia	rs911186	G/**A**	0.05357	**0.3357**	**0.6107**	
			$8.00x10^{-5}$	**$6.45x10^{-5}$**	**$3.99x10^{-5}$**	
			$6.12x10^{-5}$	**$7.14x10^{-5}$**	**$5.61x10^{-5}$**	
Type 1 Diabetes	rs2647044	**A**/G	**0.0365**	**0.1752**	0.7883	
			0.000119	**$5.41x10^{-5}$**	$4.68x10^{-5}$	
			$5.14x10^{-5}$	**$4.56x10^{-5}$**	$6.59x10^{-5}$	

Table 4.4 Disease Trait associated SNPs and mean *ZFP57* expression according to genotype. For each trait, the top *ZFP57* associated SNP is shown, with the risk allele highlighted in bold type. Frequency for each genotype in the volunteer cohort is indicated, as well as the mean normalised *ZFP57* expression and standard deviation of this figure.

4.4 Discussion

ZFP57 differential expression has been analysed in this chapter, showing a strong genetic association in primary cells. Novel isoforms of the gene have been identified and validated. Protein expression of ZFP57 has been confirmed in transformed LCLs as well as primary tissue. Bioinformatic interrogation of the *ZFP57* loci in conjunction with eQTL mapping has uncovered potential regulatory sites. Expression of *ZFP57* has been linked to possession of particular HLA-A types, and this HLA type is likely to control expression of several other genes in the surrounding loci. SNPs associated with changes in the level of *ZFP57* gene expression have been compared with known disease associations to find potential pathogenic events where ZFP57 may be playing a role.

4.4.1 Specificity of the MHC array and confirmation of variable *ZFP57* expression

In this chapter evidence has been presented that *ZFP57* is differentially expressed among HLA homozygous LCLs by qPCR and that *ZFP57* expression was independent of *MOG* expression. This added confidence to the findings of the MHC array that differential gene expression detected was the result of the implicated gene rather than an artefact of nearby genes. The specificity of the MHC custom array was also confirmed as higher expression detected by the array towards the 3' end of *ZFP57* in the COX LCL was also confirmed using exon specific qPCR.

4.4.2 Novel isoforms of *ZFP57*

The finding of different expression levels between the different exons of *ZFP57* in the COX LCL led to interrogation of the different isoforms of *ZFP57* that could be detected using RACE-PCR. Two uncharacterised isoforms of *ZFP57* were found, both of which would cause the protein product of the transcript to be truncated. This was an interesting discovery as the truncated protein, if translated, would have lost most of the DNA binding KRAB domain that is thought to be the functional domain of KRAB-ZNF proteins. Although the protein was shown to be translated in the COX, APD and PGF LCLs (although at a low level in the PGF LCL) the translation of these alternate isoforms could not be confirmed as there was no commercially available antibody that was not raised against the N-terminal portion of the protein that would be truncated. QBL and PGF LCLs were interrogated for the

different isoforms however as expected there was no detectable expression of the "skipped" isoform.

The western blot showing ZFP57 expression in three LCLs shows that additional isoforms not detected by RACE-PCR may be present in the COX LCL, although these are not found in the nuclear lysate suggesting that only one functional isoform is trafficked to the nucleus, and the other may be degraded or is a precursor to the final functional isoform. The western blot analysis also shows that even very low expression of *ZFP57* such as in the PGF LCL can still lead to a protein product. This is relatively unexpected as the published literature suggests that *ZFP57* expression and translation is reduced as the cells differentiate (Okazaki 1994). The ZFP57 protein also appears to be in a much higher concentration in the APD LCL compared to the COX LCL, despite higher *ZFP57* expression seen in COX. This implies that gene expression levels may not solely dictate the protein levels of ZFP57 in a tissue. There is no additional band found in the APD LCL western blot analysis, perhaps indicating that additional isoforms are not present.

4.4.3 Fine mapping of *ZFP57* regulatory variants

The new 288 healthy volunteer cohort helps to fine-map genetic variation affecting *ZFP57* expression beyond that seen in Chapter 3. This is likely to be due to not only the increased numbers involved in the study but also the selection criteria of the volunteers to reduce the problems of population stratification within the samples. While the top SNP showing association with *ZFP57* expression in the initial cohort of 96 volunteers (rs29228) was not associated with expression in the YRI LCLs analysed, one of the top SNPs identified in the second cohort was shown to be significantly associated with change in mean expression with genotype in both the CEU and YRI populations (rs375904). Fine mapping of expression-associated variants in one population may therefore narrow the region of association to define or closely tag the functional variant. This is also evident when looking at recombination events over the *ZFP57* loci; where strong recombination is seen the association of SNPs to expression falls. Further fine mapping in a larger sample or targeted re-sequencing of the volunteer cohort gDNA may give more insight into the true functional variants

responsible for differential *ZFP57* expression however the area tagged by the most associated SNPs (over 5kb) is still relatively large.

4.4.4 Haplotype and HLA type association

The LD structure of the *ZFP57* region was defined using haploview (Barrett 2005), confirming that as expected all three top SNPs associated with *ZFP57* were found within the same haplotype block. This also allowed the identification of rs3129063 as the only SNP to differ between haplotypes with the "A" allele at rs2535238, indicative of increased *ZFP57* expression. This SNP difference did not lead to a significant difference in mean *ZFP57* expression between the two haplotypes (Mann-Witney test, p=0.41) showing that rs3129063 was not the functional SNP for this effect.

Using the imputation data for classical HLA alleles it was confirmed that there was a significant effect on *ZFP57* expression for some HLA-A types. The HLA-A types identified as associated with increased *ZFP57* expression were HLA-A *01, *23 and *24. Although possession of a copy of HLA-B*08 appeared to be associated with altered *ZFP57* expression it was shown that this was instead an association with the HLA-A type, in that generally possession of a copy of HLA-B*08 indicated possession of an HLA-A type associated with *ZFP57* expression, rather than a separate genetic effect.

Analysis of *ZFP57*-correlated genes in B cells and monocytes showed that *HLA-A* expression varied with *ZFP57* expression ($p=1.68x10^{-10}$ in monocytes, $p=1.01x10^{-8}$ in B cells) and several other genes around the *HLA-A* region were also found to correlate with *ZFP57* when analysed on the array, such as *HLA-G* and *HLA-H*. These genes are upstream of the *ZFP57* loci and are close to *HLA-A*; it is likely that a more general *cis*-acting effect is modulating the expression of genes within or near to the *HLA-A* locus. This further confirms the association of *ZFP57* expression seen with possession of HLA-A types *01, *23 and *24.

The HLA type HLA-A*01, HLA-B*08, HLA-C*07 have been frequently associated with each other and with autoimmune disease, as seen in the COX haplotype which has been shown to increase susceptibility to SLE, T1D and myasthenia gravis (Price 1999). Additionally, HLA-A*01, HLA-B*08 has been previously associated with early onset of SLE in two different populations (Goldberg 1976), while the extended haplotype of HLA-A01-B08-C07 was associated with rapid HIV progression in the

mid-nineties and has also been associated with susceptibility to common immunodeficiency and infectious disease more generally (Kaslow 1990, Price 1999).

It was reassuring to see that expression of *ZFP57* was higher in individuals who possessed a copy of the extended COX haplotype (HLA-A1-B8-Cw7-DR3) compared with those that had a copy of the PGF haplotype (HLA-A3-B7-Cw7-DR15). This fitted with the findings of the MHC array where the COX LCL had higher *ZFP57* expression than either PGF or QBL. This finding is likely to be due to the COX haplotype containing the HLA-A*01 allele that has been shown to be associated with higher expression of *ZFP57*, and shows that MHC homozygous LCLs can be used to identify strong haplotypic effects that are also found in healthy volunteer primary cells.

BTN3A2 expression was found to have associated SNPs shared with *ZFP57* expression-associated SNPs, making it likely the two genes were somehow related. Analysis of *BTN3A2* showed that possession of the HLA-A*01 haplotype is associated with gene expression, and so the association with *ZFP57* expression may instead be the result of long-range haplotype interactions of the HLA-A locus. However, the association between the two genes appears to be negative: as *ZFP57* expression increases, *BTN3A2* expression is seen to decrease slightly and the genotype associated with higher expression of *ZFP57* is associated with lower expression of *BTN3A2*.

BTN3A2 has been previously shown to have variable gene expression with regulatory elements found over at least a 15kb stretch of DNA in *cis* (Pastinen 2004). It is a member of the immunoglobulin super-family (Williams and Barclay 1988) and has therefore been implicated in regulation of the immune response. Intriguingly a SNP shown to regulate *BTN3A2* expression has been associated with T1D (Viken 2009) in an independent finding to those of the association of the HLA-DQA1, -DQB1, -DRB1 and HLA-A, -B, -C loci that have been well established with T1D (Sheehy 1989, Hanifi Moghaddam 1998, Valdes 2005, Nejentsev 2007). Mutations that affect the ZFP57 protein have been previously associated with TND, which increases predisposition to T1D in later life as well as increasing the chance of having a relative with diabetes (Mackay and Temple 2010, Temple and Shield 2010).

4.4.5 ENCODE data to interrogate *ZFP57* regulation

This chapter has shown that *ZFP57* expression is not confined to development in all humans. Expression of both the gene and protein has been seen in several different cell and tissue types, and while expression is low it is highly variable, dependent on genotype. This is particularly apparent in the healthy volunteer cohort, where the majority of individuals do not express *ZFP57* at a detectable level; however a significant minority show reproducible expression that leads to a protein product (determined by FACS analysis). This expression is not constant in all cell types, confirmed using FACS analysis that showed significantly higher ZFP57 expression in monocytes compared to B cells from one individual. Regulatory marks investigated by ENCODE also showed cell-type specific differences in the histone modifications observed, with H3K27Ac marks (associated with active regulatory elements) strongly observed over the 5' end of *ZFP57*, as well as indications of promoter-specific histone marks (H3K4Me1) and histone modifications generally associated with regulatory elements (H3K4Me3).

Of the seven cell lines analysed for histone modifications by the ENCODE project, NHEK (a keratinocyte cell line) showed strongest evidence of the regulatory marks associated with active gene transcription and regulation. This is interesting as the keratinocyte cells have undergone substantial differentiation and thus would be expected to have very low or absent gene expression. However, gene expression data was not available for the cell line from ENCODE so whether these modifications are functional is unclear. Embryonic stem cells (hESC cell line) showed H3K4Me1 enrichment, suggesting the presence of regulatory elements. Unlike in NHEK cells, this was stronger than the H3K27Ac mark, which is associated with active regulatory elements. Additionally, expression of *ZFP57* was found using RNA-seq by ENCODE, but it was at a relatively low level. This could be due to the conditions for cell culture of hESC cells being different to conditions in vivo or because down-regulation of *ZFP57* has already begun. Early down-regulation of *ZFP57* has been previously noted when ZFP57 was determined essential in both maternal mRNA and in the early stage of the oocyte (Li 2008). Interestingly, in hESC cells, the highest enrichment of the H3K27Ac mark is seen overlapping the area showing the highest association with *ZFP57* expression levels. When viewing the ENCODE

data the most informative cell line for comparison purposes to the LCL and volunteer data was the binding data generated for GM12878, an LCL of CEU origin. GM12878 showed evidence of the H3K4Me1 mark suggesting regulatory elements were present, particularly over the first intron where most SNPs associated with *ZFP57* expression were found.

The cell line GM12878 is heterozygous for the top SNPs associated with *ZFP57* expression differences (genotype available through the HapMap project). Where a given TF identified as having a binding site modified by one of these SNPs has been assayed by ENCODE for genome-wide binding it has failed to see any of the *in silico* predictions made by JASPAR *in vivo*. This is likely to be due to the high number of false positives generated by TF binding prediction (Fickett 1996). It is also known that most gene regulation interactions involve several regulatory proteins working in a complex , and these would not be identified by DNA sequence alone (Wasserman and Sandelin 2004). ENCODE TF binding data detected SRF binding in H1hESC, CEPBP in HeLa cells, STAT3 in MCF10A-Er-Src cells, CTCF in MCR-7 cells and in H1hESC (using 2 different antibodies). All these TFs had also been investigated using ChIP-seq in the GM12878 LCL; however no binding of any was detectable.

It is unsurprising to only see the binding of SRF to *ZFP57* in hESC cells, as SRF is reported to be predominantly an activating TF, inducible in response to serum or growth factors and contributing to cell differentiation (Johansen and Prywes 1995). It is therefore likely that in embryonic stem cells, SRF plays a regulatory role in the expression of *ZFP57*. However, it is unknown if SRF could also be regulating expression of *ZFP57* in LCLs and PBMCs, especially as no binding was detected in the GM12878 LCL. CTCF is well known as a transcriptional regulator (Phillips and Corces 2009); its binding in the hESC line is unsurprising as it may play a role in *ZFP57* regulation in early development. Like SRF it is not found binding in the GM12878 LCL so it is unclear if any putative regulatory role could be extrapolated to cells that have undergone differentiation.

4.4.6 Disease association with expression of *ZFP57*

Some SNPs shown to be associated with variance in *ZFP57* expression were found to be associated with complex disease (or a proxy SNP in LD with a previously associated SNP). The strongest association with *ZFP57* expression involving a disease associated SNP was the nasopharyngeal

carcinoma associated SNP rs3129055; disease association p=7.0x10^{-11} (Tse 2009), *ZFP57* expression association p=7.15x10^{-31}, which is located upstream of *ZFP57* and falls under a peak of H3K4Me1 histone modification (determined by ENCODE ChIP-seq data) in hESC cells predominantly. This is particularly interesting due to the previous study of *ZFP57* expression in early development. rs3129055 was originally thought to be associated with the *HLA-F* gene (upstream of the SNP) when analysed by the nasopharyngeal carcinoma study (Tse 2009), but this analysis suggests that it may instead be marking an association with *ZFP57* expression downstream of the SNP. It is found in high LD with the top *ZFP57* associated SNP rs2535238 (R^2=0.838496), which is suggestive of strong linkage. Two other SNPs showing weaker association with *ZFP57* expression have also been identified associated with increased susceptibility to nasopharyngeal carcinoma; rs2517713; disease association p=4x10^{-20}, (Tse 2009) *ZFP57* expression association p=1.72x10^{-5} and rs2860580; disease association p=5x10^{-7} (Bei 2010) *ZFP57* expression association p=1.25x10^{-4}. Both of these SNPs are located at a greater distance to *ZFP57*, near *HLA-A*, and so could be the result of linkage in the MHC. As previously discussed, *HLA-A* expression has been correlated with *ZFP57* in primary B cells and monocytes, but this was thought likely to be due to a longer ranging *cis* effect than through direct linkage.

In addition to the association of *ZFP57* expression with nasopharyngeal carcinoma, an association with lung cancer and adenocarcinoma was also found. The lung adenocarcinoma susceptibility SNP showing the strongest association with *ZFP57* expression is rs4324798 (p=2.9x10^{-10}), which is found nearly 1Mb downstream from *ZFP57* and has a disease association of p=2x10^{-8} (Landi 2009). However, it is located under a strong peak for H3K4Me3 (histone modification associated with promoter elements) determined by ENCODE ChIP-seq data. Another SNP increasing susceptibility to lung cancer/ adenocarcinoma (rs3117582) replicated in three studies with association to disease of p=5x10^{12} (Wang 2008a, Broderick 2009, Landi 2009) shows much weaker association with *ZFP57* expression (p=2.0x10^{-4}) and is located upstream at much greater distance. Despite this, it may still be indicative of a real effect as the SNP is found under a robust peak of H3K4Me3, and H3K27Ac (associated with active regulatory elements).

KRAB-ZNF genes have previously been associated with various types of cancer, both by down-regulation of the gene leading to a lack of heterochromatin formation (Cheng 2010), and up-regulation leading to silencing of other tumour suppressor genes (Silva 2006). Both these examples show the fine balance that must be maintained for normal cell development, and the epigenetic ramifications for the disruption that can occur when expression of these TFs is altered. Methylation or deletion of ZNF genes on chromosome 19, where a cluster of these TFs are found is commonly seen in many cancers (Shao 2001).

HIV-1 control and AIDS progression has been associated with the HLA region by many studies (Pereyra 2010). rs8321, found within the *ZNRD1* gene, has shown the highest association with *ZFP57* expression ($p=2.75x10^{-11}$) and has been implicated in control of AIDS progression ($p=5x10^{-7}$) (Limou 2009). The SNP rs259919 with an association to HIV control of $p=3x10^{-7}$ (Fellay 2009) has also shown high association with *ZFP57* expression ($p=5.2x10^{-11}$) and is also located further upstream in the MHC class I region near the *ZNRD1* gene. Several other SNPs associated with HIV-1 control also show evidence of association with *ZFP57* expression but at much lower significance; rs3131018 ($p=2.58x10^{-3}$) and rs7758512 ($p=5.57x10^{-3}$) (Fellay 2009, Pereyra 2010). rs7758512 is again located within the *ZNRD1* gene, under a strong peak of H3K4Me1 methylation modification in the NHEK cell line only. This could imply a cell-type specific effect of the SNP, perhaps indicating that in certain cell types it could have a more significant effect on *ZFP57* expression. It is unclear how *ZFP57* would carry out a functional role in the control of HIV-1, especially in the context of the proposed mechanism of variable virus-peptide binding by HLA molecules (Pereyra 2010). The association of HIV-1 control to *ZFP57* expression may be the result instead of linkage in the MHC or by association of *ZFP57* with *ZNRD1*. When *ZNRD1* is knocked down HIV-1 progression is inhibited (Ballana 2010), so an association between *ZFP57* and *ZNRD1* could imply that *ZNRD1* is a target of *ZFP57*.

The overlap of autoimmune disease association with *ZFP57* expression SNPs was not particularly striking compared to cancer and HIV-1 control; however some disease-associated SNPs are also found to be associated with *ZFP57* expression. A GWAS for vitiligo, an autoimmune disease resulting in depigmentation through melanocyte loss, pinpointed the SNP rs6904029 (p value of disease

association =$1x10^{-21}$) (Jin 2010), which is weakly associated to *ZFP57* expression ($p=6.86x10^{-5}$). At a similar level of significance of *ZFP57* expression association are associations to systemic sclerosis, SLE and neonatal lupus erythematosus which are all inflammatory conditions (Harley 2008, Clancy 2010, Allanore 2011). T1D association of SNP rs2647044 (disease association $p=1x10^{-16}$) is seen at a lower level of significance (*ZFP57* expression association $p=4.2x10^{-3}$) but is interesting due to previous association of *ZFP57* to TND (Mackay 2008). The SNP rs2647044, identified as being associated with T1D susceptibility (Hakonarson 2007), is located in the HLA-DQ region, previously associated with several autoimmune diseases, particularly T1D (Barrett 2009a, Britten 2009).

Some SNPs showing a weak association with *ZFP57* expression have been previously associated with schizophrenia in several different studies (Purcell 2009, Shi 2009, Stefansson 2009). These tend to be found in close proximity to other ZNF genes rather than in the *ZFP57* region, suggesting that perhaps an element of global control of ZNF genes could be responsible for this association, rather than a specific effect of *ZFP57* expression in schizophrenia.

4.4.7 Conclusion

ZFP57 has proven to be a gene of interest in terms of its differential expression and regulation. SNPs that have been shown to be closely associated with *ZFP57* expression have been previously identified as associated with various diseases, and these diseases have often been previously linked the HLA-A type most associated with *ZFP57* expression. HLA-type analysis has allowed these associations to be made clear. It is possible to see this as a future way for functional studies to combine their information with GWAS leading to greater biological insight. This analysis will be used as a template for the study of other promising candidates from the previous chapter, the *HLA-DQ* genes and *HLA-C*.

Chapter 5 - Allele-specific expression of classical HLA genes

5.1 Introduction

Classical HLA genes play a critical role in antigen presentation and immunity. They are remarkably polymorphic and show striking associations with disease, notably autoimmune conditions (Fernando 2008b). However, fine mapping and resolving the functional basis of many associations is difficult. In many cases this involves structural changes in the encoded proteins but differential gene expression may also be significant (Todd 1988, Epstein 2009). In light of the observed associations with gene expression found for specific HLA genes and SNPs in Chapter 3, some genes in the classical class I and class II regions, namely *HLA-C* and the *HLA-DQ* family are investigated further.

5.1.1 HLA-DQ region

The HLA-DQ genes encode a class II antigen presenting molecule, the HLA-DQ heterodimer (see Figure 5.2), involved in antigen presentation via the exogenous pathway. The HLA-DQ heterodimer is formed from α-chains encoded by *HLA-DQA1*, and a β-chains encoded by *HLA-DQB1* which associate and are anchored in the cell surface membrane. While *HLA-DQA2* is also an α-chain molecule, it is expressed at lower levels than *HLA-DQA1*. It is unclear whether the HLA-DQB2 protein is expressed despite expression of the gene (also a β-chain encoding gene) at the transcript level, meaning that HLA-DQA2 may be expressed on the cell surface of antigen presenting cells without associating with a β-chain molecule (Rudy and Lew 1997). The genes are found in a highly polymorphic region of the MHC, particularly the membrane distal domain genes *HLA-DQA1* and *HLA-DQB1*; the HLA-DQ locus is detailed in Figure 5.1. The HLA-DQ genes are expressed primarily in antigen presenting cells such as B cells, dendritic cells and macrophages (Korman 1985).

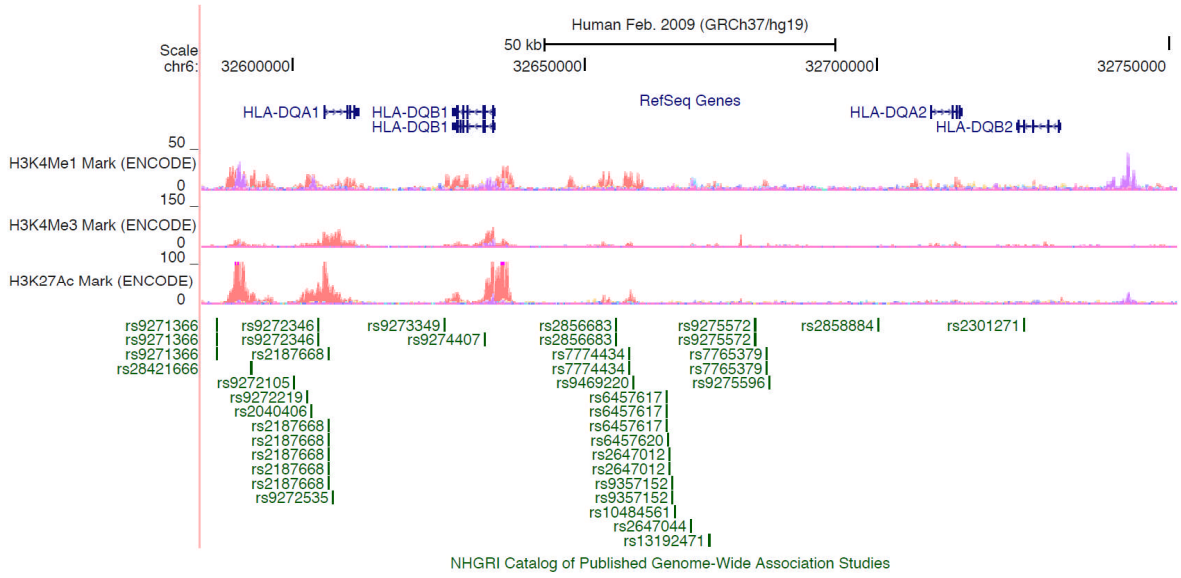


Figure 5.1: HLA-DQ gene locus. The HLA-DQ genes are found in an area of extreme polymorphism, and many variants at this locus have been associated with complex disease. SNPs shown in green are GWAS identified variants associated with various complex diseases. The HLA-DQ genes from the refseq assembly are shown, with histone modifications associated with regulation of gene expression; H3K4Me1 (often indicating regulatory elements), H3K4Me3 (often indicating a promoter region), and H3K27Ac (often indicating active regulatory elements) in seven cell lines investigated by ENCODE.

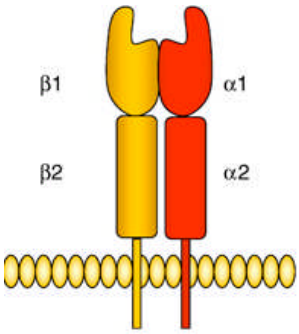


Figure 5.2: HLA-DQ protein association. Reproduced from (Cassinotti 2009). The schematic shows the association between the two different protein products of the class II genes, such as those encoded by HLA-DQ region genes, forming an antigen presenting molecule on the surface of an antigen presenting cell.

The polymorphisms found in all of these genes include non-synonymous variants and structural changes in the expressed protein that result in varying abilities to bind and present certain antigens. This is likely to be most significant for the more widely expressed *HLA-DQA1* and *HLA-DQB1* genes. *HLA-DQB1* was shown to have the most significant *cis*-associated SNPs related to its variable

expression in Chapter 3, and is the subject of further characterisation in this chapter. The gene can be found on the antisense strand in the MHC (chr6:32,627,657-32,634,466) (hg19). The HLA-DQ heterodimer of HLA-DQA1 and HLA-DQB1 has been implicated in coeliac disease, where possession of particular HLA-DQA and HLA-DQB alleles led to increased risk of disease, and intriguingly possession of two copies of the relevant HLA-DQB*02 allele appeared to be associated with enhanced disease risk at an earlier age, suggesting a dose-dependent effect of this risk allele (Murray 2007). This is particularly interesting when considering the impact of *cis*-regulated variable expression. It is thought that binding of gluten to the HLA class II molecules at the HLA-DQ heterodimer is partly responsible for the pathogenesis of coeliac disease, whereby specific changes to the residues of the dimer alter the affinity of the antigenic peptides that are presented by the molecule, resulting in some form of aberrant presentation (Qiao 2009). Coeliac disease is thought to be influenced by many factors and an individual alteration in the state of a HLA molecule is generally accepted to be insufficient to induce the disease (Green and Jabri 2006).

The HLA-DQ region has also been associated with other autoimmune diseases, notably T1D. Both increased and decreased susceptibility to T1D have been identified involving the HLA-DQA and HLA-DQB regions (Todd 1990, Redondo 2001). Three other inflammatory conditions have also shown reproducible associations with the HLA-DQ region; RA, SLE and MS (Yao 1993, Ihle 2003, Heap and van Heel 2009). This is perhaps unsurprising given the significant overlap seen between loci associated with many complex autoimmune disorders (Cotsapas 2011).

Inherited factors also have a large impact on a response to infection or susceptibility to infectious disease (Hill 2006). The HLA-DQ region as a whole has been previously implicated in several studies investigating susceptibility to infectious disease, for example the modification of outcomes when infected by *Toxoplasma gondii* (Mack 1999). A well reported association of the HLA-DQ region with susceptibility to leprosy is hard to distinguish from the HLA-DR region due to the strong LD in the region (Fitness 2002, da Silva 2009).

In light of these observations, the expression of all genes in the HLA-DQ region will be interrogated in this chapter for further understanding into their regulation and disease association.

5.1.2 HLA-C

In contrast to the HLA-DQ genes, *HLA-C* is expressed in the majority of cells rather than just antigen presenting cells. It is part of the MHC class I antigen presentation pathway, important in presenting peptides to CD8+ cytotoxic T cells via the endogenous pathway, and is encoded on chr6:31,236,530-31,239,855 (hg19). HLA-C protein is associated with β2-microglobin, however only HLA-C forms the peptide binding cleft, and is anchored into the cell surface by a transmembrane domain (see Figure 5.3). If a non-self antigen is presented by HLA-C or one of the other HLA class I molecules it identifies the cell to the immune system as being infected.

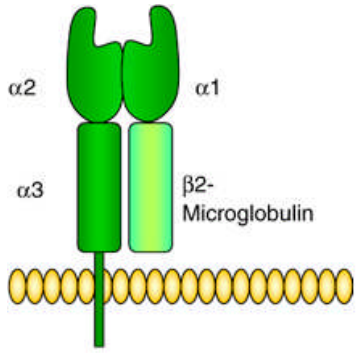


Figure 5.3: Class I antigen presenting molecule. Reproduced from (Cassinotti 2009). The schematic shows the association between the 2 different protein products of the class I genes, such as *HLA-C,* and *β2-Microglobulin*; forming an antigen presenting molecule found on the cell surface of the majority of cell types.

HLA-C is expressed at a lower level on the cell surface than the other classical class I molecules HLA-A and HLA-B, despite similar transcription levels. This is thought to be due to poor association with β2-microglobin and export to the cell surface (Neefjes and Ploegh 1988). HLA-C-specific responses and disease associations have been less described than other class I associations (Blais 2011). Like the HLA-DQ region, *HLA-C* has been associated with several diseases that involve the immune system, notably T1D (Valdes 2005), MS (Yeo 2007), and host control of HIV-1 (Fellay 2007). SNPs in the *HLA-C* region, which is also close to the genes *HCP5* and *HLA-B* (see Figure 5.8), have been implicated in the control of the viral load in HIV infected patients, however strong LD in the region makes it impossible to determine the true causal variant for this phenomenon (Trachtenberg 2009). This is also the case in the study of psoriasis, where a SNP upstream of the *HLA-C* region has been implicated as

associated with the condition, but there are many other SNPs associated in the same region and a mechanism for their effect is unclear (Liu 2008). Expression of *HLA-C* was shown to be associated with several local, likely *cis-acting* variants following eQTL mapping in the cohort of 96 healthy volunteers (Chapter 3). In this chapter, *HLA-C* and its genetic determinants of expression are analysed further.

5.2 Aims

Both the *HLA-DQ* region and the *HLA-C* region have been implicated in susceptibility to, and protection against, many human diseases. However, the mechanisms for many of these associations remain unclear. While a dose-dependent effect caused by the number of copies of the susceptibility allele may be causative, it is often hard to determine if this is due to an effect on the protein structure or the expression level of the gene of interest. By further investigating the expression of these genes it is hoped that it will become clearer whether expression levels could be responsible for disease pathogenesis or protection. The HLA-DQ region was investigated to determine if expression of all four genes is related and to identify alleles affecting expression strongly, which could in turn be linked to disease. The *HLA-C* gene has been previously linked to HIV-1 control and protein expression appears particularly important for this. The investigation into *HLA-C* gene expression could help to identify variants especially associated with gene expression that may in turn affect protein levels, as well as to fine map known HIV-1 associated SNPs in terms of their gene association. Specific aims of this chapter are:

1. To validate and fine map the observed cis-eQTL for *HLA-DQB1* and other HLA-DQ genes
2. To validate and fine map the observed cis-eQTL for *HLA-C*
3. To investigate the functional and disease significance of identified expression-associated SNPs and haplotypes

5.3 Results

5.3.1 HLA-DQ region

Initially, expression of *HLA-DQA1*, *HLA-DQA2*, *HLA-DQB1* and *HLA-DQB2* was investigated in the new volunteer cohort (detailed in sections 2.7 and 4.3.3) to determine firstly, if differential expression of these functionally linked genes was related to each other and secondly, if the association of rs1071630 with HLA-*DQB1* gene expression found in the original cohort (Chapter 3) could be replicated. Gene expression was measured using SYBR green qPCR in duplicate and was normalised to *ACTB*. Genotyping of gDNA was performed using the Illumina Infinium high-density genotyping bead arrays and association analysis was performed using PLINK as previously described (Sections 3.3 and 4.3.3). The same qPCR primers (sequences found in Table A.1) were used to ensure that the results were comparable between the sample cohorts.

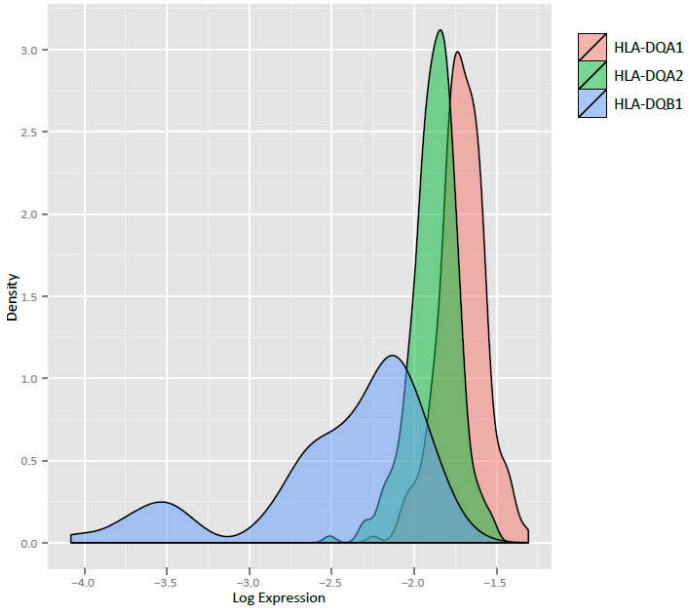


Figure 5.4: Density plot showing variation of expression levels in *HLA-DQA1*, *HLA-DQA2* and *HLA-DQB1*. Log expression of the three *HLA-DQ* genes, determined by qPCR, plotted versus the frequency density of expression.

The observed expression levels for each gene and variance within the cohort are shown (Figure 5.4). *HLA-DQB2* expression was found to be so low that it was not included in this analysis (see Figure 5.5). While both *HLA-DQA1* and *HLA-DQA2* show a relatively Gaussian distribution of expression, with little overall variation between individuals, *HLA-DQB1* expression is highly variable and seems to have two

separate groups in terms of expression levels. These are defined as "high" and "low" HLA-DQB1 expressers. High expressers do not have a Gaussian distribution of expression, implying that a third group may be present, however for the purpose of this analysis they will be considered together. Expression for each HLA-DQ gene was then correlated with the other HLA-DQ region genes to look for association of expression in the region as a whole (Figure 5.5).

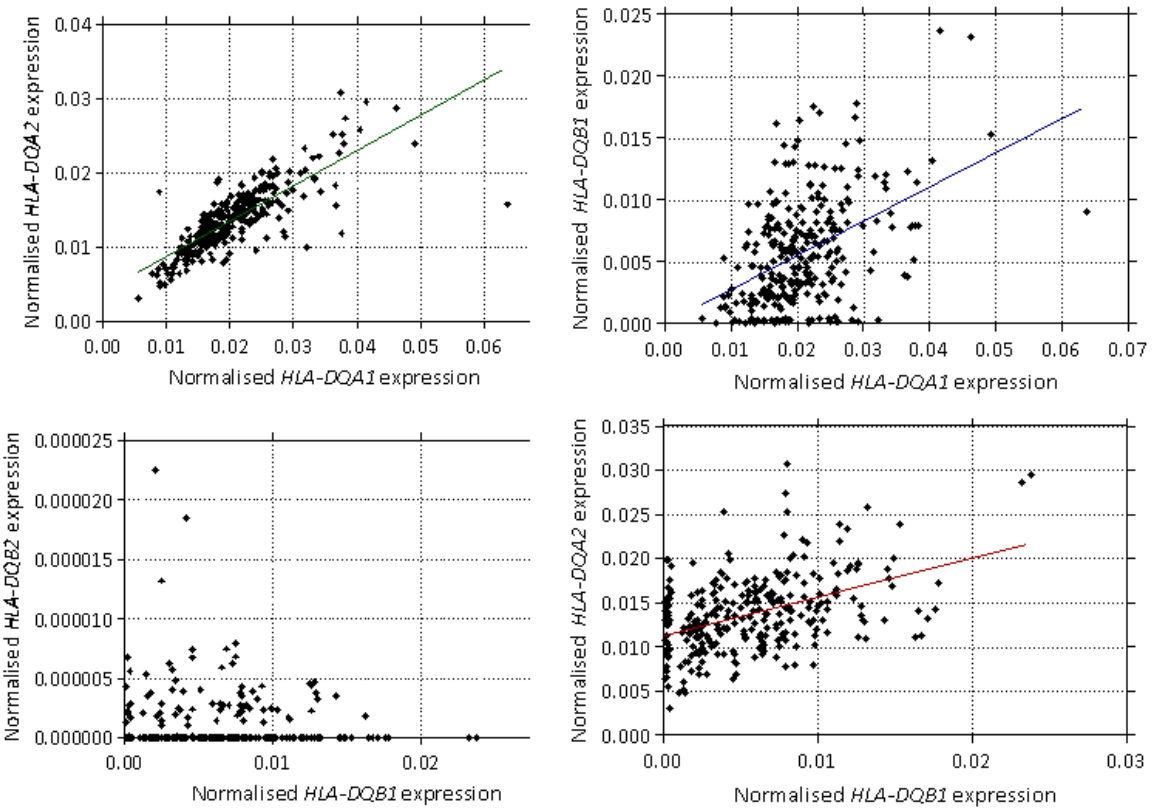


Figure 5.5: Linear regression correlation of expression of i) *HLA-DQA1* vs *HLA-DQA2* ii) *HLA-DQA1* vs *HLA-DQB1* ii) *HLA-DQB1* vs *HLA-DQB2* and iv) *HLA-DQA2* vs *HLA-DQB1*. Normalised expression of the 4 HLA-DQ genes are plotted against each other. R^2 values are i) 0.632 ($p<0.0001$, linear regression), ii) 0.210 ($p<0.0001$), iii) 0.0008 ($p=0.638$) and iv) 0.199 ($p<0.0001$).

This data shows that three out of the *HLA-DQ* genes show a significant correlation in levels of expression. The strongest correlation was seen between *HLA-DQA1* and *HLA-DQA2*. No correlation was seen between *HLA-DQB2* and *HLA-DQB1* (and between *HLA-DQB2* and any *HLA-DQA* gene, data not shown), probably due to the extremely low expression of *HLA-DQB2* rendering many samples undetectable and thus affecting the results. Genetic determinants of gene expression were then

investigated for *HLA-DQA1, HLA-DQA2* and *HLA-DQB1*, but not *HLA-DQB2* due to the high number of samples deemed "non-expressing".

5.3.2 *HLA-DQB1* expression

PLINK analysis showed that the SNP previously most associated to *HLA-DQB1* expression in the first volunteer cohort (rs1071630) was also strongly associated with expression in the second cohort ($p=2.45x10^{-54}$) (see Figure 5.6). The denser genotyping information in the second cohort allowed a stronger association to *HLA-DQB1* expression to be determined, rs9273448 ($p=3.18x10^{-66}$). Imputation over the region surrounding *HLA-DQB1* using the data from the 1000 genome project (Pennisi 2010) led to 39,975 additional SNPs for denser genotyping information to use in linear analysis performed by PLINK. The program IMPUTE2 was used to predict the genotypes at additional loci using the co-ordinates Chr6: 29600054-35639688 (Hg19), taken from the UCSC genome browser.

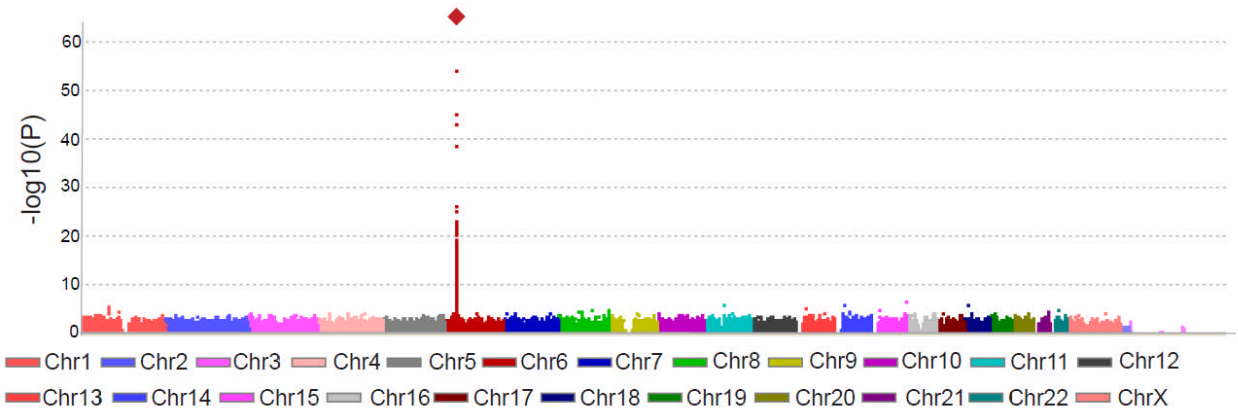


Figure 5.6: Manhattan Plot showing the association of SNPs with *HLA-DQB1* expression. The most associated SNP to HLA-DQB1 expression, rs9273448, is indicated by the larger red diamond.

In total, 3964 SNPs were shown to be associated at a genome-wide level of significance ($1x10^{-8}$) with the expression of *HLA-DQB1* in the volunteer cohort. Expression was plotted in Figure 5.7, grouped by genotype at rs9273448 and rs1071630 to demonstrate the effect of genotype on expression levels.

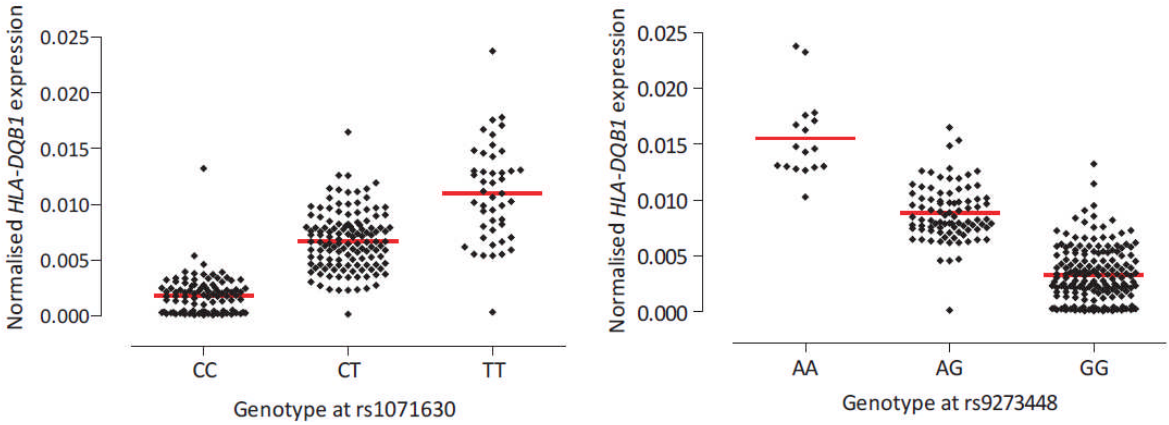


Figure 5.7: Association of genotype at i) rs1071630 and ii) rs9273448 with *HLA-DQB1* expression. Expression values are separated by genotype at i) rs1071630, the most associated SNP with *HLA-DQB1* expression in the initial cohort of 96 volunteers and ii) rs9273448, the most associated SNP with *HLA-DQB1* expression in the new cohort of 288 volunteers and plotted. Mean expression is indicated for each genotype by the red bar. Both SNPs are shown to indicate a significant difference in mean expression between the different genotypes by Kruskal-Wallis test (rs1071630; $p<0.0001$, rs9273448; $p<0.0001$).

The region isolated by the Manhattan plot as showing most association with *HLA-DQB1* expression was investigated further using a recombination plot of the locus and surrounding area (Figure 5.8). Recombination rates plotted in the locus showed that SNP association to expression decreased after peaks of recombination as expected. As the locus was so polymorphic, only imputed SNPs shown to be in high LD with the most associated typed SNPs to *HLA-DQB1* expression were included (22 in total).

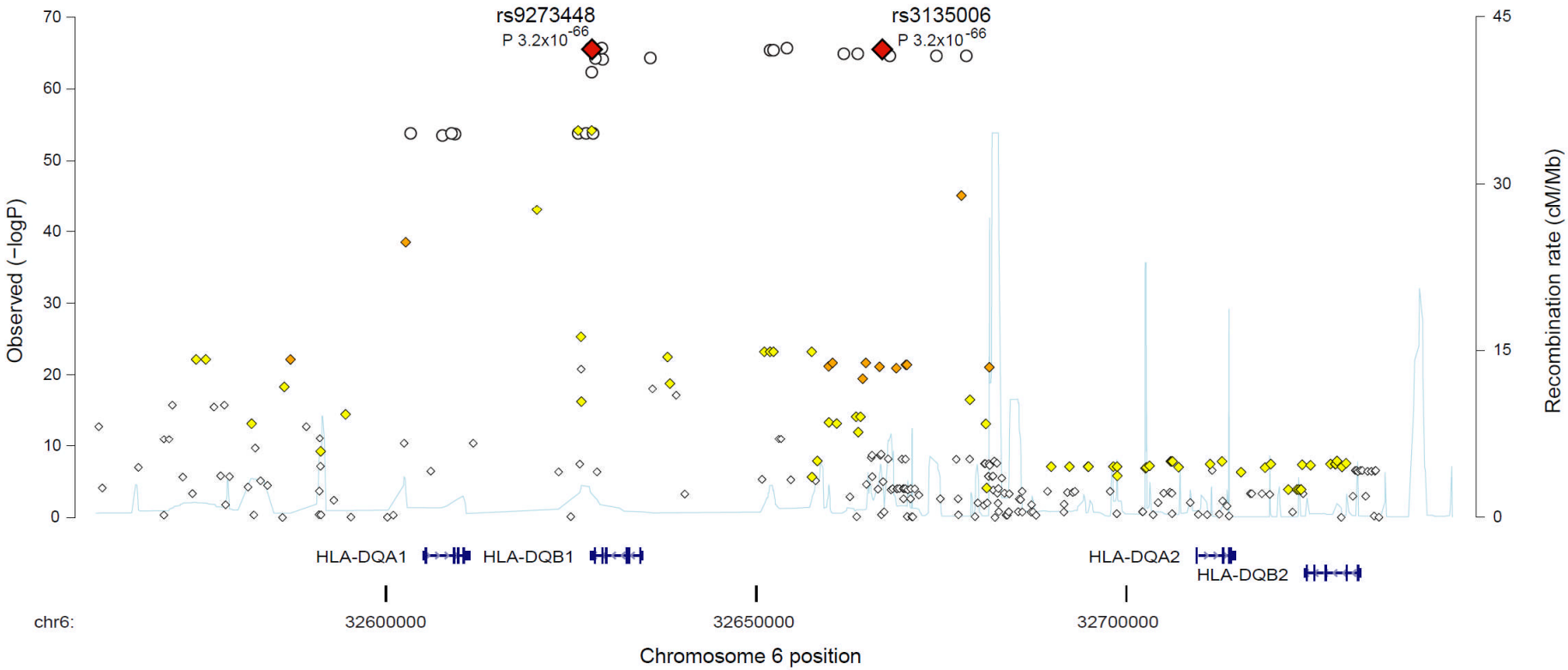


Figure 5.8: Regional association plot of the *HLA-DQB1* locus: rs9273448 and rs3135006 the most associated typed SNPs to *HLA-DQB1* expression are shown as red diamonds. Other typed SNPs in the region are indicated with a diamond showing LD to rs9273448: orange diamond R2 between 0.5≤0.8, yellow diamond 0.2≤0.5, white diamond <0.2) Imputed SNPs with R2>0.8 and so presumed to be in close LD are shown as white circles.

As *HLA-DQB1* expression appeared to have two separate distributions (see Figure 5.4), these were analysed separately to assess if there were separate genetic contributions to each group. The normalised expression of *HLA-DQB1* was used to define the groups, with samples that had log(*HLA-DQB1* expression)<-3.0 being defined as "low" expressers as previously stated. PLINK was again used to find SNPs that were associated with the difference in expression patterns. In the group with higher *HLA-DQB1* expression the top SNPs associated with expression did not change significantly. As the low expresser group had only 39 individuals, it was impossible to carry out an association analysis using expression values as there was not sufficient power to detect SNPs associated with this expression pattern. Instead a case-control association analysis was run, using the high expressers as "unaffected" samples, and the low expressers as "affected" samples. The top SNP associated with the low expresser group was rs9273363 ($p=4.48x10^{-22}$). The top SNPs associated with each expression group were compared in Table 5.1:

rs9273363	**GENO**	**A/A**	**A/C**	**C/C**
	COUNTS	20	115	145
	FREQ	0.07143	0.4107	0.5179
	MEAN	0.000294	0.004128	0.007776
rs9273448	**GENO**	**A/A**	**A/G**	**G/G**
	COUNTS	17	87	176
	FREQ	0.06071	0.3107	0.6286
	MEAN	0.01553	0.00882	0.003278

Table 5.1: SNP frequencies and mean expression of *HLA-DQB1*. rs9273363 (associated with the low expresser group of *HLA-DQB1*) and rs9273448 (associated with the high expresser group of *HLA-DQB1*) are compared in terms of their frequencies and associated expression in the volunteer cohort.

5.3.3 *HLA-DQA1* expression

As *HLA-DQB1* expression was associated so strongly with genetic variation, and *HLA-DQA1* expression was correlated with *HLA-DQB1* (see Figure 5.5), it was unsurprising that *HLA-DQA1* was shown to also have strong genetic determinants of expression when analysed using PLINK. Results from this analysis are shown in the Manhattan plot in Figure 5.9.

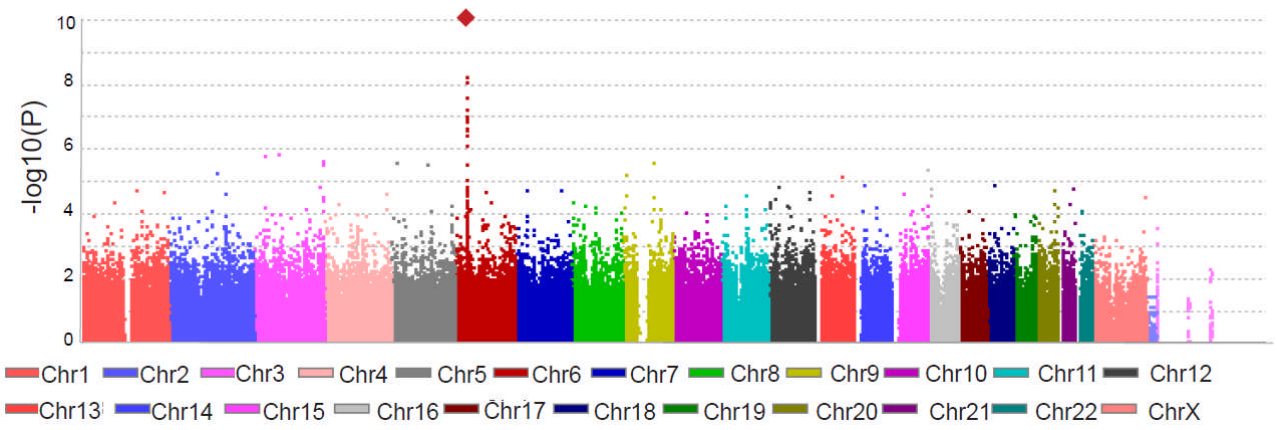


Figure 5.9: Manhattan Plot showing the association of SNPs with *HLA-DQA1* expression. The most associated SNP with *HLA-DQA1* expression, rs17843604, is indicated with the larger red diamond ($p=7.83x10^{-11}$).

The most associated SNP with *HLA-DQA1* (rs17843604) expression was used to group the volunteers by genotype to look at how expression varied in Figure 5.10. In order to see if SNPs that affected *HLA-DQB1* expression also affected other genes, *HLA-DQA1* expression was plotted by genotype at rs9273448 (the most associated SNP with *HLA-DQB1* expression).

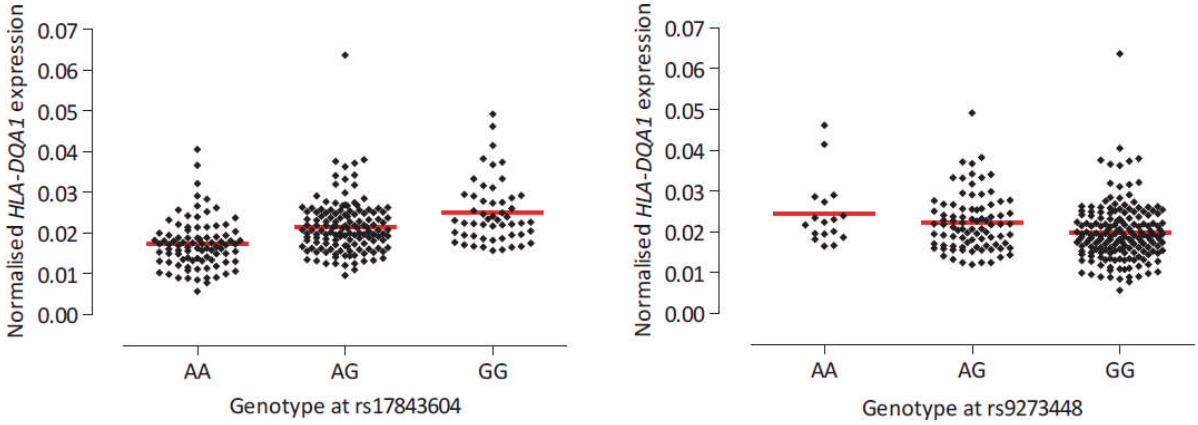


Figure 5.10: Association of i) genotype at rs17843604 and ii) genotype at rs9273448 with *HLA-DQA1* expression. Expression values are separated by genotype at i) rs17843604, the most associated SNP with *HLA-DQA1* expression and ii) rs9273448, the most associated SNP with *HLA-DQB1* expression. (p value of association with *HLA-DQA1* expression=$5.84x10^{-4}$). Mean expression is indicated for each genotype by the red bar. Both SNPs are shown to indicate a significant difference in mean expression of *HLA-DQA1* between the different genotypes by Kruskal-Wallis test (rs17843604 $p<0.0001$, rs9273448 $p=0.021$).

5.3.4 *HLA-DQA2* expression

The same SNP associated most with *HLA-DQA1* expression was also found to be most strongly associated with *HLA-DQA2* expression (rs17843604) following eQTL analysis (Figure 5.11). This reinforces the finding that *HLA-DQA1* and *HLA-DQA2* were the most correlated genes of the HLA-DQ region in terms of their expression.

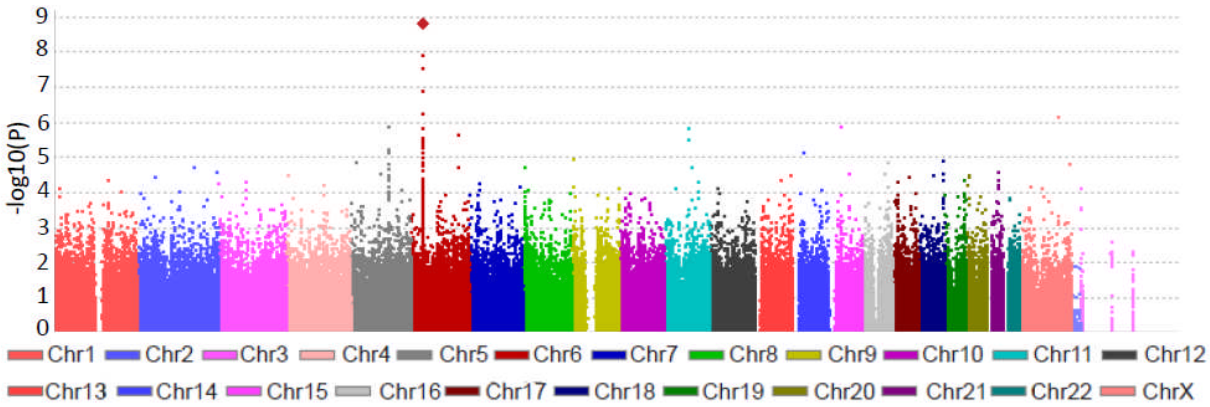


Figure 5.11: Manhattan Plot showing the association of SNPs with *HLA-DQA2* expression. The most associated SNP with *HLA-DQA2* expression, rs17843604, is indicated with the larger red diamond.

HLA-DQA2 expression according to genotype at rs17843604 and rs9273448 (to assess how the SNP strongly associated with *HLA-DQB1* expression affected other HLA-DQ genes) were plotted in Figure 5.12.

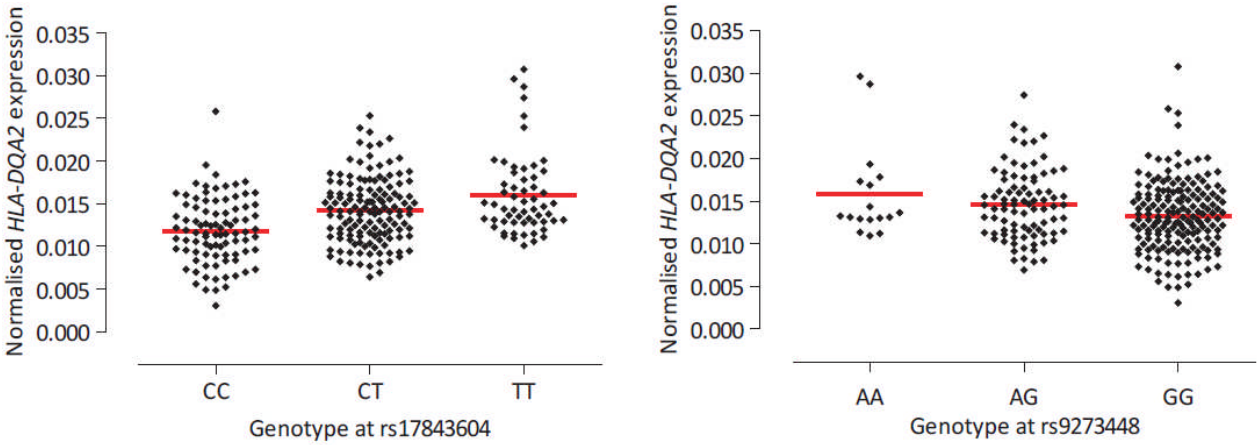


Figure 5.12: Association of i) genotype at rs17843604 and ii) genotype at rs9273448 with *HLA-DQA2* expression. Expression values are separated by genotype at i) rs17843604, the most associated SNP with *HLA-DQA2* expression ($p=1.58x10^{-9}$) and ii) rs9273448, the most associated SNP with *HLA-DQB1* expression (p(association to *HLA-DQA2*)=$1.27x10^{-3}$) and plotted. Mean expression is indicated for each genotype by the red bar. Both SNPs are shown to indicate a significant difference in mean expression of *HLA-DQA2* between the different genotypes by Kruskal-Wallis test (rs17843604 p<0.0001, rs9273448 p=0.0298).

5.3.5 Haplotype association with *HLA-DQB1* expression

As *HLA-DQB1* showed the strongest association of genetic variants with gene expression of the *HLA-DQ* genes, it was investigated further. The top SNPs associated with both the increased expression of *HLA-DQB1* and the group of *HLA-DQB1* low expressers were analysed for evidence of a haplotype association.

Both SNPs were found to be relatively rare, with a homozygous minor allele frequency of 7% (rs9273363) and 6% (rs9273448) respectively. The expression of *HLA-DQB1* in volunteers with these minor alleles in a homozygous state was significantly different to expression in volunteers that had one or two copies of the major allele. This implied that they could be from separate genetic associations or haplotypes associated with expression variation of *HLA-DQB1*. However, for both high and low expression of *HLA-DQB1*, top SNPs associated with expression were outside of any LD blocks defined by Haploview (see Figure 5.13). This suggests that the haplotype marked by this variant is not well defined and may be close to an area of high recombination.

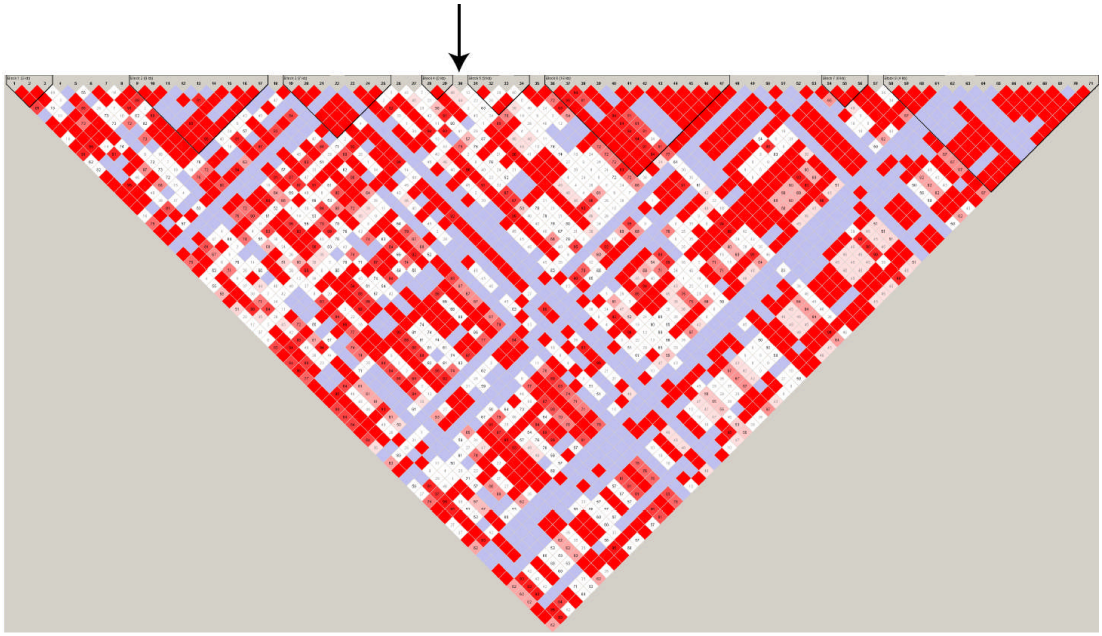


Figure 5.13: Haplotype structure over *HLA-DQB1* expression associated variants. A haploview plot shows the LD structure of the region around the SNPs most associated with *HLA-DQB1* expression. Red diamonds show allelic association (measured by the D' statistic) between two SNPs with an LOD>2; the darker the shade of red, the higher the value of D', where 1 is the highest value. White squares indicate no statistically significant evidence of LD but have an LOD>2. Blue squares show no LD and have an LOD<2 between SNPs. The arrow indicates the approximate area of the most associated SNPs. Using SNPs from the HapMap project, no strong LD can be found in the area implying haplotypes are impossible to define over this region.

As haplotype blocks beyond the relatively large area defined in Figure 5.8 cannot be determined for the region containing the most associated SNPs, further haplotype-based analysis of *HLA-DQB1* expression is not possible. The classical HLA types however can be interrogated to see if they impact the expression of *HLA-DQB1* significantly and mark a longer ranging haplotypic effect. As HLA-DQB is one of the six main HLA-loci that have been typed by imputation for the volunteer cohort (see section 4.3.6), it was investigated whether HLA-DQB type was associated with *HLA-DQB1* gene expression. Association of all classical HLA types with *HLA-DQB1* expression was also analysed to help define the genetic variation responsible for differential gene expression (see Figure 5.14), particularly HLA-DQA and HLA-DR alleles which are both close to the HLA-DQB locus.

5.3.6 HLA type association with *HLA-DQB1* expression

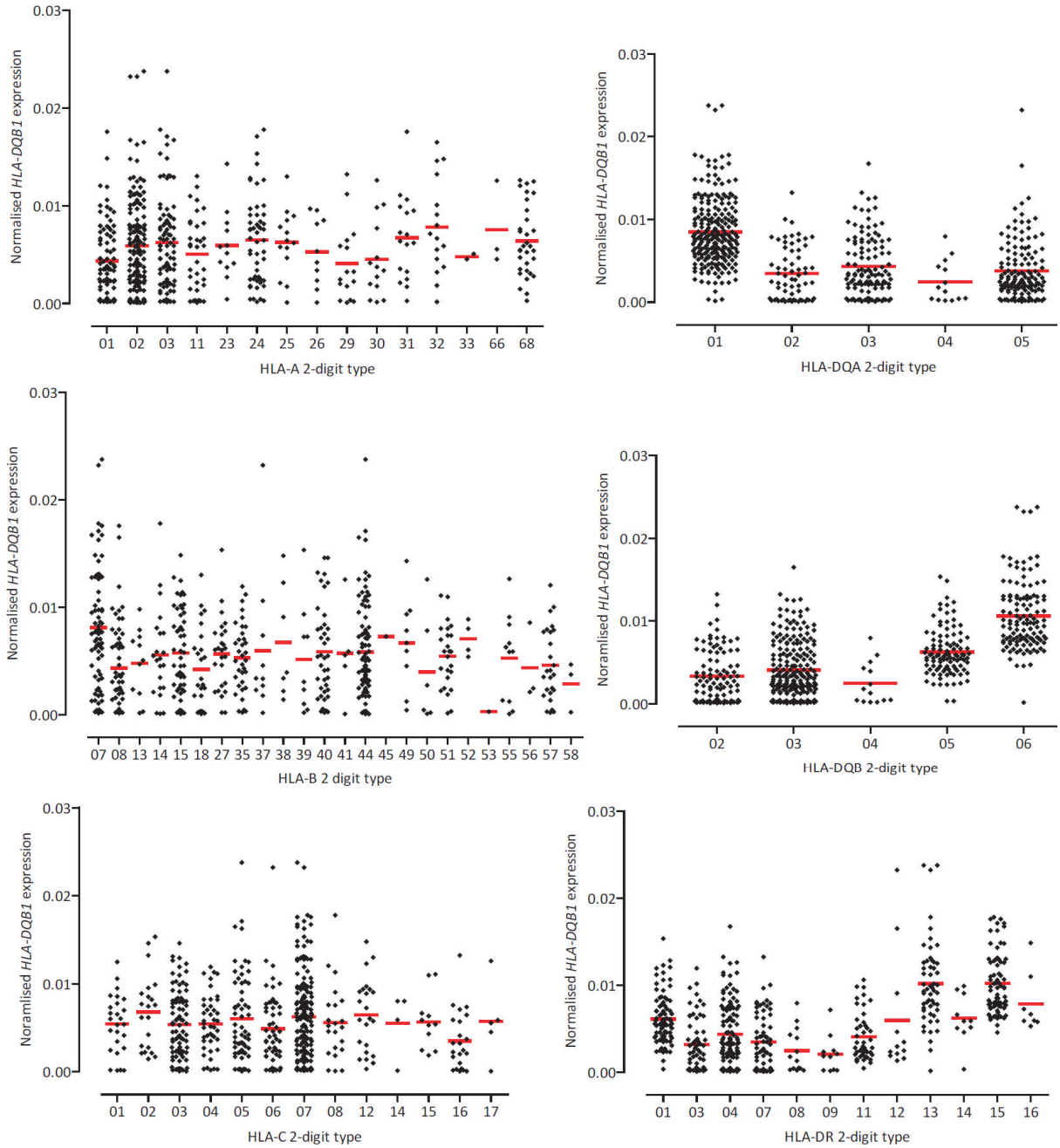


Figure 5.14: *HLA-DQB1* expression and 2 digit HLA-types. The average *HLA-DQB1* expression associated with a copy of each different HLA type is shown for six regions across the MHC for 288 healthy volunteers together with the mean expression values (denoted by the red bar). While HLA-A, HLA-B and HLA-C alleles do not show any association with *HLA-DQB1* expression differences, class II alleles do show evidence of association ($p<0.0001$ when analysed using a Mann-Whitney test) involving HLA-DQA*01, HLA-DQB*06 and HLA-DR*13/15.

HLA-DQB alleles could be segregated according to mean expression values of *HLA-DQB1*, with possession of a copy of HLA-DQB*06 in particular indicating likelihood of higher expression. As

increased expression was also seen in volunteers who possessed a copy of HLA-DQA*01 or HLA-DR*13/14/16, further analysis was undertaken to determine if the HLA-DQB allele was responsible for these differences as well (see Figures 5.15 and 5.16), or if there was a separate genetic effect.

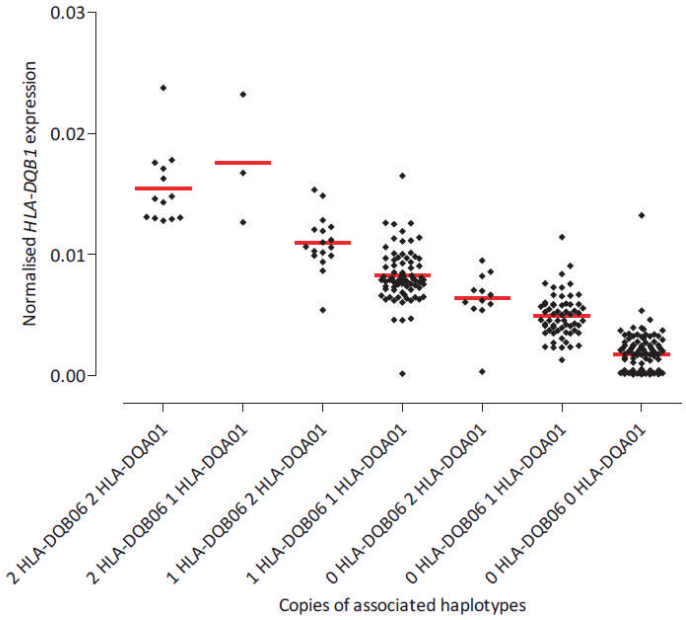


Figure 5.15: *HLA-DQB1* expression and HLA-DQB and HLA-DQA allele copy number and gene expression. Volunteers are grouped according to copy number of HLA-DQB*06 and HLA-DQA*01 and *HLA-DQB1* expression levels are plotted. Possession of copies of either allele influences the mean gene expression level, indicated by the red bar ($p<0.0001$ when analysed using a Kruskal-Wallis test).

This analysis shows that despite no clear haplotypic association being found over the region at the SNP level with *HLA-DQB1* expression, there are strong effects apparent on analysis of HLA type involving the classical class II alleles. Although HLA-DR type appears to influence expression, this is most likely the result of association between the HLA-DQB and HLA-DR regions, as HLA-DR type does not influence gene expression directly (Figure 5.16). There was also no clear HLA-type association with the group of low *HLA-DQB1* expressers, suggesting that the SNPs that are associated with low *HLA-DQB1* expression mark a particular functional variant.

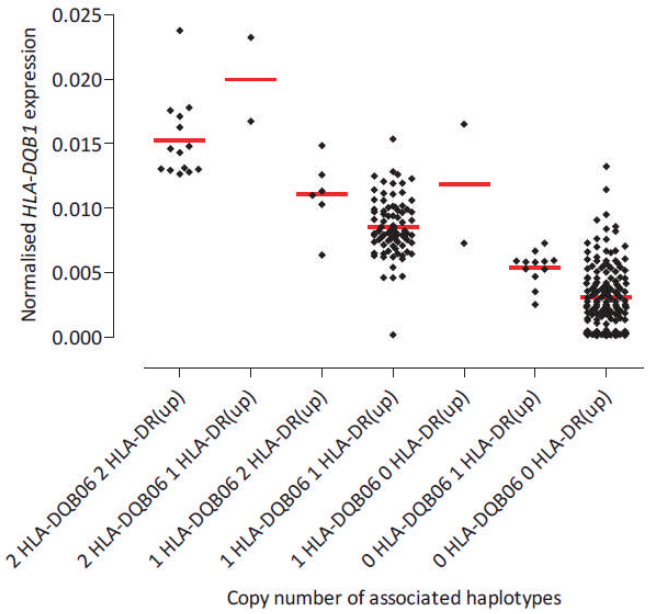


Figure 5.16: *HLA-DQB1* expression-associated-allele (HLA-DQB and HLA-DR) copy-number and gene expression. Volunteers were grouped according to copy number of HLA-DQB*06 and HLA-DR*13/15/16 alleles and their *HLA-DQB1* expression is plotted. Mean expression is indicated for each group by the red bar. While possession of HLA-DQB*06 clearly influences gene expression ($p<0.0001$ between the HLA-DQB types using a Kruskal-Wallis test), the number of copies of the HLA-DR alleles do not necessarily have a separate effect ($p=0.0785$ between the 1 HLA-DQB*06 with varying HLA-DR types using a Kruskal-Wallis test, $p=0.003$ between 0 HLA-DQB*06 with varying HLA-DR types using a Mann-Whitney test).

5.3.7 *HLA-DQ* gene expression and association with disease

Having shown that *HLA-DQB1* expression is highly variable according to genotype and can also be associated with possession of particular HLA types, its association with disease was investigated. As previously noted, the haplotype HLA-A3-B7-Cw7-DR15 carried by the PGF LCL is associated with several autoimmune diseases, and includes the HLA-DQB*06 allele, shown here to be associated with increased *HLA-DQB1* expression. By comparing the SNPs most associated with *HLA-DQB1* expression with the GWAS catalogue of disease associated SNPs, possible functional information about *HLA-DQB1* and its role in disease could be determined. This was carried out in the same manner as previously described in section 4.3.15 and the results detailed in Table 5.2. If diseases associated with the HLA-A3-B7-Cw7-DR15 haplotype, such as T1D protection, and MS or SLE susceptibility, are shown to be particularly associated with *HLA-DQB1* expression this may also give an insight into how this haplotype is involved in disease susceptibility.

Disease trait	*HLA-DQB1* associated SNP	p value association with *HLA-DQB1* expression	p value GWAS disease association	Reference
Type 1 diabetes	rs1063355	$9.79x10^{-55}$	$6x10^{-129}$	(WTCCC 2007, Cooper 2008)
	rs2647044	$7.69x10^{-9}$	$1x10^{-16}$	(Hakonarson 2007)
Multiple sclerosis	rs3135388	$5.17x10^{-23}$	$4x10^{-225}$	(Hafler 2007, De Jager 2009)
	rs9271366	$6.48x10^{-23}$	$4x10^{-17}$	(Bahlo M 2009, Nischwitz 2010)
	rs3129934	$1.67x10^{-18}$	$9x10^{-11}$	(Comabella 2008)
Asthma	rs9273349	$9.79x10^{-55}$	$7x10^{-14}$	(Moffatt 2010)
	rs204993	$2.81x10^{-8}$	$2x10^{-15}$	(Hirota 2011)
	rs3117098	$3.68x10^{-7}$	$5x10^{-12}$	(Hirota 2011)
Immunoglobulin A Deficiency	rs9271366	$6.48x10^{-23}$	$3x10^{-33}$	(Ferreira 2010)
	rs2187668	$3.8x10^{-7}$	$2x10^{-33}$	(Ferreira 2010)
Follicular lymphoma	rs2647012	$4.21x10^{-20}$	$2x10^{-21}$	(Smedby 2011)
Hepatocellular Carcinoma	rs9275572	$3.55x10^{-17}$	$6x10^{-9}$	(Kumar 2011)
Hodgkin's lymphoma	rs6903608	$3.05x10^{-10}$	$3x10^{-50}$	(Enciso-Mora 2010)
Alopecia areata	rs9275572	$3.55x10^{-17}$	$1x10^{-35}$	(Petukhova 2010)
Systemic lupus erythematosus	rs9270856	$1.94x10^{-16}$	$1x10^{-12}$	(Han 2009)
	rs2301271	$5.15x10^{-8}$	$2x10^{-12}$	(Chung 2011)
	rs2187668	$3.8x10^{-7}$	$6x10^{-28}$	(Chung 2011), (Hom 2008)
Inflammatory bowel disease	rs9271366	$6.48x10^{-23}$	$2x10^{-70}$	(Okada 2011)
	rs9271488	$1.84x10^{-13}$	$1x10^{-8}$	(Kugathasan 2008)
	rs477515	$1.21x10^{-11}$	$1x10^{-8}$	(Kugathasan 2008)
Systemic sclerosis	rs6457617	$9.48x10^{-13}$	$2x10^{-37}$	(Radstake 2010, Allanore 2011)
	rs443198	$1.37x10^{-8}$	$9x10^{-21}$	(Gorlova 2011)
	rs3129763	$7.98x10^{-8}$	$1x10^{-11}$	(Gorlova 2011)
Nephropathy	rs9275596	$1.04x10^{-21}$	$2x10^{-26}$	(Gharavi 2011)
Rheumatoid arthritis	rs6457617	$9.48x10^{-13}$	$1x10^{-9}$	(Julia 2008), (WTCCC 2007)
	rs9268853	$3.15x10^{-12}$	$5x10^{-109}$	(Eleftherohorinou 2009)
	rs9272219	$5.11x10^{-11}$	$1x10^{-45}$	(Eleftherohorinou 2009)
Ulcerative colitis	rs2395185	$2.22x10^{-12}$	$5x10^{-22}$	(Asano 2009, Silverberg 2009)
	rs9268853	$3.15x10^{-12}$	$1x10^{-55}$	(Anderson 2011)
	rs9268877	$1.06x10^{-9}$	$4x10^{-23}$	(Barrett 2009b), (Franke 2008)
Coeliac disease	rs2187668	$3.8x10^{-7}$	$1x10^{-50}$	(Dubois 2010)
	rs2187668	$3.8x10^{-7}$	$1x10^{-19}$	(van Heel 2007)
Leprosy	rs602875	$9.82x10^{-10}$	$5x10^{-27}$	(Zhang 2009)
Hepatitis B	rs2856718	$7.96x10^{-9}$	$4x10^{-37}$	(Mbarek 2011)
	rs7453920	$2.73x10^{-8}$	$6x10^{-28}$	(Mbarek 2011)

Table 5.2: Disease traits and their associated SNPs identified by GWAS studies that have association to *HLA-DQB1* expression. The table shows selected disease-trait associated SNPs have been subsequently linked to *HLA-DQB1* expression.

There are several diseases that have been strongly linked to particular SNPs that either show association with *HLA-DQB1* expression or are in strong linkage with SNPs associated with *HLA-DQB1* expression. The list of associated diseases is extensive, so only those SNPs showing association with *HLA-DQB1* expression above a genome-wide confidence limit have been included. In general these SNPs have been linked to mainly autoimmune diseases although there is also an association with infectious disease in the case of hepatitis and leprosy. Diseases that have previously been associated with the HLA-A3-B7-Cw7-DR15 (PGF) haplotype are also shown to have SNP associations that are linked with *HLA-DQB1* expression.

5.3.8 *HLA-DQB1* and Leprosy

The top SNP associated with *HLA-DQB1* expression in the first cohort of volunteers analysed was rs1076320 and this was associated with expression at an even more significant level in the second cohort of volunteers ($p=2.45x10^{-54}$). This SNP had been previously implicated in susceptibility to leprosy in a GWAS (Wong 2010). Additionally, a second GWAS showed that another Leprosy-associated SNP was also associated with *HLA-DQB1* expression (rs602875, see Table 5.3) (Zhang 2009). This study also showed that the SNP rs9271366 was associated with Leprosy susceptibility. However, rs9271366 was found to be refractory to genotyping in their cohort, and so rs602875 was used as a proxy SNP. When examined in the healthy volunteer cohort, rs9271366 was found to be associated with *HLA-DQB1* to a much greater extent, ($p=7.24x10^{-23}$). In the healthy volunteer cohort, the SNP rs602875 was not genotyped with high quality; 57 individuals were unable to be genotyped at this location. This could explain why rs602875 appears to be so much less associated with *HLA-DQB1* expression despite the high LD between it and rs9271366.

The SNP rs602875 found by the Zhang study and rs1071630 determined by the Wong study to be associated with leprosy susceptibility were also shown to be associated with *HLA-DQA1* expression; rs1071630 ($p=3.66x10^{-9}$) and rs602875 ($p=1.51x10^{-3}$) although the significance of the association of rs602875 does not reach genome-wide significance. It is notable that although some association is seen the p values are much less significant than those seen for association with *HLA-DQB1* expression and so it is hypothesised that association with *HLA-DQB1* expression is driving the signal seen in the

GWAS study. The SNP rs9271366 has no association with *HLA-DQA1* expression (p=0.77), making this theory even more likely.

In order to determine if *HLA-DQB1* expression could be directly linked with leprosy susceptibility, SNPs identified by both previous GWAS studies as identified with leprosy were compared for their association to *HLA-DQB1* and *HLA-DQA1* gene expression as well as the strength of association to leprosy susceptibility and whether the effect of the minor allele was protective or deleterious in Figure 5.17 and Table 5.3.

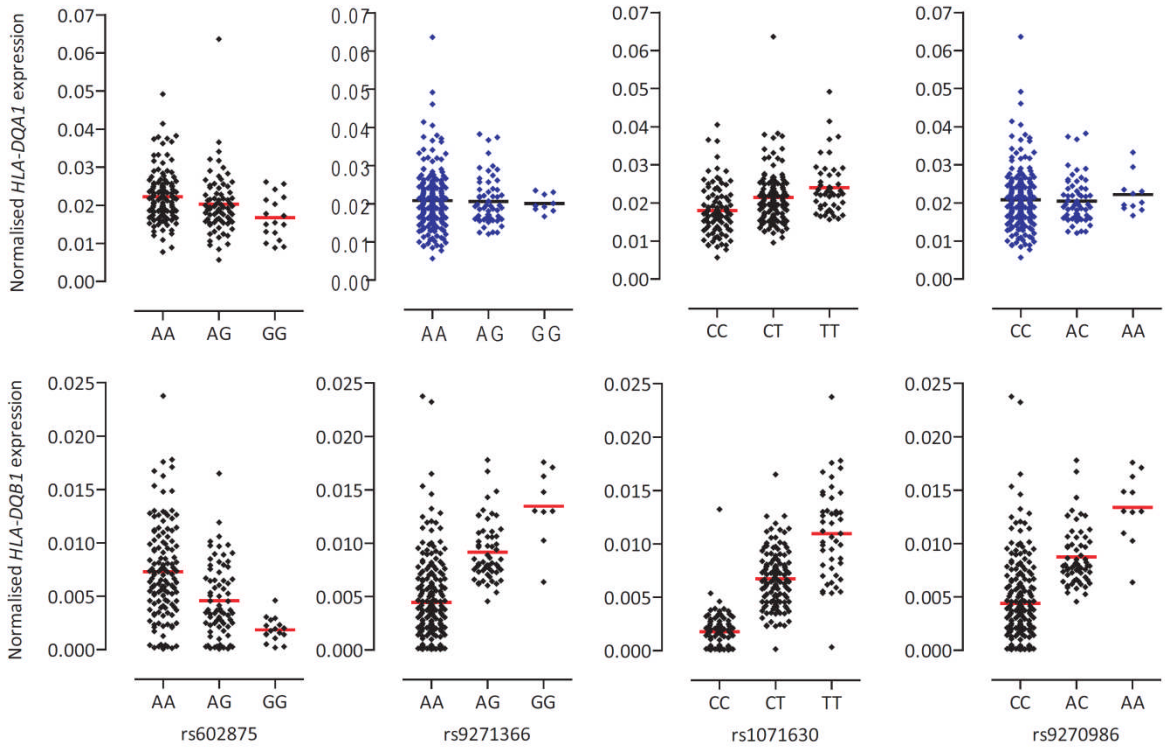


Figure 5.17: Expression of *HLA-DQA1* and *HLA-DQB1* for leprosy-associated SNP genotypes. Expression of *HLA-DQA1* and *HLA-DQB1* in the healthy volunteer cohort is determined by qPCR and is shown according to genotype at four SNPs most associated with leprosy susceptibility in two separate studies; rs602875, rs9271366, rs1071630 and rs9270986. Expression that is found to be significantly different according to genotype ($p<0.05$ by Kruskal-Wallis test) is shown in black (in all examples of differential *HLA-DQB1* expression $p<0.0001$, while differential *HLA-DQA1* p values are: rs602875 p=0.0014 and rs1071630 $p<0.0001$), while expression that shows no significant difference associated with genotype is coloured in blue.

SNP	Genotype	Position	Putative Gene of regulation	p value *HLA-DQB1* expression	Associated fold change in *HLA-DQB1* expression	p value *HLA-DQA1* expression	Associated fold change in *HLA-DQA1* expression	OR of Leprosy association	p value of Leprosy association	Reference
rs3104369	C/T*	32602482	*HLA-DQB1*	$3.31x10^{-39}$	2.58	$8.67x10^{-8}$	1.24	2.01 (1.53-2.64)	$2.20x10^{-7}$	(Wong 2010)
rs477515	A*/G	32569691	*HLA-DQB1*	$1.21x10^{-11}$	0.47	0.0001593	0.84	0.41 (0.29-0.57)	$4.80x10^{-7}$	(Wong 2010)
rs9270856	A*/G	32570839	*HLA-DQB1*	$1.94x10^{-16}$	1.96	0.05309	1.06	Not recorded	0.0955	(Zhang 2009)
rs9270986	A*/C	32574060	*HLA-DQB1*	$6.79x10^{-23}$	2.14	0.5319	1.02	2.45 (1.83-3.29)	$1.40x10^{-9}$	(Wong 2010)
rs1071630	C/T*	32609126	*HLA-DQB1*	$2.45x10^{-54}$	4.39	$3.66x10^{-9}$	1.26	2.3 (1.85-2.86)	$4.9x10^{-14}$	(Wong 2010)
rs9271366	A/G*	32586854	*HLA-DQB1*	$7.24x10^{-23}$	2.18	0.8142	0.98	2.35	$1.94x10^{-17}$	(Zhang 2009)
rs602875	A/G*	32573629	*HLA-DQB1*	$9.82x10^{-10}$	0.45	0.001513	0.83	0.67	$5.0x10^{-27}$	(Zhang 2009)
rs9273448	A*/G	32627747	*HLA-DQB1*	$3.18x10^{-66}$	3.02	0.0005848	1.18	NA	NA	NA

Table 5.3: SNPs previously associated with Leprosy and their association to *HLA-DQA1* and *HLA-DQB1* expression. SNPs shown to be associated with susceptibility to leprosy in 2 separate GWAS analyses are shown with both their association to *HLA-DQA1* and *HLA-DQB1* expression and the fold change in expression seen in *HLA-DQB1* with regard to the minor allele (indicated with *). The p value of their association with leprosy susceptibility, odds ratio for minor allele frequency in disease populations and study reference are also indicated.

All of the leprosy-associated SNPs were shown to have strong association above genome-wide significance with *HLA-DQB1* expression and for all SNPs the minor allele shown to be enriched in the leprosy cohorts studied is found to be associated with an increase in *HLA-DQB1* expression. Where the minor allele is found to be depleted in disease cohorts (rs602875 and rs477515) and thus associated with protection, it is associated with a decrease in expression of *HLA-DQB1*. All SNPs analysed show markedly more significant association with *HLA-DQB1* expression compared to *HLA-DQA1* expression, with more than two times greater fold change in *HLA-DQB1* associated with genotype than *HLA-DQA1*. Thus, the likelihood is that linkage in the HLA-DQ-DR locus is causing the association with *HLA-DQA1* expression where is occurs. Susceptibility to leprosy is likely to be associated with *HLA-DQB1*, narrowing the wider locus implicated by both the Zhang and Wong studies.

5.3.9 Further genotyping in a leprosy cohort

In order to test the theory that *HLA-DQB1* expression was directly related to the leprosy susceptibility of individuals, 78 cases and 122 controls were selected from a previously described cohort to study leprosy in New Delhi, India (Malhotra 2005). Five SNPs from across the *HLA-DQA1/HLA-DQB1* region were genotyped, using a Sequenom Mass Array primer extension (Jurinke 2001, Jurinke 2005) assay by Dr Tom Parks, that had been associated with *HLA-DQB1* expression in the healthy volunteer cohort (see Table A.6 for primer sequences). The top SNP associated with *HLA-DQB1* expression rs9273448 was not typed as a sequenom assay was unable to be designed for this SNP. However, the imputed SNPs rs9273440 and rs3135006, in perfect LD with rs9273448, were typed using the sequenom assay; results are shown in Table 5.4.

SNP		*HLA-DQB1* expression in Caucasians			association with leprosy in Indians from New Delhi			
		freq.	relative expression	*P*-value	case freq	control freq	OR (95% CI)	*P*-value
rs17843603	G	0.42	4.889	7.31×10^{-51}	0.69	0.45	2.79 (1.89-4.34)	3.6×10^{-6}
rs9273410	A	0.46	0.403	1.41×10^{-29}	0.22	0.40	0.42 (0.27-0.68)	3.0×10^{-4}
rs9273440	T	0.79	0.390	3.18×10^{-66}	0.07	0.11	0.61 (0.29-1.27)	0.182
rs3134970	A	0.79	0.390	1.41×10^{-29}	0.07	0.11	0.59 (0.29-1.25)	0.166
rs3135006	A	0.79	0.390	3.18×10^{-66}	0.07	0.11	0.62 (0.30-1.29)	0.198

Table 5.4: *HLA-DQB1* expression associated SNPs and their association with leprosy susceptibility. 5 SNPs known to be associated with increased *HLA-DQB1* expression were genotyped in a leprosy cohort and control group. They were analysed to determine if they were over-represented in the case group compared to the controls using PLINK to perform Pearson's χ^2 allelic test and logistic regression.

The variant rs17843603 was shown to be significantly associated with disease status ($p=3.6 \times 10^{-6}$), and is known to be strongly associated with *HLA-DQB1* expression ($p=7.31 \times 10^{-51}$). SNP rs9273410 is also shown to have significant disease association. Both ORs of susceptibility to leprosy highlight the skewed prevalence of the alleles in the leprosy population compared to the healthy controls. The remaining three loci were not significantly associated with disease status. However, the minor allele frequency was only 11% in the control population and so statistical power to assess their potential imbalance was limited. This was a large contrast to the allele incidence in the Caucasian healthy volunteer population where it was found at a much higher frequency. In the Caucasian population these alleles were associated with lower expression and so would presumably be associated with decreased leprosy susceptibility if an association was present. There is a small reduction in frequency of the allele in the case population compared to the controls however it does not reach statistical significance.

Haplotypes for the Indian cohort were also reconstructed using Haploview (Barrett 2005) defining haplotypes as previously described using the confidence interval definition (see Section A.4.1). A GCC haplotype at rs17843603, rs9273410 and rs9273440 was found to be especially associated with disease, with a case frequency of 0.62 compared to a control frequency of 0.36 (OR 2.88), $p=4.7 \times 10^{-7}$.

Interestingly, the SNP rs107843603 was found to be in near direct linkage with rs1071630 (r^2 = 0.92, D' = 0.98), the SNP identified as being strongly associated with leprosy (Wong 2010).

5.3.10 *HLA-C*

HLA-C has been previously implicated in immune related disease such as HIV and psoriasis (Fellay 2007, Liu 2008). The gene shows variable gene expression associated with genotype as demonstrated in section 3.3.12. To further fine map the variants involved in control of gene expression qPCR was used to analyse the expression of *HLA-C* in the cohort of 288 genotyped healthy volunteers. An eQTL study was performed using PLINK as previously described in section 4.3.3, and the results are shown in the Manhattan plot in Figure 5.18.

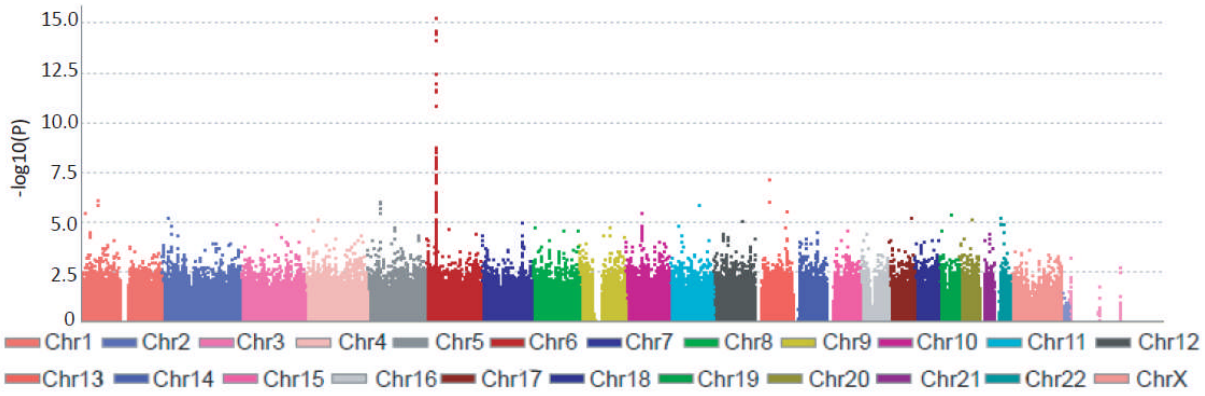


Figure 5.18: Manhattan Plot showing the association of SNPs with *HLA-C* expression. *HLA-C* expression in 288 healthy volunteer PBMC samples was determined using qPCR and normalised to *ACTB*. Genotyping of gDNA was performed using the Illumina Infinium high-density genotyping bead arrays and association analysis was performed using PLINK.

A clear peak of association to expression is seen over variants around the *HLA-C* locus on chromosome 6, while no SNPs at other locations reach genome-wide significance. The top SNP associated with *HLA-C* expression (rs10484554) was used to group volunteers according to genotype, and expression was compared between the three groups. This was also performed for the SNP rs9264942 which was previously shown to be identified with HLA-C cell surface expression and HIV-1 progression (Thomas 2009), and was shown to be associated with *HLA-C* gene expression in 96 healthy volunteers (see Figure 5.19).

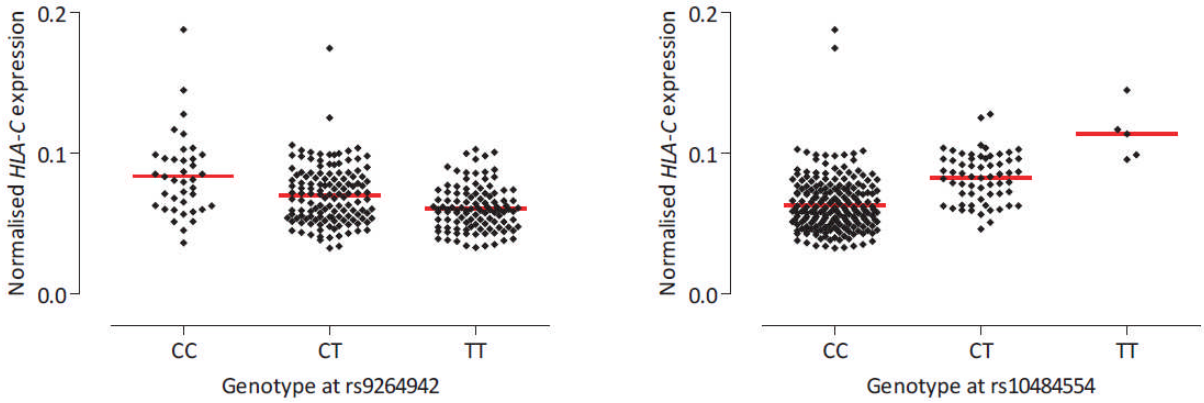


Figure 5.19: Association of i) genotype at rs9264942 and ii) genotype at rs10484554 with *HLA-C* expression. *HLA-C* expression is measured using SYBR green qPCR in duplicate and normalised to *ACTB*. Expression values are separated by genotype at i) rs9264942, ($p=7.36x10^{-9}$) and ii) rs10484554 ($p=5.19x10^{-16}$), and plotted. Mean expression is indicated for each genotype by the red bar. Both SNPs indicate a significant difference in mean expression of *HLA-C* between the different genotypes by Kruskal-Wallis test ($p<0.0001$).

When recombination over the *HLA-C* locus was considered, a large region was identified as having variants associated with variable expression of the gene, which contained many other genes in addition to *HLA-C* (see Figure 5.20). However, peaks of recombination bordered the most associated SNPs, and the most associated SNPs were located close to one another upstream of *HLA-C*.

In both cases, a significant difference in mean gene expression could be determined according to genotype showing the strong genetic influence on *HLA-C* expression. While *HLA-C* expression was known to vary according to genotype at rs9264942, it was unclear whether this was likely to be a causal SNP or the result of haplotypic structure in the region. As other SNPs surrounding *HLA-C* are associated with gene expression at greater significance, it is likely that the rs9264942 SNP is tagging a functionally important variant rather than directly influencing expression itself (see Figure 5.20).

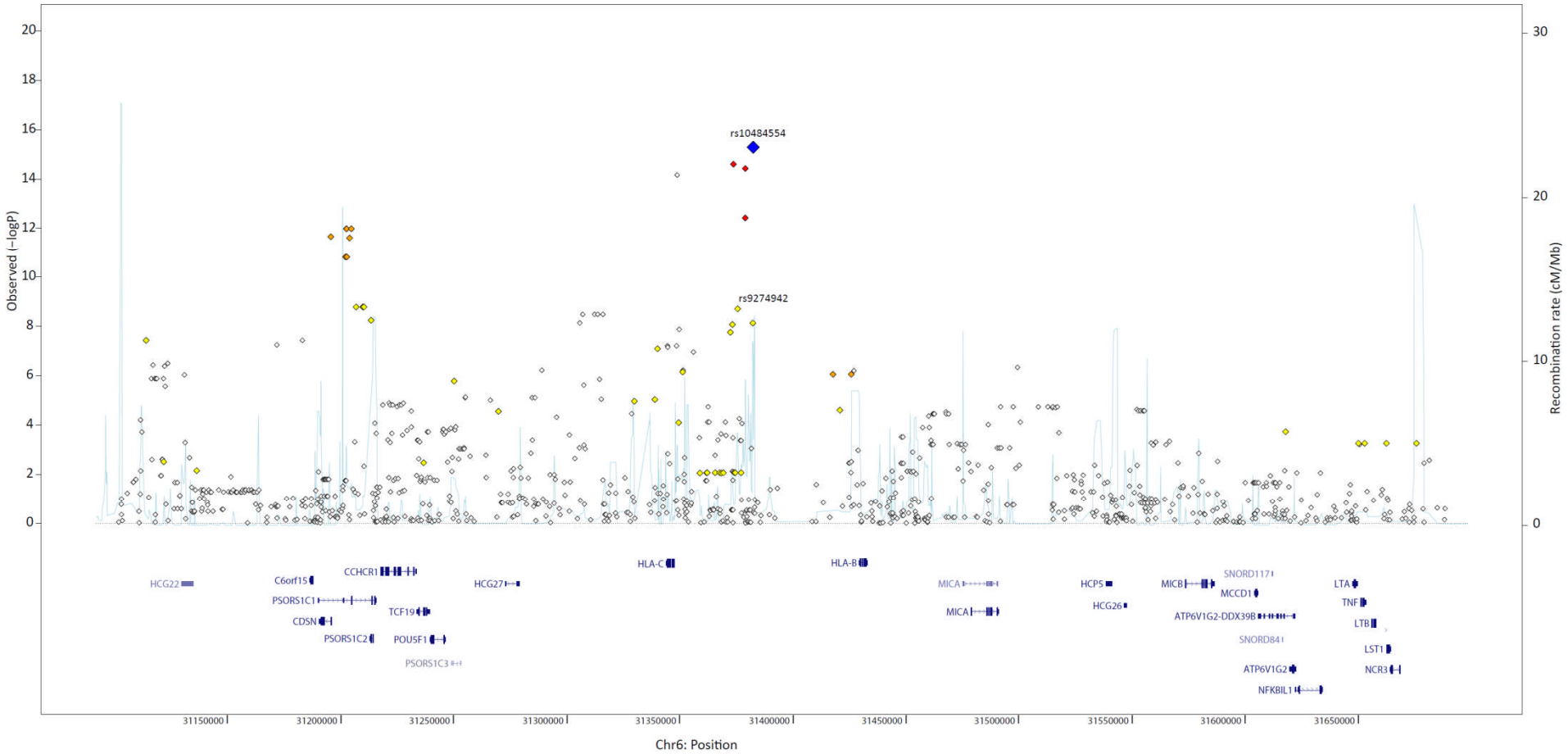


Figure 5.20: Recombination plot showing the *HLA-C* locus and surrounding area. The SNP most associated with *HLA-C* expression, rs10484554, is shown in blue. rs9264942, a previously associated SNP to *HLA-C* expression is indicated, and is found to be in low LD with the most associated SNP (R2=0.266). SNPs are shown in red (LD 0.8≤1.0 to rs10484554), orange (LD 0.5≤0.8), yellow (LD 0.2≤0.5) and white (LD 0.0≤0.2).

5.3.11 *HLA-C* expression association with complex disease

Several SNPs shown to be associated with *HLA-C* expression at a level of genome-wide significance by this study have been strongly associated with disease traits by GWAS studies. The GWAS catalogue was used to find SNPs associated with disease that had been also shown to be associated with *HLA-C* expression in this study, and the results are shown in Table 5.5:

Disease trait	*HLA-C* expression associated SNP	p value *HLA-C* association	p value disease association	Reference
AIDS progression	rs10484554	$5.19x10^{-16}$	$6.00x10^{-8}$	(Limou 2009)
	rs9264942	$7.36x10^{-9}$	$6.00x10^{-32}$	(Fellay 2009)
Psoriasis	rs10484554	$5.19x10^{-16}$	$4.00x10^{-214}$	(Liu 2008, Strange 2010)
HIV-1 control	rs3815087	$1.04x10^{-12}$	$8.00x10^{-8}$	(Fellay 2009)
	rs9264942	$7.36x10^{-9}$	$3.00x10^{-35}$	(Pereyra 2010)
	rs9264942	$7.36x10^{-9}$	$6.00x10^{-12}$	(Fellay 2009)
Follicular lymphoma	rs6457327	$3.69x10^{-8}$	$7.00x10^{-6}$	(Conde 2010)
	rs6457327	$3.69x10^{-8}$	$5.00x10^{-11}$	(Skibola 2009)
Stevens-Johnson syndrome and toxic epidermal necrolysis (SJS-TEN)	rs3815087	$1.04x10^{-12}$	$3.00x10^{-7}$	(Genin 2011)

Table 5.5: GWAS defined disease associated SNPs that show association with *HLA-C* expression and their related traits. The table shows disease traits that are hypothesised to be associated to *HLA-C* expression by shared associated SNPs determined by GWAS and eQTL analysis of healthy volunteers.

The SNP rs9464942 that had been shown to have HIV-1 control association was unsurprisingly shown to be one of the most associated SNPs with *HLA-C* gene expression given its previous identification in variation of HLA-C cell surface expression (Thomas 2009). The SNP that showed most association with *HLA-C* expression variation (rs10484554) was also found to be associated with AIDS progression and another SNP showing strong association to *HLA-C* expression, rs3815087, was associated with HIV-1 control. Interestingly, two SNPs associated with *HLA-C* expression were found to be associated with different skin conditions, psoriasis and Stevens-Johnson syndrome/ Toxic Epidermal Necrolysis (Liu 2008, Genin 2011). Both are immune-related conditions that are likely to be triggered by an external influence, such as drug administration or infection, although in the case of psoriasis the mechanism by which this occurs is far from clear (Schön and Boehncke 2005). In previous work in the genetics of psoriasis, two MHC loci have been identified as associated strongly, both *HLA-C* and

HCP5. *HCP5* is an endogenous retroviral element that has also been associated with HIV-1 progression, although the mechanism for this association is unclear. Weak LD is found between the SNPs associated with *HLA-C* and those disease associated SNPs found within *HCP5* although separate effects of the loci have been previously demonstrated (Fellay 2007). One way that the *HCP5* variant could be acting is via differential expression, therefore *HCP5* expression was assayed in the healthy volunteer cohort, to see if the same variants were affecting *HCP5* expression that affected *HLA-C* expression. However no association with expression was found on eQTL analysis for *HCP5* when expression of this gene was assayed by qPCR (see Section A.4.2 – Figure A.8).

5.3.12 *HLA-C* expression and HLA allele association

Classical MHC alleles were investigated for their association with *HLA-C* expression, as the *HLA-C* locus has been associated with several diseases, such as psoriasis and HIV-1 progression. As SNPs associated with *HLA-C* expression were also disease associated (determined by GWAS, see Table 5.5), the HLA types were considered to determine if any longer-range effects could be responsible for gene expression differences (Figure 5.21).

HLA-C expression is associated with HLA-C alleles, most notably with HLA-C*06. The HLA-B*57 allele also shows elevated gene expression of *HLA-C* in comparison to other 2-digit types.

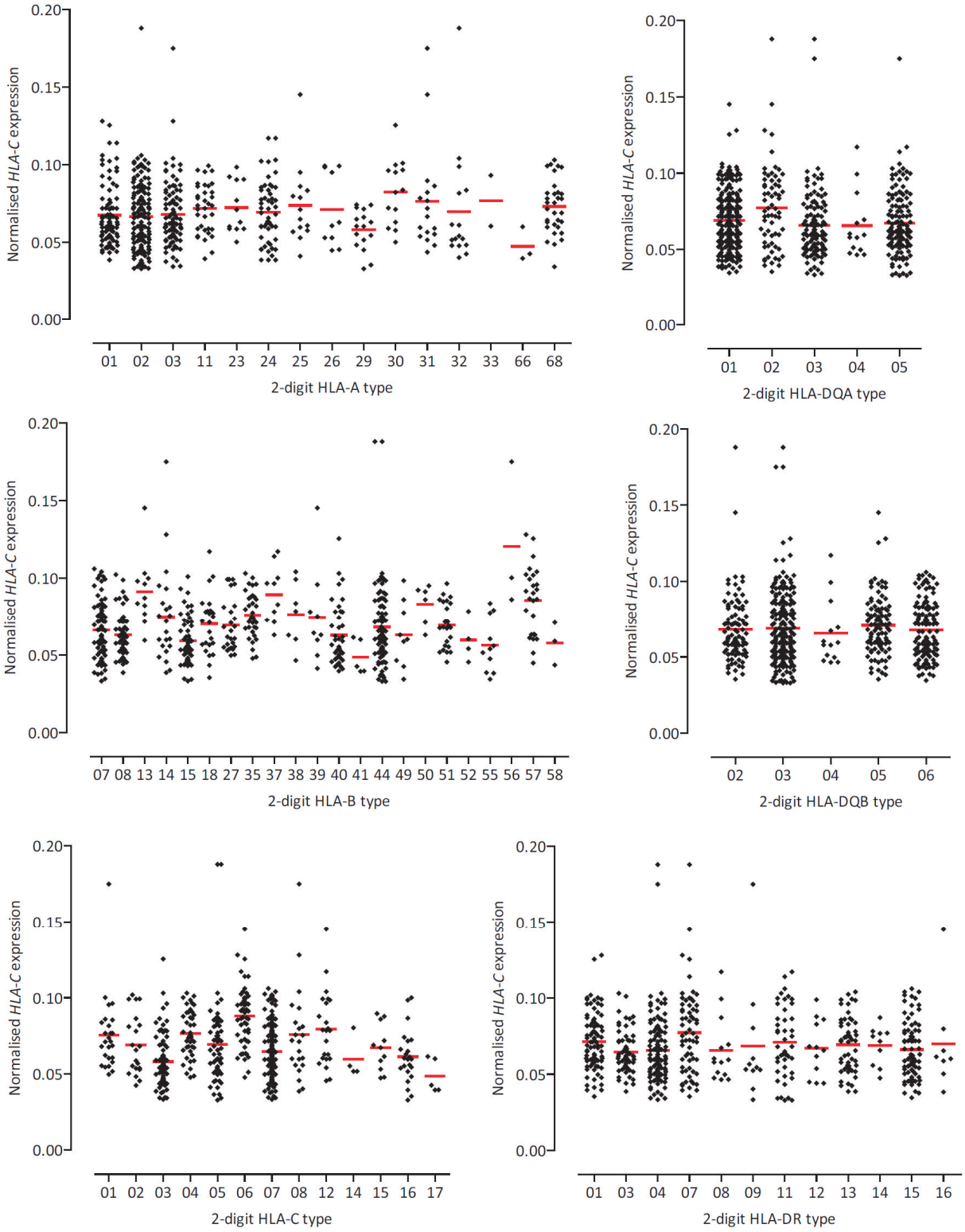


Figure 5.21: *HLA-C* expression according to 2-digit HLA types. The *HLA-C* expression associated with a copy of each different allele is shown for six regions across the MHC. Mean expression is indicated for each group by the red bar. ASE is evident according to HLA-C 2-digit type, where those individuals possessing a copy of HLA-C*06 show higher *HLA-C* expression when compared with other HLA-C types ($p<0.0001$ when analysed using a Mann-Whitney test). HLA-B type may also play a role in influencing *HLA-C* expression, individuals possessing HLA-B*57 were shown to have significantly higher *HLA-C* expression ($P<0.0001$ when analysed using a Mann-Whitney test).

To determine if this was an effect of the LD structure of the region or a true effect of the HLA-B*57 allele expression of *HLA-C* was analysed by copies of HLA-C*06 and possession of HLA-B*57 in Figure 5.22. The haplotypic effect of HLA-C*06 seen when analysing all *HLA-C* 2-digit types is confirmed, but the presence of HLA-B*57 does not appear to alter *HLA-C* expression.

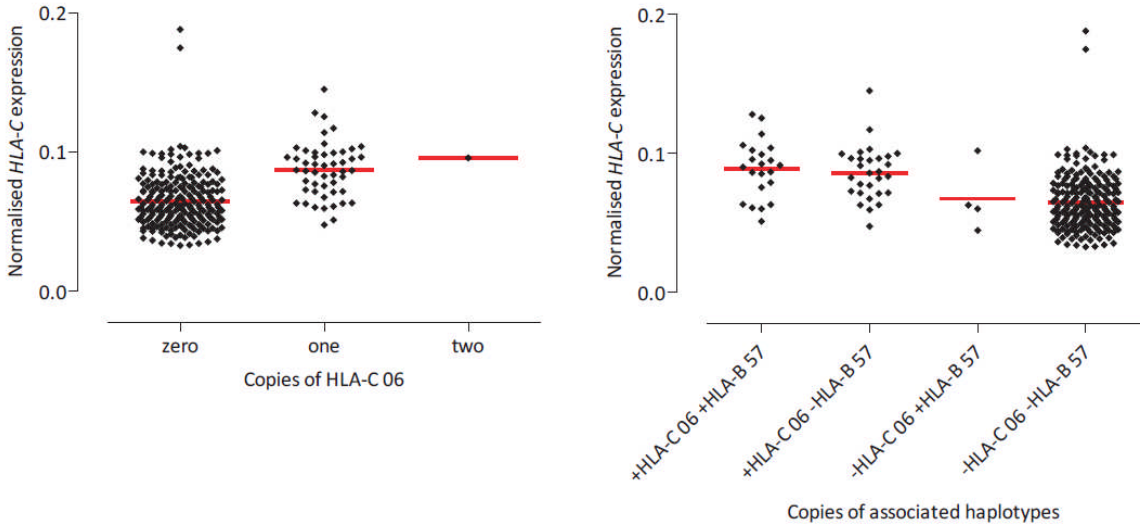


Figure 5.22: *HLA-C* association with HLA-C*06 and HLA-B *B57 2-digit type. *HLA-C* expression in the healthy volunteer cohort is shown according to i) copies of HLA-C*06 type and ii) possession of HLA-C*06 and HLA-B*57. Mean expression is indicated for each group by the red bar.

5.3.14 *HLA-C* expression-associated haplotype reconstruction

Having shown that *HLA-C* expression was associated with the HLA-C 2-digit type HLA-C*06, and that the T minor allele for rs10484554 was over-represented in those volunteers possessing the HLA-C*06 allele, it was predicted that rs10484554 was likely to be found in an area of strong LD that was in turn associated with the HLA-C*06 serological type. Haploview was used to assess the LD in the area most associated with gene expression, shown in Figure 5.23:

Figure 5.23: No clear LD over most associated SNPs with *HLA-C* expression. The SNP most associated with *HLA-C* expression, rs10484554 is indicated in the regional LD plot, but no clear linkage can be seen around this SNP. Upstream and downstream of the region are haplotype blocks that appear to be associated with each other however this association breaks down over the area of rs10484554.

While regions of strong LD were found nearby to the SNP rs10484554, it was located in an area with several peaks in the recombination rate (see Figure 5.20), and LD was reduced.. Additional haplotype analysis of the SNPs in the region was thus impossible to perform. It is therefore possible that rs10484554 could be acting in a functional manner itself, rather than tagging an area containing the functional variant, or that there are other causative SNPS that have not been identified in the region. It is also likely that the HLA-C*06 allele is associated with *HLA-C* expression via possession of the rs10484554 T allele, rather than due to a wider effect of the whole allele or the particular effect of variants at the binding cleft in HLA-C.

5.4 Discussion

5.4.1 DQ region

Some evidence of association with expression for HLA-DQ genes was found in the initial volunteer cohort (see Chapter 3) and the link between genetic variation and variable gene expression of these genes was further investigated as described in this chapter. The much greater numbers of individuals and the much denser genotyping data available for the second cohort of volunteers provided substantially increased power to resolve variants associated with variable gene expression. While it was expected that the HLA-DQ genes would all be closely related in terms of expression, the r^2 values showed that the gene expression was not in perfect correlation. *HLA-DQA1* was not most strongly correlated to *HLA-DQB1* expression despite their formation of a class II antigen presenting heterodimer. Instead *HLA-DQA2* was the most strongly correlated gene to *HLA-DQA1* expression and the same SNP (rs17843604) was found to be most associated with expression of both genes. This suggests tightly controlled co-regulation of these genes. *HLA-DQB2* could not be accurately assessed due to the extremely low expression of the gene. It was therefore impossible to define genetic variants that may affect its expression accurately, and to accurately correlate expression to the expression of the other HLA-DQ genes.

eQTL analysis showed that many of the same SNPs are correlated to the expression of *HLA-DQA1, HLA-DQA2* and *HLA-DQB1* in varying degrees. It is therefore striking that *HLA-DQB1* shows a much stronger genetic association with variable gene expression in this cohort than either of the other two genes. *HLA-DQB1* expression variation was hypothesised to play a functional role due to this much more variable expression than other HLA genes. The two separate distributions of *HLA-DQB1* expression in the volunteer cohort were also intriguing and suggested the possibility of different disease susceptibility for either those with either high or low expression of *HLA-DQB1*.

5.4.2 *HLA-DQB1* and disease association

In order to investigate the relationship between HLA-DQB1 and disease, association of *HLA-DQB1* expression-related SNPs were compared with disease SNPs found in GWAS. Many of the disease associations that were found involved diseases that have long been associated with the HLA-DQ

region. For example, both T1D and coeliac disease had disease-associated SNPs showing convincing associations with *HLA-DQB1* expression (Hakonarson 2007, van Heel 2007, WTCCC 2007, Cooper 2008, Dubois 2010). However, many of these SNPs have not previously been definitively associated with any particular gene in the HLA-DQ region; instead the HLA-DQ-DR locus as a whole has been put forward as the associated region. Often the association is thought to be in linkage to an effect on the protein product, such as alteration of binding sites or the level that the protein can be trafficked to the cell surface. Study of the gene expression of the HLA-DQ genes has clearly marked *HLA-DQB1* as the most associated gene to the genetic variation, and shows that there is a wide range of expression of *HLA-DQB1* possible. This implicates the level of *HLA-DQB1* expression as an important variable trait in disease and shows the value of combining GWAS analyses with other functional work.

5.4.3 HLA type and *HLA-DQB1* expression

It has previously been hard to separately dissect the effects of the HLA-DQ and HLA-DR regions due to the strong LD and highly polymorphic nature of the region (Fernando 2008a). While SNP-free primers according to the latest dbSNP release could be designed for the DQ genes this proved impossible for the HLA-DR genes. However by analysing gene expression in the light of HLA-DQ and HLA-DR 2-digit types, it showed that gene expression of *HLA-DQB1* was not affected by the HLA-DR locus. This strengthened the assertion that HLA-DQB1 was playing an important role in disease pathogenesis without additional influence of the surrounding loci. While this does not rule out a combination of both HLA-DQ and HLA-DR alleles in the region affecting disease susceptibility differently, it does imply that *HLA-DQB1* expression acts independently of the HLA-DR locus and so suggests a mechanism for how some of the associated variants may be influencing disease.

This analysis also allows further interrogation of known disease associations in order to predict the genes that may be functionally responsible for the associations. For example, the HLA-A3-B7-Cw7-DR15 haplotype (PGF) that is carried by around 10% of the European population is known to be associated with MS susceptibility and protection from diabetes (Stewart 2004). This haplotype contains the HLA-DQB*06 allele, which was shown to be highly associated with the increased expression of *HLA-DQB1* in the healthy volunteer cohort. When individuals possessing a copy of the

full PGF ancestral haplotype (HLA-A3-B7-Cw7-DR15) from the volunteer cohort were compared to all other volunteers, gene expression was found to be significantly increased in the "PGF" individuals for *HLA-DQB1* expression, but not for the other HLA-DQ genes. MS susceptibility was also found to be associated with *HLA-DQB1* related SNPs, suggesting that the mechanism for susceptibility to this autoimmune condition could be through expression of *HLA-DQB1*. HLA-DQB1 cannot explain all the disease associations previously found with the HLA-A3-B7-Cw7-DR15 haplotype however; T1D susceptibility was found to have SNPs in common with SNPs associated with increased *HLA-DQB1* expression. This is at odds with the HLA-A3-B7-Cw7-DR15 haplotype and its strong protection against T1D (Larsen and Alper 2004), highlighting the different molecular mechanisms that contribute towards genetic disease associations.

5.4.4 Disease associations, *HLA-DQB1* and *HLA-C*

When the GWAS catalogue was used to search for SNPs that had association to diseases and also to *HLA-DQB1* expression, leprosy was one of the diseases shown to be associated with *HLA-DQB1* expression related SNPs (see Tables 5.2 and 5.3). However, the SNP included in the GWAS catalogue was not the SNP shown to have the strongest association with *HLA-DQB1* expression (rs602875), but was in fact associated at a more modest level with *HLA-DQB1* expression. This was because in the study, rs9271366 was found to be refractory to genotyping (Zhang 2009) and thus another marker was used. It is clear that these SNPs do not have the same association with the expression of *HLA-DQB1* from the large order of magnitude difference in the p value of association. Without using SNPs not published in the GWAS catalogue, the stronger association may have been lost and an association with leprosy susceptibility may not have been investigated further.

Considering the role of *HLA-DQB1* in antigen presentation it is perhaps unsurprising that up-regulation of *HLA-DQB1* should be associated with increased risk of autoimmune conditions. However increased expression is also associated with infectious diseases such as hepatitis (Mbarek 2011) and leprosy as previously discussed. *HLA-C*-increased expression is associated with maintenance of viral set-point in HIV-1 and non-progression of the disease (Thomas 2009), a contrast to the increased *HLA-DQB1* expression association with disease susceptibility. This highlights how

the different antigen presenting pathways may have quite separate effects on disease susceptibility. Both *HLA-DQB1* and *HLA-C* expression have also been associated with various cancers such as follicular lymphoma (in both cases) and Hodgkin's Lymphoma (*HLA-DQB1*). This could suggest a role of immune-related genes in cancer pathogenesis.

Although the presence of the SNP rs9273363 was shown to be strongly associated with the expression of *HLA-DQB1* in the volunteer group that had low expression levels of this gene, this SNP or proxies to it were not implicated as associated with any known GWAS hits. However, other SNPs associated at a genome-wide level of significance with both low and high expression of *HLA-DQB1* were found to have associations to the same diseases suggesting that a dose control of *HLA-DQB1* may be important in different disease traits ranging from autoimmune to infectious as well as various types of cancer incidence. Effect of gene dosage of HLA-DQ genes in autoimmune disease has been shown to play a role in severity of disease, for example copy number of HLA-DQA*05 and HLA-DQB*02 impact upon both the risk and severity of coeliac disease (Murray 2007). Here the study of gene expression in combination with genotype and HLA type gives insight into how copy number could mechanistically be influencing disease susceptibility.

To further investigate the association of *HLA-DQB1* expression with leprosy, five SNPs shown to be associated with *HLA-DQB1* expression through the healthy volunteer analysis were genotyped in an Indian cohort of leprosy sufferers and healthy controls (data courtesy Tom Parks). Of these, two SNPs (rs17843603 and rs9273410) were found to be significantly associated with leprosy, while the other three were suggestively associated albeit at a non-significant level. This was extremely encouraging as it showed the value of using gene expression analysis to investigate known disease association. It was also reassuring to see that variants identified in a Caucasian population could be informative in the Indian population and that clear differences were seen between the cases and controls in the frequency of the *HLA-DQB1* associated SNPs. Further analysis in a leprosy cohort could include assessment of *HLA-DQB1* expression levels to see if they were significantly elevated in the disease cohort compared to controls.

5.4.5 Genetic association of *HLA-C* expression

Following the analysis of *HLA-C* expression in the initial volunteer cohort of 96 individuals it was not clear whether there were particularly strong genetic associations seen in relation to the control of *HLA-C* levels in healthy individuals. However, as genotyping was sparse over the area and *HLA-C* was one of the genes seen to have high variability between the three MHC homozygous LCLs analysed using the MHC array, it was taken forward for further analysis in the second volunteer cohort. The larger sample size allowed a much clearer region of association to be defined and previously associated SNPs to *HLA-C* expression were confirmed in this new cohort, for example the association of rs9264942 with *HLA-C* levels (Thomas 2009). *HLA-C* is an interesting candidate due to its disease associations and relatively recent evolution; it is found only in great apes and humans (Kulpa and Collins 2011).

An analysis of association with *HLA-C* expression levels by HLA type showed that possession of the allele HLA-C*06 was associated with higher expression levels. Analysis of *HLA-C* expression has shown how gene expression analysis in conjunction with known alleles can help to break down ambiguity of association in the MHC. For example, the HLA-B*57 allele was shown not to be associated with the change in *HLA-C* expression levels where previously the different effects of particular HLA-B and HLA-C allele possession had been difficult to distinguish from one another. The HLA-B/C block has been previously associated with several diseases (Trachtenberg 2009) so to be able to define more precisely the effect of variants will help to determine how these disease associations may be functioning. While the variants associated with HIV-1 viral load and progression have been said to be attributable to HLA-A and HLA-B HLA types, this study has shown that they are strongly associated with *HLA-C* expression (Catano 2008).

5.4.6 Lack of *HCP5* expression association with *HLA-C*

With this in mind expression of *HCP5,* a close by gene that like *HLA-C* had been implicated in psoriasis susceptibility through GWAS, was investigated to see if SNPs associated with *HLA-C* expression were also affecting expression of *HCP5* (Liu 2008). Wide variation in expression of *HCP5* was not seen in the volunteer cohort and no strong genetic association could be found, therefore we hypothesised

that variants identified as associated with *HLA-C* expression were only affecting expression of *HLA-C*. This implies that *HCP5* association with psoriasis is not brought about by modulation of gene expression, and is a separate effect from that of *HLA-C* or is the result of linkage in the region. HLA-C*06 has been implicated as the primary allele for psoriasis susceptibility through resequencing of exons, so this may implicate the LD of the MHC in this association rather than an actual effect of *HCP5* (Nair 2006). A variation in *ERAP1* has also been associated with psoriasis, but only when the HLA-C*06 allele is also present, suggesting that an interaction occurs between these two molecules (Strange 2010). This is particularly interesting given that ERAP1 is associated with antigen processing and class I peptide presentation. It is possible that increased expression in *HLA-C* as well as the structural variation of the HLA-C*06 allele is responsible for this association.

5.4.7 Conclusion

Following on from the observations of variable gene expression in the first healthy volunteer cohort, the new, larger healthy volunteer cohort has been used to more precisely map variants associated with gene expression of HLA-DQ genes and *HLA-C*. After showing that genetic variation was associated with *HLA-DQB1* and *HLA-C* expression most strongly, the effect of HLA-type on both of these genes has also been investigated. This has confirmed that the HLA-C*06 haplotype previously associated with psoriasis is found more often in individuals expressing *HLA-C* more highly. Disease association with gene expression has been investigated for both *HLA-DQB1* and *HLA-C*, implicating expression of *HLA-DQB1* in susceptibility to leprosy and *HLA-C* in psoriasis and control of HIV-1.

Chapter 6 – mRNA-sequencing to investigate ASE levels

6.1 Introduction

Advances in next generation sequencing (NGS) have led to an assumption that RNA-seq will replace array technologies for genome-wide gene expression analysis as cost decreases and methodologies improve. RNA-seq avoids the need for alternative splicing, RNA editing, differential allelic expression and fusion transcripts to be taken into account when designing an array-based study by defining the transcriptional boundaries and capturing expressed SNPs to allow allelic detection and discrimination (Costa 2010). This chapter explores the use of RNA-seq for allele-specific quantification with a particular focus on the MHC to assess the applicability of this approach to a highly polymorphic region.

6.1.1 RNA-seq methodology

The technique of RNA-seq is detailed in Figure 6.1. cDNA libraries are made from an RNA sample either from total RNA or using a selector such as polyA+ to ensure mRNA is captured. Sequencing adaptors are added to the cDNA, which is then sequenced in either a single-end read or a paired-end read to give the sequence information. The resulting sequence reads are aligned to an appropriate reference genome for analysis of gene expression levels. Advantages of RNA-seq include the ability to detect non-coding RNAs and transcripts that were previously unidentified as no hybridisation is required. As sequences are mapped to the reference genome in an unambiguous manner, normalisation of data is not necessary unlike in array analysis. Saturation, where the gene is expressed too highly to give an accurate measure, does not occur in RNA-seq, meaning it has a greater dynamic range for expression analysis (Wang 2009). Additionally, the use of RNA-seq allows investigators to resolve expression of specific alternatively spliced isoforms far beyond the specificity of an array (Pastinen 2010).

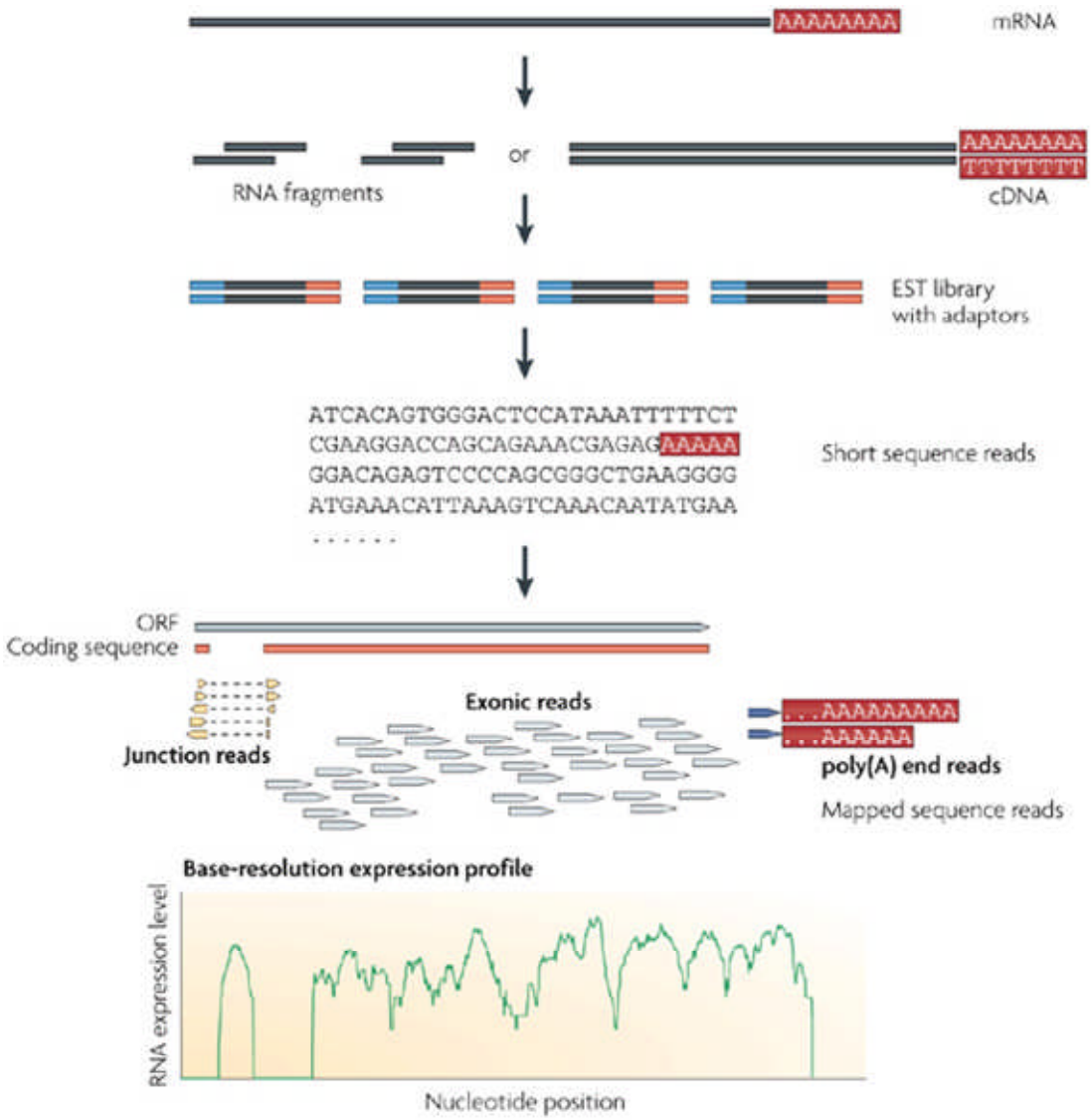


Figure 6.1: Schematic showing the process of mRNA-seq. Reproduced from (Wang 2009). mRNA is isolated by polyT capture and then fragmented or made into a cDNA library then fragmented. A library of Expressed Sequence Tag (EST) cDNA is then generated with adaptors ligated onto each end. The EST cDNA library is then sequenced using a suitable platform. The sequence results are then mapped to a suitable reference sequence. Mapped short sequence fall into three categories; junction reads that span exonic junctions, exonic reads that are contained within one exon and poly(A) end reads that have the poly A tail connected to part of an exon.

6.1.2 Use of RNA-seq in gene expression studies

RNA-seq has already been implemented in several studies in both humans and model organisms including yeast (Nagalakshmi 2008), Drosophila (McManus 2011) and mice, where ASE has been shown to differ through development of the mouse blastomere (Tang 2011). As previously discussed in Section 1.1.10, LCLs from the HapMap project have been analysed using mRNA-seq with

encouraging results. A greater discovery of eQTL was possible than with traditional arrays and the dynamic range was comparable with a sequencing depth of 10 million (Montgomery 2010, Pickrell 2010). A recent study has also suggested that RNA editing is a more common phenomenon than previously thought. Combining DNA- and RNA-seq showed many differences between the DNA sequence and the sequences of the transcripts (Li 2011), however this result is currently controversial (Chakravarti 2011). While it remains to be seen if this assessment is the result of experimental or analytic bias, it remains an interesting question at a point when increasing numbers of studies are starting to routinely use sequencing technologies. Despite these studies, a systematic analysis of the use and limitations of RNA-seq has not been performed and the analysis pipelines that are used vary extensively between experiments. Comparative analysis across five different RNA-seq data sets led to the conclusion that RNA-seq data could be biased in terms of transcript length, GC composition and sequence polymorphism, and that a method to reduce this bias gave more consistent results particularly when comparing data sets (Zheng 2011).

6.1.3 Concerns over use of RNA-seq for gene expression detection

Although detection of ASE is an attractive prospect when using RNA-seq, the sensitivity of the technique has not been fully tested and the lower bound of allelic difference in expression that can be detected with confidence by RNA-seq is unknown. When using mRNA-seq, where the mRNA is reverse transcribed, the reverse transcriptase enzyme will fall off the mRNA causing bias towards the 3' end of the gene, as the method of capture for mRNA is to use polyT to pair with the polyA tail of the mRNA (Costa 2010). The read depth, or coverage, required for acceptable analysis of gene expression using RNA-seq data also depends on the expression level of the gene of interest. Finally, bias in read mapping has been highlighted as a potential problem particularly in terms of choice of reference genome, use of random hexamer priming in library generation, and GC content of transcripts (Hansen 2010, Zheng 2011).

In terms of using RNA-seq for gene expression analysis in the MHC, there are several potential problems. Firstly, the repetitive nature of much of the DNA sequence in the MHC leads to difficulty in mapping the short reads generated by an RNA-seq experiment. Secondly, RNA-seq can give biased

results between different genes. While this is not a problem when comparing expression of one gene between different samples, it may cause problems in the MHC as the different haplotypes can lead to biases in the mapping with a number of reference sequences available, leading to inaccurate quantification of levels of gene expression. Studies using RNA-seq to analyse gene expression including in the MHC (Montgomery 2010, Pickrell 2010) have focussed only on individual point analysis of expression. This makes analysis over an entire haplotype difficult, a problem that was investigated using the MHC custom array to analyse gene expression (Vandiedonck 2011) but that requires further thought if MHC gene expression is to be analysed using RNA-seq. The technique of mapping reads to a reference genome may also be problematic in the MHC due to its diversity, where short reads are difficult to align to a reference and may in fact be so different from the reference that they bear no resemblance to it (Holcomb 2011).

6.1.4 Testing the accuracy, sensitivity and reproducibility of RNA-seq

In order to address some of these concerns, particularly in response to the potential for read mapping and choice of reference sequence leading to bias in detection of ASE, an mRNA-seq experiment investigating ASE was initiated (see Figure 6.2). In this experiment two of the HLA-homozygous LCLs, PGF and COX, previously interrogated with the MHC array, were used (Vandiedonck 2011). This had three main advantages:

i) gene expression across the MHC had been intensely studied previously in these LCLs;

ii) both LCLs had been completely sequenced across 4.75Mb of the MHC (Stewart 2004) and these sequences formed part of the GRCh37 human genome release (http://www.genomereference.org) allowing accurate and specific read mapping;

iii) from genotyping using the HumanCVD array in the Knight lab a significant majority of the rest of the genome in these LCLs was found to be homozygous (Vandiedonck, unpublished data), allowing the study of RNA-seq and ASE to extend beyond the MHC.

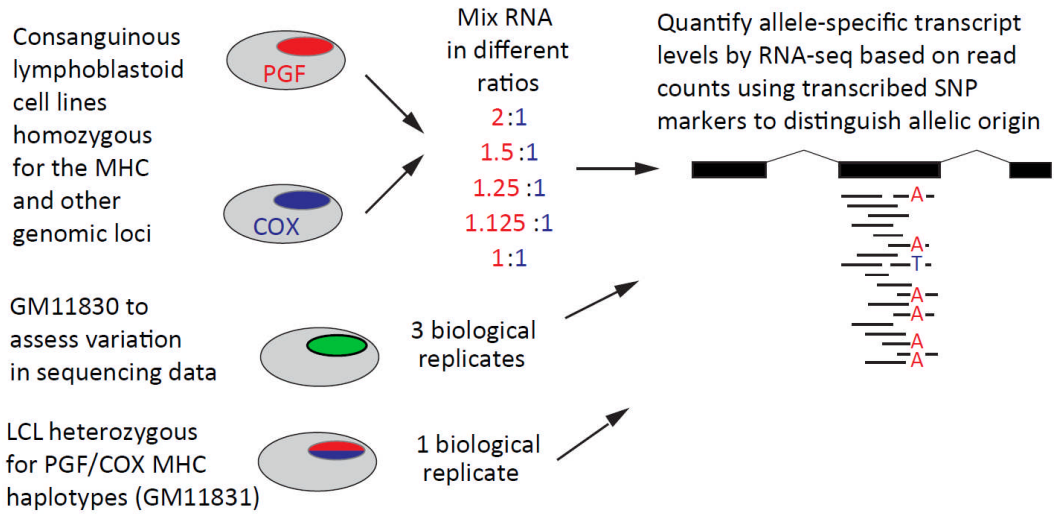


Figure 6.2: Experimental design for analysis of allele specific RNA-seq and its limitations in gene expression quantification. MHC homozygous cell line total RNA is mixed in differing ratios before mRNA library prep and sequencing. Differential ratios can then be analysed to determine the limits of using RNA-seq in analysis of ASE.

"Synthetic heterozygote" samples were produced by mixing the RNA from the two LCLs in a series of different ratios (Figure 6.2). This allowed the allelic imbalance that could be confidently and accurately detected to be determined. Additionally, an LCL from the CEU HapMap panel that was a "natural heterozygote" i.e. contained similar haplotypes to the PGF and COX LCLs, one on each chromosome, was also selected and sequenced following the same protocol to assess in vivo allelic imbalance and the variation associated with these haplotypes. For this, the LCL NA11831 was identified which has the haplotypes (HLA A1-B8-Cw7-DQA5-DQB2-DR3) and (HLA A1-B7-Cw7-DQA1-DQB6-DR15, almost identical to the two haplotypes of interest: PGF (HLA A3-B7-Cw7-DQA1-DQB6-DR15) and COX (HLA A1-B8-Cw7-DQA5-DQB2-DR3). To ensure that sequencing information beyond the MHC was useful, RNA samples of the individual LCLs COX and PGF alone were also sequenced for comparison. Genomic DNA from both LCLs was exome sequenced to allow additional "synthetic heterozygote" sites to be identified over the entire transcribed genome. The exome data from the PGF and COX LCLs would also allow assessment of how effective the exome capture and subsequent sequencing in the MHC was, as the full sequence for this region was already known.

6.1.5 Reproducibility of RNA-seq

In order to assess the reproducibility of RNA-seq results, a further CEU LCL, NA11830 was used to generate RNA samples for sequencing to determine the variation seen between biological replicates of the same sample. The LCL was grown up in three separate cultures to be able to assess if the variability of RNA-seq reads and mapping was consistently low to allow future use in experiments where small changes in expression may be expected and so variance in the technique for discovery would render the results impossible to analyse.

6.2 Aims

In this chapter data are presented to assess the utility of mRNA-seq for allele-specific discrimination with a particular focus on the MHC. To do this, HLA-homozygous LCLs were to be analysed, either alone or in different mixed allelic ratios together with heterozygous LCLs. Specific aims were to define:

1. The effects of reference sequence choice for mapping on assaying allele-specific expression. This included the question of whether whole exome sequencing should be routinely performed with RNA-seq when quantifying allele-specific expression of a sample with no gene sequence information to improve reliability of read mapping especially in relation to the MHC.
2. The lowest threshold of allelic difference (ratio) at which allele-specific expression quantification could be discriminated and if this varies depending on abundance of the transcript.
3. The likelihood of alternative isoforms affecting the accuracy of measurement of gene expression, in particular focusing on the MHC.

6.3 Results

6.3.1 Sample Preparation and Data Alignment

RNA was extracted from one biological replicate of the COX, PGF and GM11831 LCLs harvested in mid-log growth phase, and gDNA was also harvested from the COX and PGF LCLs. Three biological replicates of the LCL GM11830 were prepared in the same manner. The RNA was quantified using the Qubit RNA assay and COX and PGF gDNA was assayed using the Qubit DNA assay to ensure the most accurate concentration reading was taken. GM11831 and three replicates of GM11830 were submitted directly for sequencing, while the COX and PGF RNA were submitted either separately or mixed in ratios as detailed in Figure 6.2 and Table 6.1.

Cell Line	Material	Ratio	No. of Replicates	Type of sequencing
PGF	gDNA	NA	1	Exome-seq
COX	gDNA	NA	1	Exome-seq
PGF	RNA	NA	1	RNA-seq
COX	RNA	NA	1	RNA-seq
PGF:COX	RNA	1:1	1	RNA-seq
PGF:COX	RNA	1.125:1	1	RNA-seq
PGF:COX	RNA	1.25:1	1	RNA-seq
PGF:COX	RNA	1.5:1	1	RNA-seq
PGF:COX	RNA	2:1	1	RNA-seq
GM11830	RNA	NA	3	RNA-seq
GM11831	RNA	NA	1	RNA-seq

Table 6.1: LCL samples submitted for sequencing. Samples submitted for sequencing and the type of sequencing used in the study are indicated, with the number of biological replicates for each sample.

Libraries were generated from each of the RNA samples using polyT capture of mRNA. Once the mRNA had been selected the RNA was fragmented and then cDNA libraries were made. The cDNA was end-repaired, A-tailed and adaptor-ligated before amplification for sequencing. Sequencing was performed using the Illumina GAIIx. Exome sequencing was performed on exome DNA captured using RNA baits, again using the Illumina GAIIx. Libraries were prepared and sequenced by the High Throughput Genomics Group. Reads from the mRNA-sequencing were aligned to the reference (Hg19) using Stampy, a statistical algorithm for mapping Illumina reads (Lunter and Goodson 2011).

COX and PGF haplotypes were defined exactly across the MHC, as only reads which could be mapped unambiguously were left in the analysis. Beyond the MHC, heterozygotic sites were defined by calling the variants and interrogating bases that were highlighted as having alternate possibilities. Exome sequencing also allowed these sites to be identified separately from the mRNA-seq data. Read pile-ups were then compared between the alternate bases directly allowing the differing ratio between the samples to be assessed.

In the mRNA-seq data, between 97.2-98.1% of all reads were successfully mapped to the reference genome. The alignment of the mRNA-seq data to chromosome 6 revealed an apparently low percentage of the total mapped reads (between 1.5-1.9% of all reads). This may be due to the use of all eight MHC haplotype references that are available as part of the Hg19 reference genome. Where a read found in the MHC is present in more than one haplotype it will be assigned a low mapping score due the inability to unambiguously define the location. This will result in many reads from chromosome 6 being removed from further analysis. Those that remain in the analysis may still have a low mapping score, if another locus is quite similar to the one defined by the mapper, and due to the repetitive nature of the MHC this is likely to affect many transcripts. Alternatively, extremely highly expressed genes, such as housekeeping genes, that are located on other chromosomes may skew the percentage coverage over the whole genome. In this case, a relatively small number of genes could account for a large proportion of the total reads, and thus would reduce the percentage of reads found on chromosome 6.

Of the reads from the exome sequencing data, between 99.3-99.6% were mapped to the reference genome successfully. The GC content of the reads should be higher than the average across the genome due to higher GC content in exomes. GC content was between 45.6 ± 10.5-45.9 ± 10.4%, above the genomic average of 40%. Finally, good coverage over defined exons was observed in the data, as between 39.6-40.1% of all reads were found in exons. Defined exons are a relatively restricted parameter, meaning that many more reads are likely to be exonic, but that these exons were not defined during analysis.

6.3.2 Detection of different allelic variants in *GAPDH*

Initially, the housekeeping gene *GAPDH* was investigated to find coding sequence differences between the COX and PGF LCLs using the exome sequencing data (see Figure 6.3). Four exonic single nucleotide differences were discovered between the LCLs and these were then analysed in the five mixed ratios of PGF and COX RNA, the mixed samples creating a series of 'SNPs' at these sites of single nucleotide differences between the LCLs referred to in subsequent text as 'ratio SNPs' or 'raSNPs'. Coverage varied between the raSNPs, immediately highlighting how the detection of ASE over a whole gene would be affected by the position of informative SNPs.

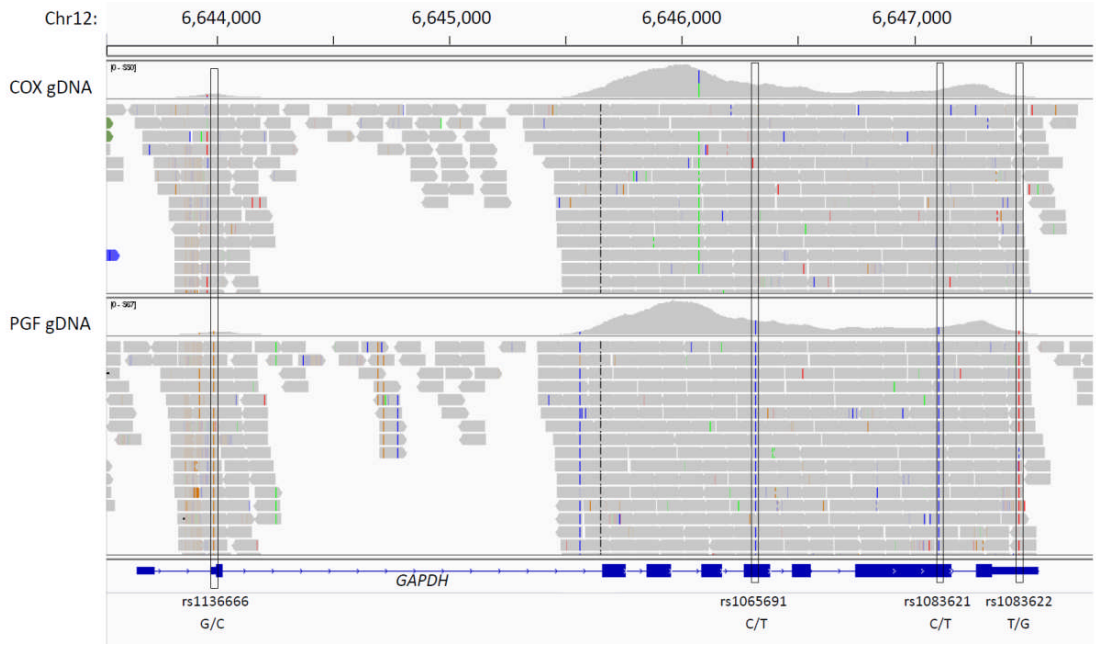


Figure 6.3: Position of *GAPDH* raSNPs. Mapped exome-seq gDNA data was analysed using IGV software to visualise exonic sequence against the reference genome Hg19. Exonic homozygous variants that occurred in only one of either the COX or PGF LCLs were selected in the GAPDH gene as they created an artificial heterozygote that could be analysed to assess sensitivity and variability of detection of ASE. raSNPs in the *GAPDH* gene are indicated by circling with a grey box.

To confirm that the raSNPs identified by exome sequencing were correct, they were also checked in the single mRNA-seq data for each LCL. After confirming a single variant present in each cell line mapped unambiguously in the mRNA-seq data, the ratios of the alleles were compared between the five mixed RNA samples in Figure 6.4.

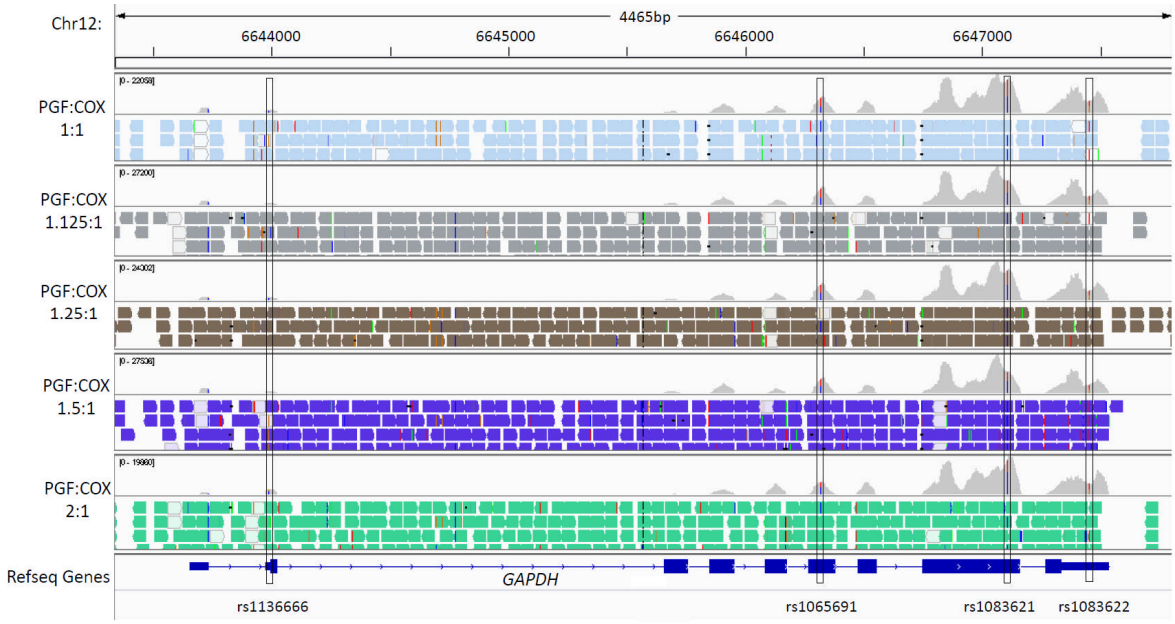


Figure 6.4: *GAPDH* mRNA-seq tracks and differential coverage identified by raSNPs between the COX and PGF LCLs. mRNA-seq reads that have been mapped for the five mixed samples using Stampy are aligned to the Hg19 genome and visualised in IGV. SNPs between COX and PGF are indicated by circling with a grey box.

Read counts at each raSNP were analysed to determine if ASE could be quantified as expected between the different mixed ratio samples. Good levels of correlation were found between the trends of observed and expected values (Table 6.2 and Figure 6.5) where high read depth coverage was present; ASE differences followed the trend predicted from the mixed ratios. For the raSNP where coverage was lower (rs1136666), the 2:1 allelic ratio showed higher quantities of the PGF LCL SNP compared to the 1:1 ratio as expected but there was no clear association between the different mixes and the amount of each allele seen (Figure 6.5).

***GAPDH* SNP**	**rs1136666 G:C**	**rs1065691 C:T**	**rs1083621 C:T**	**rs1083622 T:G**
Mix PGF/COX 1:1	498:1572 (0.32)	3471:5147 (0.67)	9063:9493 (0.95)	2764:3945 (0.70)
Mix PGF/COX 1.125:1	304:1442 (0.21)	5318:6284 (0.85)	8238:8661 (0.95)	2793:3516 (0.79)
Mix PGF/COX 1.25:1	497:1414 (0.35)	4676:4995 (0.94)	9659:8638 (1.12)	2735:3439 (0.80)
Mix PGF/COX 1.5:1	335:1438 (0.23)	5250:4868 (1.08)	8395:6378 (1.32)	3152:3248 (0.97)
Mix PGF/COX 2:1	1174:1618 (0.73)	3771:2613 (1.44)	11270:6403 (1.76)	3168:2566 (1.23)

Table 6.2: Read counts for the *GAPDH* SNPs between COX and PGF RNA. Read counts at each individual variant in *GAPDH* between COX and PGF were obtained using IGV. The expected ratio in each sample is shown as the sample name and the observed ratio is indicated for each raSNP in brackets.

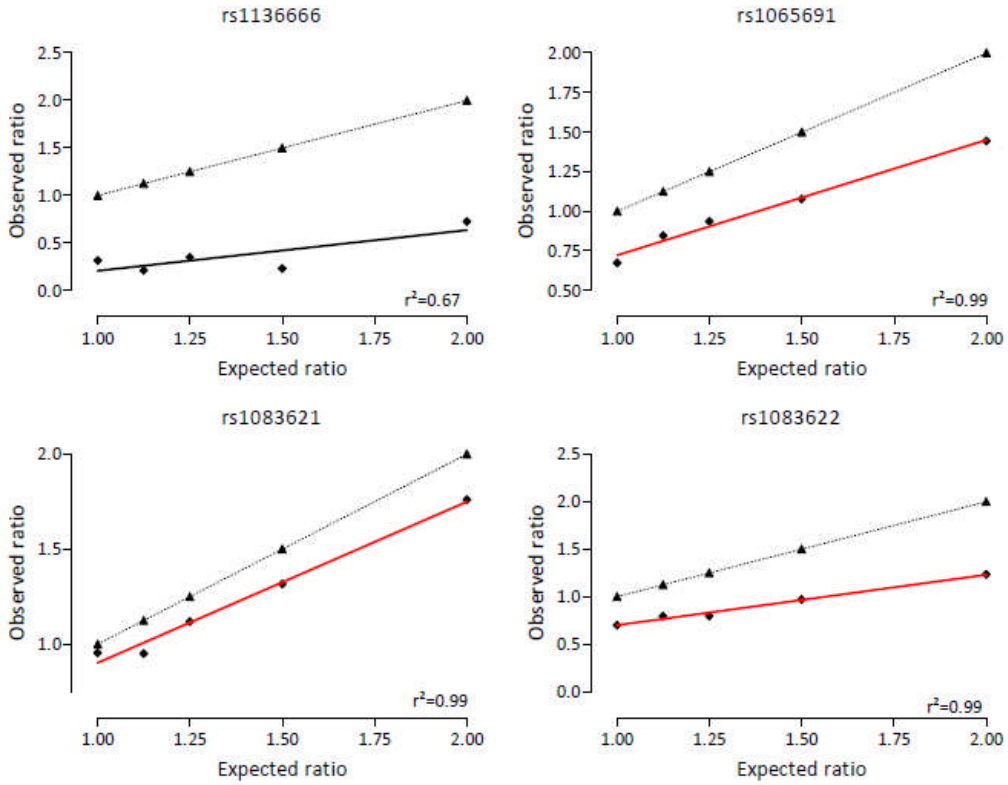


Figure 6.5: Observed and expected ratios of expression between COX and PGF RNA in the *GAPDH* gene. Four raSNPs found between the PGF and COX LCLs were investigated between the different mixed ratio samples. The raSNPs were identified by exome sequencing, and ratios were determined using read counts; rs1136666 between 335 and 1618 reads; rs1065691 between 2613 and 6284 reads; rs1083621 between 6378 and 11270 reads; rs1083622 between 2566 and 3945 reads. The read counts found in the LCL RNA for COX and PGF samples alone were as follows: rs1136666, 3128 in COX (C) and 1279 in PGF (G); rs1065691, 8645 in COX (T) and 7451 in PGF (C); rs1083621, 19177 in COX (T) and 15028 in PGF (C); rs1083622 10049 in COX (G) and 4262 in PGF (T). No significant association between the ratios was seen for rs1136666 when observed and expected ratios were analysed using linear regression (p=0.096), but a significant association was seen for all other SNPs; rs1065691 (p=0.0007), rs1083621 (p=0.0007), rs1083622 (p=0.0006). The black dotted line indicates the expected line if the observed were to exactly match the expected ratio, while the red line indicates the observed ratios.

Despite a lower coverage than other raSNPs found in *GAPDH*, rs1136666 still had quite high counts for both alleles at all ratios of RNA mixtures analysed (see Table 6.2). An approximation was used to define the variance of sampling from the total read number at each variant based on the Bernoulli distribution. This allowed the raSNP rs1136666 to be analysed to determine if high variability in sampling over the SNP accounted for the lack of association between the observed and expected ratio. Estimated standard deviation is plotted in Figure 6.6; the variation between the predicted

possible ratios show variance in sampling caused by the reads is unlikely to affect the association of the observed ratios with the expected ratios to the extent seen for this SNP.

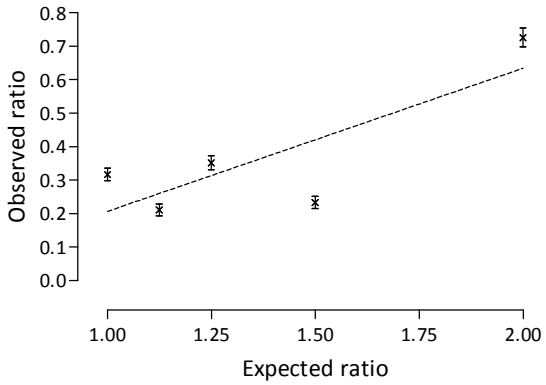


Figure 6.6: Observed and expected ratios of expression between COX and PGF RNA at raSNP rs1136666. Upper and lower estimates for each ratio were estimated by finding the predicted extremes of read counts for each allele according to a Bernoulli distribution (Standard deviation = ½√N, where N= total number of reads at the raSNP). These were plotted as error bars to the experimentally determined ratio between each allele at rs1136666.

6.3.3 Effect of differential isoforms on determination of ASE

Different known transcript isoforms of *GAPDH* were investigated to see if this could explain the lower association with the expected ratios seen in the observed RNA expression estimates for this raSNP. The low counts were particularly seen for the variant seen in the PGF LCL RNA (G), with a lowest read count of 335 compared to the lowest read count of 1414 seen for the COX variant (C), suggesting that any isoform difference may predominantly affect detection of expression in PGF. When the alternative isoforms were retrieved using Ensembl (www.ensembl.org), it was clear that rs1136666 was found very near the 5' end of most isoforms, often outside the protein coding sequence (see Figure 6.7). Different isoforms between the LCLs could help to explain the poor association between the observed and expected ratios seen for raSNP rs1136666, where ASE of particular exons causes the ratio to be skewed between the alleles of the raSNP.

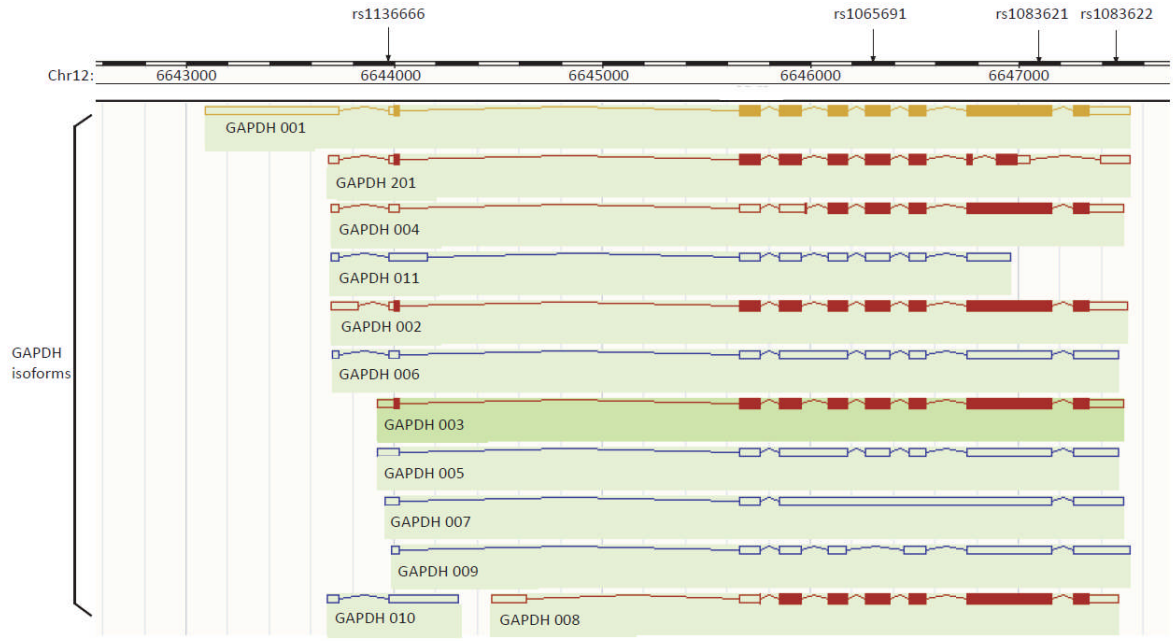


Figure 6.7: Alternative transcripts of the *GAPDH* gene and location of homozygous coding SNPs found between COX and PGF. Alternative isoforms of *GAPDH* are indicated with boxes representing exons, and lines introns. The full length transcript is shown in yellow, and transcripts that create a protein product are coloured in red. Untranslated transcripts are shown in blue. In all cases protein coding sequence is indicated by blocked colour within the exon box. There are several different isoforms of the *GAPDH* gene created by alternative splicing and intron retention or inclusion. Expression over the three raSNPs at the 3' end of the gene follows the expected ratios more closely than expression over the raSNP found at the 5' end of the gene (rs1136666).

6.3.4 Further analysis of raSNPs in Non-MHC genes

Following confirmation that ASE could be detected with relatively high accuracy when read counts were high, further genes with high expression and raSNPs created by the mixing of COX and PGF were investigated to assess this further. Firstly, another gene with high read counts over both variants was selected for analysis; *RPS20* (encoding a ribsomal protein forming part of the 40S subunit), which has at least 200 reads over each allele at the raSNP between COX and PGF (chr8:56985815). This showed a good level of correlation between observed and expected ratios (Figure 6.8).

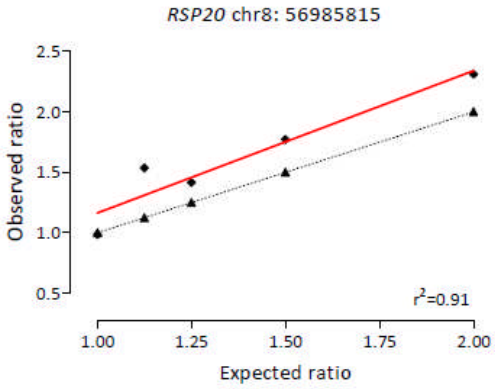


Figure 6.8: Observed and expected ratios of expression between COX and PGF RNA in a highly expressed gene. Two raSNPs identified between the PGF and COX LCLs using exome sequencing were investigated between the different mixed ratio samples in *RPS20*. The PGF LCL has a C at chr8:56985814 and COX has a T, and each variant had between 355 and 2509 reads in each mixed RNA sample. COX RNA alone had 791 reads of the T allele, and PGF RNA alone had 811 reads. There was significant association between the observed and expected ratios when analysed with linear regression (p=0.012). The black dotted line indicates the expected line if the observed were to exactly match the expected ratio, while the red line indicates the observed ratios.

Although the observed ratio of the different alleles did not match exactly the expected ratio, there was a clear relationship between the amounts of PGF compared to COX RNA in the sample. Like the *GAPDH* analysis the sample with the PGF:COX ratio of 1.125:1 was found outside the line of best fit (see Figure 6.8). This may indicate a limit in the ability to discern small difference in ASE even among highly expressed genes. The *POLR2A* gene (encoding the largest subunit of RNA polymerase II) was then investigated as an raSNP between the LCLs was identified, chr17:7399867; G in PGF and A in COX that had lower read counts per allele. In this case the counts over each allele ranged from 48-119. A previous study has suggested that ASE can only be accurately determined if 50 reads for both alleles can be identified (Heap 2010). This result showed that the association between the different ratios at this variant was even greater than the analysis of *RPS20*, despite its higher read counts. This implies that a level of around 50 reads can give as accurate a result as far higher numbers of a given transcript.

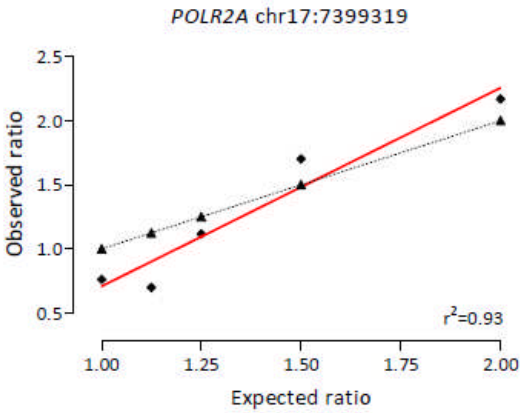


Figure 6.9: Observed and expected ratios of expression between COX and PGF RNA in a moderately expressed gene. A variant was identified in *POLR2A*, chr17:7399319, between the COX and PGF LCLs using exome sequencing. Analysis of the ratios found between the COX and PGF alleles showed that the observed ratios closely followed the expected ratios when analysed with linear regression (p=0.0068). The black dotted line indicates the expected line if the observed were to exactly match the expected ratio, while the red line indicates the observed ratios.

Another gene with around 50-100 reads per allele for at least some samples was identified, *NADK*, and two raSNPs in this gene were investigated to confirm this finding.

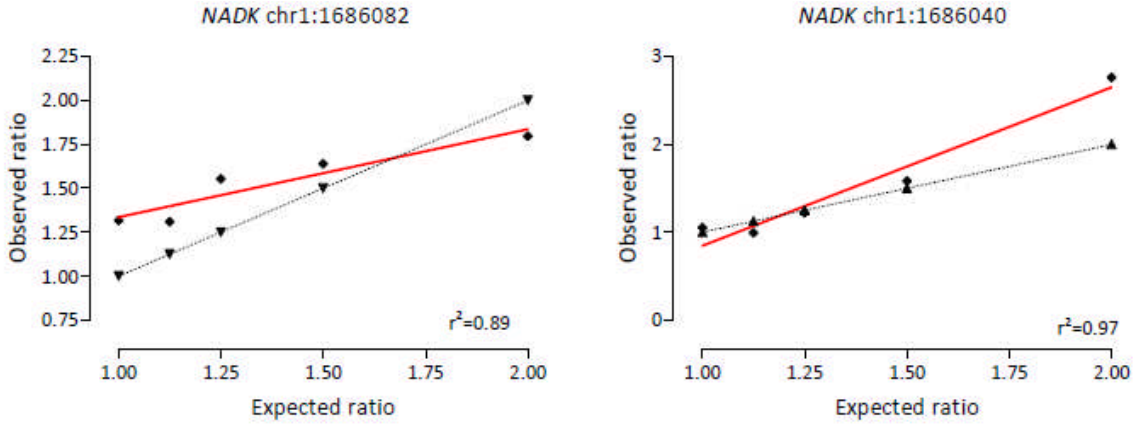


Figure 6.10: Observed and expected ratios of expression between COX and PGF RNA in the *NADK* gene. Two raSNPs found between the PGF and COX LCLs were investigated between the different mixed ratio samples. The SNPs were identified by exome sequencing; alleles at chr1:1686082 (not defined in dbSNP) are G in PGF, A in COX. Variants at rs4751, at chr1:1686040 are T in PGF and G in COX. Reads over both were significantly associated with the expected ratio when analysed with linear regression (chr1:1686082, p=0.018; rs4751, p=0.004) showing good detection of ASE. The black dotted line indicates the expected line if the observed were to exactly match the expected ratio, while the red line indicates the observed ratios.

Like the example in *POLR2A*, the observed and expected ratios are significantly associated in the mixed RNA samples. This changes however when smaller read counts over different alleles are investigated. For example, the gene *NOTCH1* was explored over three raSNPs, and only one was found to have a significant association between the expected and observed ratios of the allele counts. In this gene the read counts for the SNPs ranged from 5-77, however counts over rs2229975 were lowest, ranging from 5-28, and this showed least association between the observed and expected ratios of the PGF and COX alleles (Figure 6.11).

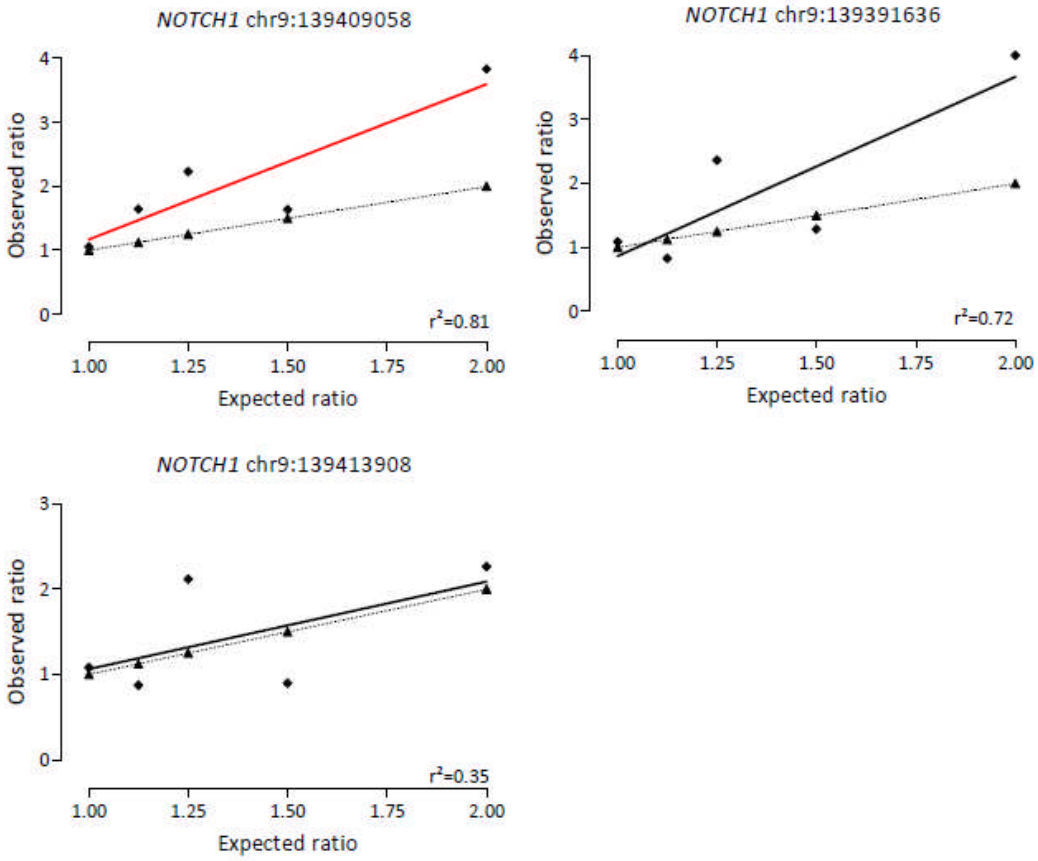


Figure 6.11: Observed and expected ratios of expression between COX and PGF RNA in the *NOTCH1* gene. Two raSNPs found between the PGF and COX LCLs were investigated between the different mixed ratio samples. The SNPs were identified by exome sequencing; the raSNP at chr9:139409058, (not defined in dbSNP) G in PGF, A in COX. At chr9:139391636 the raSNP is rs2229974, G in PGF, A in COX and chr9:139413908 the raSNP is rs2229975, C in PGF and T in COX. Reads over chr9:139409058 were analysed with linear regression and were significantly associated with the expected ratio (chr9:139409058, p=0.037) while at rs2229974 although the r^2 value was indicative of a weak association it was not statistically significant (p=0.069). For rs2229975 there was least evidence of association between the observed and expected ratios (p=0.29). The black dotted line indicates the expected line if the observed were to exactly match the expected ratio, while the red or black solid line indicates the observed ratios.

Another two low expressed genes, *GUSB* and *CDK1*, were analysed in the same manner (Figure 6.12). In *CDK1* the raSNP rs1871446 had no read counts above 45, with most counts in the samples found in the range of 11-30. The raSNP found in *GUSB*, chr7:65429360 had similarly low counts for both of the alleles, between 11-56 counts.

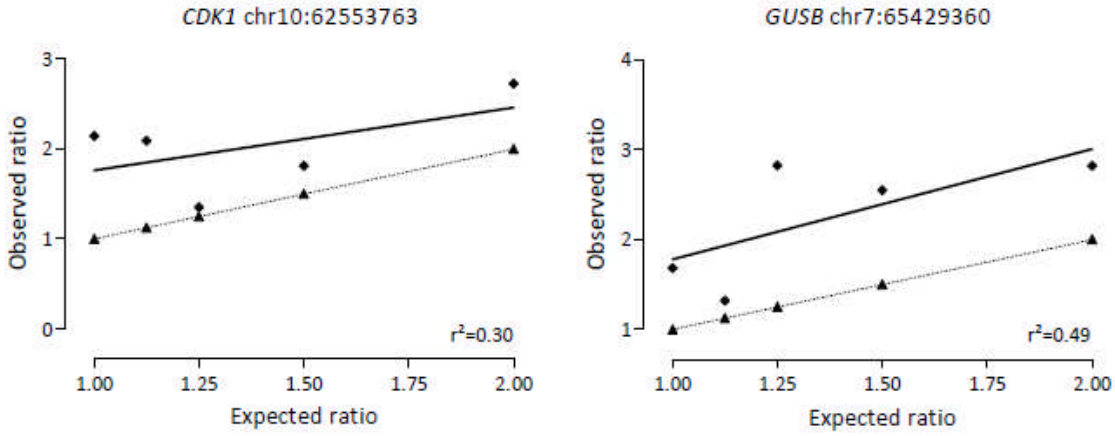


Figure 6.12: Observed and expected ratios of expression between COX and PGF RNA transcripts of *CDK1* and *GUSB*. Both genes contained raSNPs located using exome sequencing data (*CDK1* – rs1871446, *GUSB* – chr7:65429360, undefined by dbSNP). There was no significant association between the observed and expected ratios when analysed using linear regression (*CDK1*, p=0.33; *GUSB*, p=0.19) although the overall trend showed the PGF allele increasing as more PGF RNA was used to make the mixed samples. The black dotted line indicates the expected line if the observed were to exactly match the expected ratio, while the black solid line indicates the observed ratios.

6.3.5 Assessment of variation in mRNA-seq data between biological replicates

The mRNA-seq data for the LCL GM11830 allowed comparison between three biological replicates to assess the reproducibility of detection of ASE. Initially the four raSNPs found in *GAPDH* were analysed in GM11830, where a 1:1 ratio of each allele was expected. Although the observed ratio did not necessarily match the predicted ratio, it did correspond to the ratios observed in Figure 6.5. Each raSNP showed good reproducibility between each of the biological replicates. The difference between the ratios observed could be the result of differential expression between exons of the genes and alternatively spliced isoforms.

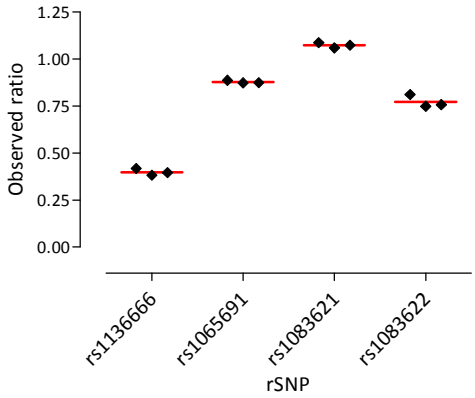


Figure 6.13: Observed ratios between the raSNPs found in *GAPDH* in mRNA-seq for GM11830. Three biological replicates of mRNA-seq of the LCL GM11830 are assessed for biological reproducibility. All observed ratios show good reproducibility between the biological replicates. Standard deviation between replicates at each raSNP is as follows: rs1136666 SD=0.018, rs1065691 SD=0.0076, rs1083621 SD=0.014, rs1083622 SD=0.034. Read counts for these raSNPs ranged from; rs1136666, 670 and 2225 reads; rs1065691, 2966 to 5420 reads; rs1083621, 10032 to 12671 reads; rs1083622, 3134 to 4552 reads.

As previously discussed, *GAPDH* shows high gene expression, where no read count for any allele is below 670 reads. *GUSB* and *POLR2A* were therefore analysed to look at genes with lower total read counts for raSNPs. The highest read count for any allele of an raSNP in these genes was 55, and variability was far higher than seen between biological replicates assessed over *GAPDH* raSNPs. It is apparent that lower read counts make it more difficult to be certain that the value for a given allelic difference is accurate.

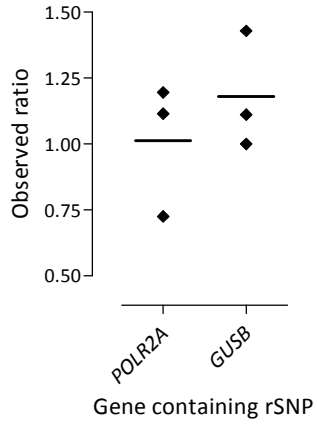


Figure 6.14: Observed ratios between the raSNPs found in *POLR2A* and *GUSB* in mRNA-seq for GM11830. Three biological replicates of mRNA-seq of the LCL GM11830 are assessed for biological reproducibility at chr17:7400815 in the gene *POLR2A* and at chr7:65425894 in the gene *GUSB*. The biological replicates show more variability than raSNPs analysed in *GAPDH*. Standard deviation between the ratios at each raSNP is 0.25 in *POLR2A*, and 0.22 in *GUSB*. Read counts at the raSNPs were between 35 and 55 reads in *POLR2A*, and between 14 and 20 reads in *GUSB*.

6.3.6 Analysis of mRNA-seq in the MHC

Coverage of RNA-seq reads over the gene dense MHC was lower than expected, with less than 2% of all reads in each sample mapping to chromosome 6. This was likely to represent problems with mapping in the area. Looking in the MHC at a highly expressed gene, *HLA-B*, immediately revealed these problems. *HLA-B* was first investigated using exome sequencing to determine if SNPs that were known to differ in alleles between the COX and PGF LCLs could be mapped accurately.

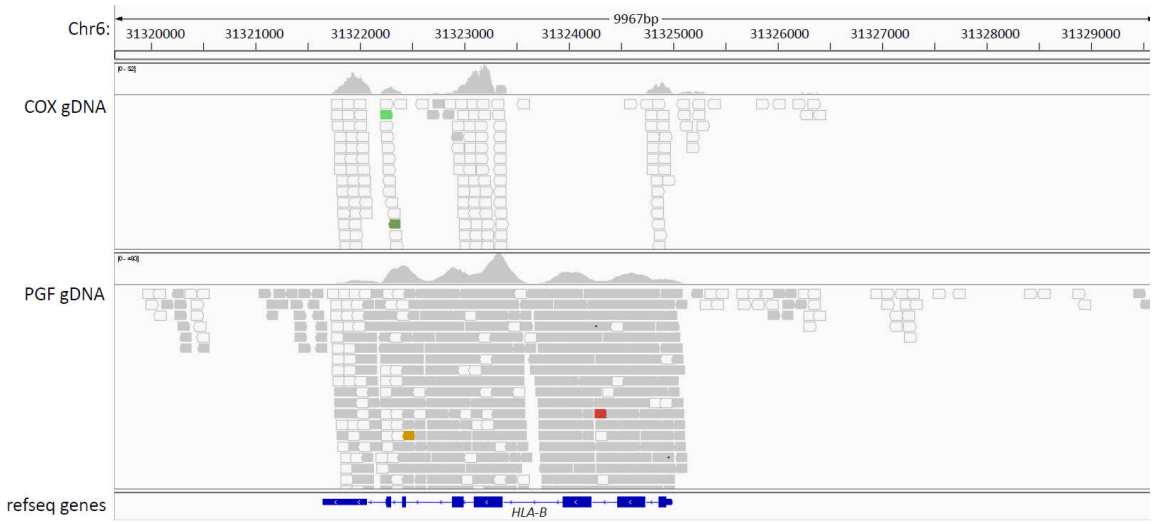


Figure 6.15: Exome sequencing data mapped over *HLA-B* in COX and PGF gDNA. Reads mapped with confidence are shaded in grey; reads shown by the white boxes reflect lack of confidence in the mapping and suggest that another locus is similar to the reference here.

No raSNPs were called in the *HLA-B* gene using Stampy, likely to reflect mapping bias in the MHC. Lower coverage seen in COX is likely to be the result of ambiguity in mapping. PGF may have a more unique sequence over *HLA-B*, while COX may have stronger similarities to other haplotypes represented in the Hg19 reference sequence; therefore reads in COX are less accurately mapped. As no raSNPs could be determined using this form of mapping, ASE could not be analysed.

Chromosome 6 was remapped using only one of the eight possible reference sequences over the MHC, again using Stampy. In the relatively highly expressed *HLA-B*, comparison of the known sequences in both the COX and PGF LCLs revealed that 19 exonic raSNPs should have been created by mixing the COX and PGF RNA. When the COX sequence was used to remap the data, no further

raSNPs were identified compared with using all eight haplotypes. However, using data mapped only to PGF allowed detection of some of the raSNPs (see Figure 6.16).

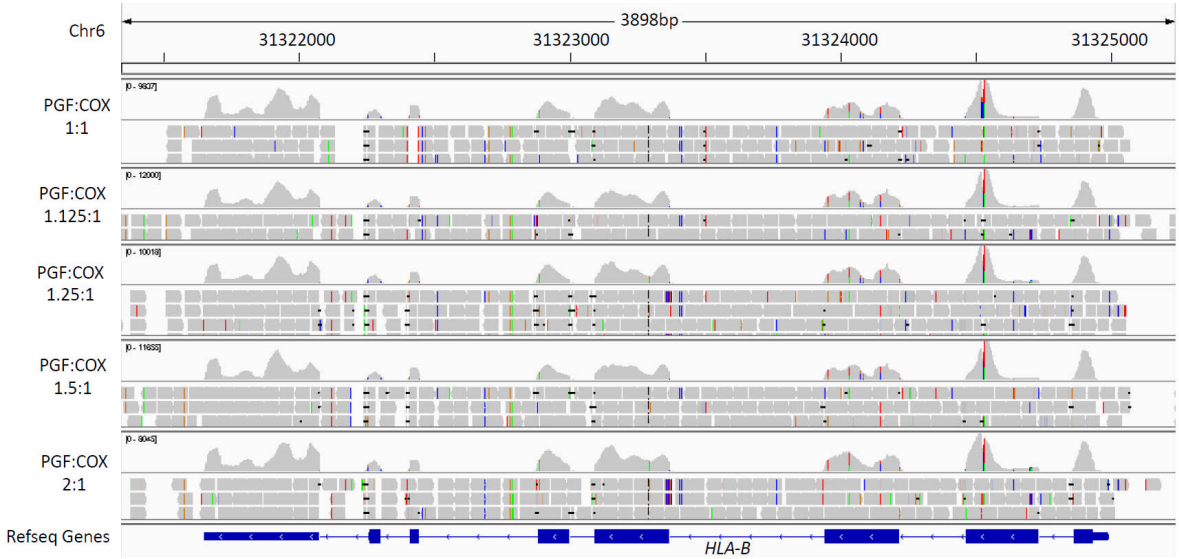


Figure 6.16: mRNA-seq coverage over *HLA-B*. *HLA-B* coverage is shown in mRNA-seq data for five mixed COX and PGF RNA samples mapped to the PGF reference sequence. 13 raSNPs that should be created by mixing PGF and COX RNA are identified in all five samples, and another two raSNPs are not called in all five mixed samples, but have read counts for both alleles. However, the final three raSNPs are not identified at all.

Remapping the MHC using individual reference sequences, either COX or PGF, rather than the composite of all eight sequences used in Hg19 led to detection of more SNPs in several genes, however these were not necessarily predicted raSNPs created by mixing PGF and COX, particularly when the COX sequence was used to map the data. In general, more SNPs were identified when only the PGF sequence was used to map reads. Another MHC gene, *HCP5*, was analysed to see if known raSNPs between COX and PGF could be identified in a less polymorphic gene using COX to map the data as well. A known raSNP should be created at chr6:31433558 and this was determined in the mixed ratio samples mapped to PGF, but not those mapped to COX. In general, mapping quality was relatively low when the COX sequence was used as can be seen in Figure 6.17, but was much improved when the PGF sequence was used to map the data (see Figure 6.16).

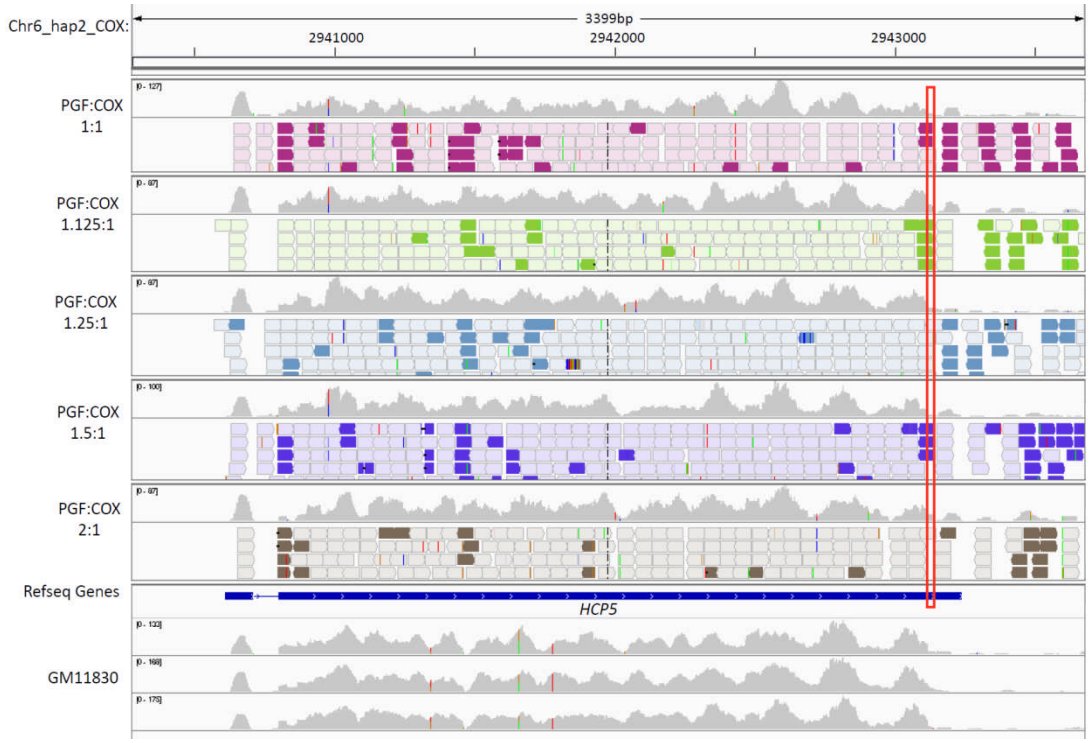


Figure 6.17: mRNA-seq coverage over *HCP5* in mixed COX and PGF samples and GM11830. Mixed COX and PGF RNA samples mRNA-seq coverage mapped to the COX reference sequence only. Reads that have an acceptable mapping score are indicated by blocked colour, with lighter coloured shading showing poor mapping quality reads. The known raSNP between COX and PGF is not identified in any of the mixed samples (predicted position outlined in a red box). Two heterozygous SNPs are found consistently in all biological replicates of the LCL GM11830.

Problems using the COX sequence to map data may reflect duplications and repetitive sequences found in the MHC. When areas mapped using the COX sequence with particularly low mapping quality are checked using a BLAT search, several alternate locations with high sequence similarity (>80%) are identified. However these occur equally throughout all possible haplotypes over the MHC and so it is unclear why mapping quality should be so much lower when COX is used compared to PGF. This may reflect an unrecognised mapping bias with respect to the MHC.

The LCL GM11830 had two heterozygous SNPs in *HCP5* identified by mRNA-seq, rs2255221 and rs2263318, which were found in all three biological replicates mapped both to PGF and COX sequences. High variability between the ratios was seen in both examples, which could be a result of greater difficulty in mapping in the MHC, or due to the relatively low read counts, particularly in the data mapped using the COX sequence. However, the ratios for each SNP were very similar between

both data sets. Additionally, another heterozygous SNP defined in the gene *HLA-DRB1* in GM11830 (rs17880292) had much larger read counts when mapped using either the COX or PGF sequence and was far more variable than the replicate ratios seen in the *GAPDH* gene, with a similar magnitude of counts (see Table 6.3). This SNP did not have similar ratios between the results from the differently mapped data, implying that mapping variability may have a large effect on accuracy in the MHC. Read counts were much lower when mapped to the COX sequence compared with the PGF sequence for the SNPs in *HCP5*, but this was not the case for the SNP in *HLA-DRB1* (see Table 6.3). This could indicate that *HLA-DRB1* is a location that is equally well mapped using both the COX and PGF sequence, in comparison to the *HCP5* gene.

		Replicate 1	Replicate 2	Replicate 3	Average ratio	Standard deviation
rs2255221	G (mapped using PGF)	73	102	91		
	T (mapped using PGF)	61	107	91	**1.05**	**0.13**
	G (mapped using COX)	10	25	39		
	T (mapped using COX)	16	31	22	**1.07**	**0.62**
rs2263318	A (mapped using PGF)	99	116	100		
	G (mapped using PGF)	71	114	90	**1.17**	**0.20**
	A (mapped using COX	42	46	39		
	G (mapped using COX)	49	34	30	**1.17**	**0.27**
rs17880292	C (mapped using PGF)	1769	1979	1899		
	T (mapped using PGF)	637	976	981	**2.25**	**0.46**
	C (mapped using COX)	1777	1927	1796		
	T (mapped using COX)	1246	1671	1607	**1.23**	**0.17**

Table 6.3: Heterozygous SNPs in GM11830 and their read counts per allele. Read counts for alternate alleles at rs2255221, rs2263318 (both found in HCP5) and rs17880292 (found in HLA-DRB1) are compared for the three biological replicates of GM11830 mapped using a single reference sequence, either COX or PGF. The average ratios and standard deviation between the detected alternate alleles are indicated for this data.

Finally, *ZFP57* was investigated as it was a far less polymorphic gene, but still located in the MHC. From the data presented on the raSNPs generated outside of the MHC, it was clear that for genes with very low expression, picking up variants can be challenging. The full sequence of *ZFP57* in both PGF and COX is known due to the sequencing of the MHC by the MHC haplotype project (Stewart 2004). When looking at expression of *ZFP57* in the LCLs separately it was possible to determine low but detectable expression of *ZFP57* in the COX LCL as expected from the findings of the MHC custom

array analysis (Vandiedonck 2011) and qPCR analysis of *ZFP57* expression in LCLs (see Chapter 3). In contrast, very little expression was detected in the PGF LCL, again as predicted by the MHC custom array and qPCR data. Two known raSNPs that differ between the PGF and COX LCLs (rs2535238 and rs2747421) are located over the beginning of exon 2 of *ZFP57*. The 5' end of *ZFP57* is significantly less expressed than the 3' end, shown using qPCR (see Chapter 3), and agreeing with this observation there is no coverage of either SNP in the mRNA-seq data for either LCL. However, RNA expression is seen over the full length of *ZFP57* in the COX LCL, beyond the Refseq annotation of the gene, and including the first exon (shown in Figure 6.18).

When looking at the mixed ratios of the LCLs, there is so little expression of *ZFP57* that it was not possible to investigate any relationship between expression and the different mixed ratio samples. Interestingly, in the LCL with HLA types similar to PGF and COX, GM11831, there appears to be approximately half of the expression seen over *ZFP57* in the COX LCL (Figure 6.18). When individual sequences over the MHC are used to re-map the data, more reads are aligned over *ZFP57*, however the numbers are still very low (an average coverage between 4 and 14 reads at any point). Several variants are identified that are not reproducible throughout the five mixed samples, or the individual RNA for COX or PGF, implying that accuracy of these calls is not high.

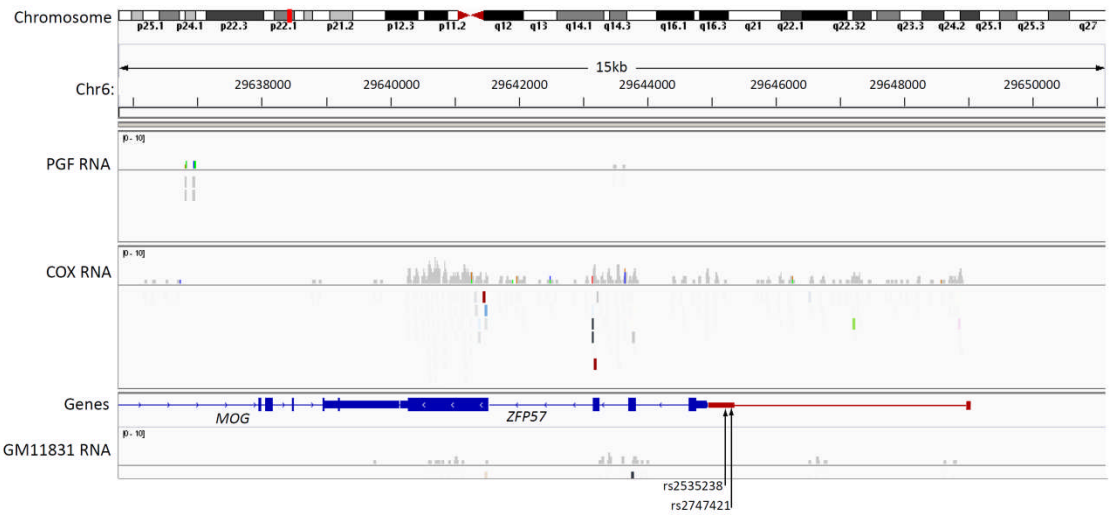


Figure 6.18: mRNA-seq in separate LCLs over *ZFP57*. Coverage of reads over *ZFP57* is visualised using IGV. The refseq isoform of *ZFP57* is shown in blue, while the full extended transcript is indicated in red. Expression is seen in the COX LCL but not PGF and in the LCL GM11831 very low expression is detected.

6.3.7 Reference-free mapping in the MHC

Because the MHC region was impossible to analyse effectively using the reads mapped with Stampy, a reference-free mapping method was attempted for areas in the MHC based on algorithms using coloured De Bruijn graphs. The data was mapped by Zamin Iqbal using the Cortex software, adapted from the method outlined in his recent publication (Iqbal 2012). This method of mapping uses a sequencing read as a "word", which is then joined with other "words" that overlap at its edges. In this way a sequence is built up. When a variant occurs, it forms a new edge that allows it to be identified; in the simplest form, a SNP will occur as a bubble in an otherwise homologous contig. While a reference sequence is not necessary for this method, it can enhance results; for example, a reference sequence allows the detection of homozygous alterations, which would otherwise not be highlighted as variants.

HLA-B was investigated using this technique as it had several known exonic raSNPs created between PGF and COX, to see if it could in principle be used to analyse data. For this purpose, the main focus was to check that using this method to define variants led to an association between the observed and expected ratios of raSNPs in the gene. Because of the different ratios seen across the *GAPDH* gene when using traditional mapping methods, one exon of *HLA-B* where variants were defined, exon three, was chosen to analyse the ratios in all five samples. In order to assess COX or PGF specific RNA, reads were only kept in the analysis if they could unambiguously be defined as either COX or PGF. This allowed LCL specific coverage to be defined over the exon.

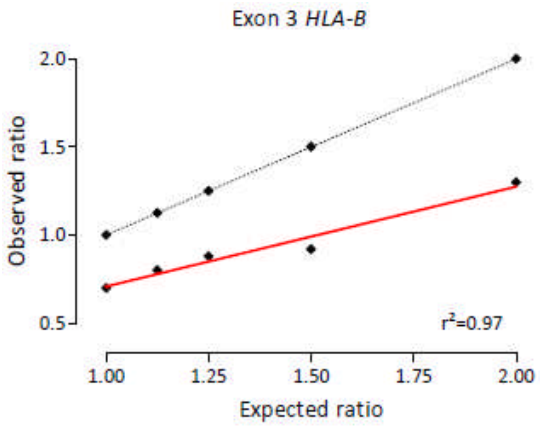


Figure 6.19: Observed and expected ratios of expression of alternate raSNPs in exon 3 of *HLA-B*. Coverage over exon 3 in *HLA-B* was defined as "PGF" or "COX" using sequence variation mapped by Cortex to separate reads that covered this region in each of the five mixed RNA samples. A significant association was seen between the observed and expected ratios of coverage over the exon (p=0.0027). The black dotted line indicates the expected line if the observed were to exactly match the expected ratio, while the red line indicates the observed ratios.

As yet, not all the known variants in *HLA-B* have been identified using the reference-free mapping, however quality of mapping scores will have led to some true variants being filtered out. To put this in context, using reference-free mapping to define all raSNPs between COX and PGF beyond the MHC has identified only half of the raSNPs analysed in this chapter found outside of the MHC (raSNPs found in *NADK*, *POLR2A* and *NOTCH1*). Despite these caveats it is likely that further work using this method will allow more accurate determination of variants in the MHC than traditional mappers, especially when considering samples where the MHC sequence is unknown.

6.4 Discussion

The introduction of NGS technology to the study of gene expression has allowed the possibility of detecting novel transcripts routinely, as well as sensitive detection of low expressed transcripts. When Illumina RNA-seq has been compared to gene expression detection using Affymetrix arrays previously, it has been shown that a single lane of sequencing is highly reproducible and gives comparable results to analysis using an array when considering annotated coding genes (Marioni 2008). A variety of different model organisms have had their transcriptomes analysed using RNA-seq and the approach is increasingly used in human studies. However, when setting up this study it was apparent that many controversies surrounded data analysis and interpretation, particularly for ASE.

6.4.1 Analysis of gene expression with RNA-seq

While RNA-seq has the potential to identify low and medium abundance transcripts, a suitable sequencing depth for their detection must be achieved (Mortazavi 2008). This means that when samples are multiplexed, a loss of deeply sequenced data will reduce the ability of RNA-seq to determine expression for some genes. While we did not multiplex the mixed ratios of RNA, or the single LCL RNA, it was clear that there were several genes where sufficient depth was not obtained to confidently analyse gene expression. Although *ZFP57* expression was detected in the COX LCL for example, there were very few reads and it was not possible to compare the expression across the five mixed PGF and COX samples. This suggests that greater read depth would be necessary to analyse genes with very low expression, and implies that for samples where genes with low expression are of significant interest multiplexing may not lead to suitable data. The use of RNA-seq with no further modification as a method suitable to determine differential expression has been questioned due to its difficulty in accurately defining rare transcripts. A recent study has suggested that as sequence depth increases, noise will also increase and so make analysis challenging; to this end they suggest use of their technique NOIseq to assess variability of read counts and improve prediction of differential expression (Tarazona 2011). The principle behind NOIseq aims to remove any bias associated with read count number in the identification of differentially expressed genes by

normalising the fold change of gene expression and the absolute change in gene expression for any genes analysed in comparable samples.

6.4.2 Accurate detection of ASE

ASE measurement has traditionally relied upon coding or exonic SNPs that allow accurate quantification of alternate alleles (Knight 2006). This remains the case for mRNA-seq, as regulatory noncoding variants are not expressed and thus cannot be used to differentiate transcripts. While use of coding SNPs is not a problem when a gene is expressed at a high level, when only a small number of reads are available for low expressed genes it can be hard to tell if variation is due to a sequencing error rather than ASE. Calling the presence of a SNP from RNA-seq data alone can be highly problematic where one allele may be expressed at a significantly lower level. The ability to use exome sequencing data from genomic DNA in these situations to identify sites where a variant should be evident would clearly be beneficial.

Heap and colleagues analysed ASE in primary T cells from four individuals in both resting and activated states. They identified 1371 unique transcripts with evidence of ASE, however around 50% of these were deemed to be due to SNP bias. They determined that 50 reads covering the SNPs in question would be required to detect a two-fold variation in gene expression where the significance level was 0.001 (Heap 2010). Because of this, it would take extremely deep sequencing to ensure that genes expressed at a low level were truly exhibiting ASE, rather than reflecting a sequence bias. This was apparent in our data, where genes with low expression were more difficult to analyse. This also reinforces the usefulness of sequence information, either via exome sequencing or genotyping and imputation, in conjunction with mRNA-seq for samples as stated previously.

In the analysis of our data we were able to find variants between the PGF and COX LCLs outside of the MHC using variants detected in the exome data that were homozygous in each LCL. This allowed the analysis of the ratios generated using the mixtures of COX and PGF RNA. In general where high read counts of alternate alleles were available, numbering over 50 of each variant, the observed ratios that were found matched closely to the expected ratios. If they were different to the expected ratios, they still followed the correct trend taking the 1:1 ratio as the starting point. In contrast,

where read counts were low the observed ratios were far more varied, highlighting the difficulty with low gene expression analysis.

In the case of analysing *ZFP57* expression using mRNA-seq reads mapped with Stampy, no expression over PGF could be detected, and no expression over known coding variants could be detected either. Therefore it was impossible to show using this gene that low expression could still be determined in an allele specific manner. Despite this, comparing the absolute read counts in the COX LCL with the GM11831 LCL (that contains HLA types analogous to both COX and PGF) showed that *ZFP57* expression was reduced in the GM11831 LCL compared to COX. This matched the prediction for *ZFP57* expression in GM11831 made based on the MHC array data.

6.4.3 Bias in RNA-seq experiments

Interestingly, a study analysing the effect of read mapping bias on ASE showed that masking known SNPs could remove any reference sequence bias. However, overall the accuracy of mapping did not improve, indicating that many SNPs confer a mapping bias towards one particular allele which in turn will cause a bias in the detection of ASE (Degner 2009). The presence of SNPs in any given read when compared with a reference sequence make it more likely that the read will be removed from mapping (Fang and Cui 2011). Although mapping algorithms can tolerate limited mismatches to their reference sequences, if many SNPs are present in the read it will be discarded. This presents particular problems for highly polymorphic areas of the genome such as the MHC. For example, data is mapped with low quality at many loci in the MHC (see Figure 6.17).

Mapping bias and detection of genes with low expression are not the only problems with analysis of RNA-seq data. The process of mRNA-seq assumes that transcripts will be sampled randomly creating short shotgun sequences. Longer transcripts in a genome will therefore have more reads than shorter ones and this can create a bias towards detection of longer transcripts (Mortazavi 2008). Two separate studies have shown that the longer the gene, the higher the read count, and this is not corrected by normalising read counts to gene lengths (Oshlack and Wakefield 2009, Bullard 2010). Differences in the sequence can also lead to bias in detection of expression via GC content changes, and the effect of random hexamer priming used in the Illumina sequencing method (Hansen 2010).

In their study of 69 Nigerian LCLs, Pickrell and colleagues noted that GC content confounds gene expression analysis between samples and corrected for this by binning exons according to GC content at 3' and 5' ends (Pickrell 2010). This again highlights the importance of sequence information in analysing the mRNA-seq data.

Alternative splicing can add another layer of complexity to the analysis of gene expression using mRNA-seq. Detection of alternative splicing using mRNA-seq has been reported by several studies, however again for reasons previously discussed moderate to high expression is required to ensure that results are not biased by either the sequence composition or by mapping of the reads. A 2008 study highlighted how mRNA-seq identified expression of 25% more genes than conventional array analysis, while also being able to define nearly 100,000 splice junctions (Sultan 2008). This study also showed that exon skipping was the most common form of alternative splicing. Studying the effect of alternative splicing between the two LCLs in our experiment was difficult due to relatively few genes where there were several different SNPs spread across a gene. In *GAPDH* the SNP found at the 5' end of the gene was not found in as many reads as the other SNPs in *GAPDH*, and also showed far less convincing association with the expected ratios of gene expression. This also implies that in some cases ASE may in fact be alternative splicing, where one allele appears to have lower expression, but in fact an excluded exon is responsible for any difference seen.

6.4.4 Analysing the MHC using RNA-seq

MHC variants and analysis of differential expression has not been extensively studied using mRNA-seq, therefore it is hard to compare analysis of this experiment to that of previously published studies. Using the mapper Stampy it was clear that many variants were not detected, even when the individual full MHC sequences of both PGF and COX were used to map the transcripts. The highly polymorphic nature of this region, and the large amount of alternative splicing and variable gene expression that is known to characterise the MHC made comparison of expression within the MHC impossible. In particular, variants identified within the MHC were often not found consistently throughout all five of the mixed ratio samples leading to no confidence in the accuracy of the call, for example in *HCP5*, where the known variant between COX and PGF is found when data is mapped

solely against the PGF sequence, but not when mapped using the COX sequence. Some raSNPs at separate locations are called in some of the mixed ratio samples when using either sequence to map the data. However, analysis of data mapped using the COX sequence alone seems to pick up variants less frequently.

Analysis of *HLA-B* demonstrated the wider phenomenon of reference sequences causing serious mapping bias in the MHC. raSNPs were relatively successfully identified when mapping was carried out using only the PGF sequence, but when using the COX sequence no raSNPs were identified across the gene at all. While it is comforting that using the PGF, or reference, sequence gives the most accurate results, several known raSNPs could still not be identified. It was concerning that if detailed sequence information was not known, or if the sample varied dramatically from the reference sequence then many variants would be missed. This also highlighted that mapping bias in the MHC has not been fully explored, as different results were found when COX was used as the sole reference sequence compared to the single use of the PGF sequence.

Because of this we also attempted reference-free mapping using *de novo* assembly (Iqbal 2012). This was predominantly to assess gene expression within the MHC; however, in the future it will be important to know how this technique would perform in other loci so that a universal method for analysis of mRNA-seq or total RNA-seq data that included the MHC could be pioneered. It is notable that currently no standard practice exists that encompasses the design, implementation and analysis of an RNA-seq experiment; yet as more RNA-seq data is generated the ability to compare data sets will become more important (Auer 2011).

Using *de novo* assembly to analyse the five mixed RNA samples allowed the reads that could be used to identify either COX or PGF sequences specifically to be kept for analysis of their ratios in the samples. All reads that could not be assigned to either COX or PGF alone once assembled were removed from the analysis. When *HLA-B* was studied using this technique raSNPs between the COX and PGF RNA could be determined. A pilot study in exon three of *HLA-B* showed promising results as a significant association between the observed and expected ratios was found, but it was notable that when looking over the whole gene instead this association disappeared. This highlights the

importance of taking alternative isoforms and bias in coverage of genes into account in gene expression analysis. It is encouraging that *HLA-B* expression may be able to be determined in this manner, as it is a polymorphic gene that is not easily studied using array or qPCR based methodologies. For genes such as this, it is of great importance to establish a reliable technique for mapping expression data. It is clear however that there is much further work necessary, beyond the scope of this thesis, to identify any confounding factors in the analysis of the MHC using reference free mapping and to ensure that ASE can be confidently assessed using RNA-seq.

It is unclear whether the most effective method for analysing RNA-seq data over the MHC will be to use a single technique to map reads and call variants or whether a combination of analysis using traditional mapping to a reference as well as reference free mapping will give more accurate and consistent results. More work, especially in the use of a single reference sequence compared with the composite reference sequence over chromosome 6, is required to see whether the increased variation of results in the MHC can be reduced. Finally, further examples beyond the pilot study of *HLA-B* using reference free mapping are necessary to confirm that the promising results seen so far are reproducible and applicable in a biologically relevant setting where full sequence information and known ratios of variants are not available.

6.4.5 Conclusion

This study confirms that mRNA-seq can consistently define ASE when expression of a gene is high. Where sequence information for a gene is known, this enables detection of ASE with high specificity. However the MHC remains a particular challenge in both the initial mapping of reads and the interpretation of gene expression data due to sequence and structural variants. While the promise of mRNA-seq is great, further work is required to ensure that all data generated can be accurately interpreted without bias of any sort and to minimise the data analysis pipeline to allow it to be more easily introduced in larger scale gene expression investigation.

Chapter 7 – Characterisation of genome-wide protein-DNA binding by ZFP57

7.1 Introduction

As highlighted in Chapter 3, the biology of ZFP57 is still relatively unknown. While thought to be important in the maintenance of methylation patterns and to act as a TF, no known binding sites or binding motif has been established in humans, and no functional work outside its role in development has been reported. Previous studies have identified aberrant methylation in neonatal T1D patients with *ZFP57* mutations (Mackay 2008); and the function of ZFP57 has been studied in mice embryos indicating that ZFP57 plays an essential role in development by maintaining maternal and paternal methylation patterns (Li 2008). With this is mind this chapter investigates the role of ZFP57 in LCLs, focusing on identifying sites of protein-DNA interaction across the genome. The LCLs chosen for this work were the PGF and COX LCLs described in Chapters 3 and 4, where evidence was presented that COX expresses *ZFP57* at a much higher level than PGF, which was included as a negative or low expressing control. A further LCL, APD, was also included as it appeared to have the high levels of ZFP57 on western blot analysis (Figure 4.21). The primary methodology chosen for this work was chromatin immunoprecipitation, an important technique for analysing protein-DNA interactions in cells (Massie and Mills 2009) which can be analysed using high-throughput sequencing using ChIP-seq, an approach discussed further in this introduction.

7.1.1 ChIP-seq methodology

To carry out TF ChIP, an antibody raised against the protein of interest is used to immunoprecipitate cross-linked sonicated-chromatin that is bound to the TF. The selected chromatin can then be used to analyse the location of binding of the TF, and in the case of ChIP-seq this is achieved by sequencing the immunoprecipitated chromatin. The technique is described in more detail in Figure 7.1. The benefit of ChIP-seq is that theoretically genome-wide coverage of TF binding can be achieved (Park 2009). ChIP-seq has also improved the study of histone modifications that are crucial in the understanding of epigenetic mechanisms. By using an antibody raised against a particular histone

modification, cross-linked chromatin with that modification can be immunoprecipitated and sequenced in the same way as TF ChIP. The resolution of ChIP-seq at a genome-wide level gives it an advantage over other methodologies such as ChIP-chip meaning it is more likely that unknown binding motifs of TFs can be identified by comparing the binding sites identified across the genome. The amount of material required is also typically much smaller for ChIP-seq experiments than ChIP-chip, making it a more realistic possibility for TFs that are not expressed highly.

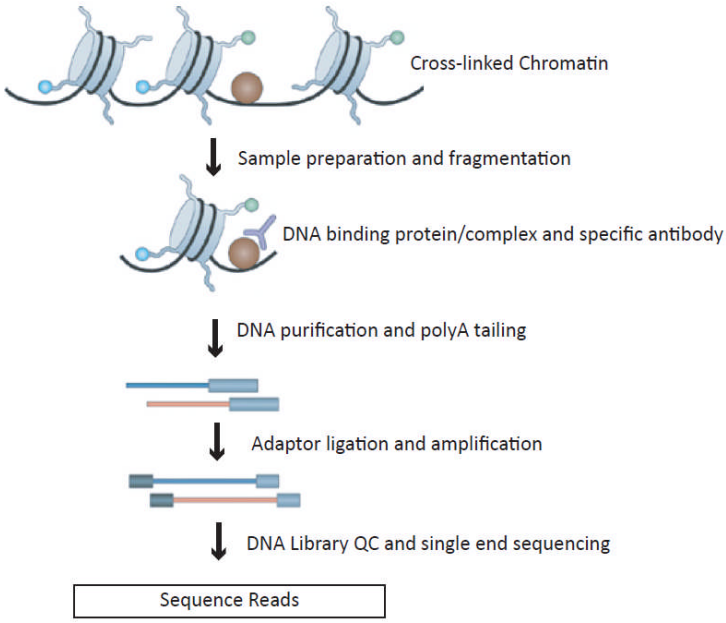


Figure 7.1: Schematic showing the methodology behind ChIP-seq experiments. Adapted from (Park 2009). Cells are cross-linked to stabilise protein-DNA interactions as described in Materials and Methods, Section 2.8. Cross-linked chromatin is extracted from the cells and is sonicated to produce fragments of around 200-1000bp in length. The chromatin is incubated with an antibody specific to the protein of interest and this is used to isolate the DNA that is binding this protein or protein complex. Isolation of the antibody-bound DNA can be carried out by various methods, for example using protein-coated magnetic beads. DNA:protein cross-links are reversed and the protein is digested to leave only the DNA of interest. This is then end-repaired, polyA-tailed, adaptor-ligated and amplified to form a DNA library of the bound DNA. This procedure is also carried out for an input sample of chromatin that is either treated with no antibody or using a mock/ non-DNA binding protein antibody to use in analysis. QC is carried out to ensure that the library preparation has been successful. Sequencing is performed using an appropriate platform and the short sequence reads are aligned to an appropriate reference sequence.

However, there are disadvantages to the use of ChIP-seq. Primarily, the cost of the experiment is high compared with that of ChIP-chip, and if genome-wide information is not required, the use of a customised microarray may be as informative (Park 2009). The volume of data generated by ChIP-

seq is extremely large, and requires specialist analysis; additionally, sequence alignment may cause problems in the data analysis, particularly where the genome sequence of the sample being studied is unknown. The quality of the antibody used affects any ChIP experiment, therefore it is important to be confident in the ability of the selected antibody to be used in immunoprecipitation experiments. As only one antibody was available for ZFP57 and this had not been tested in an immunoprecipitation experiment, a positive control that would give additional useful information was investigated.

7.1.2 KAP1 as an additional control for ZFP57 ChIP-seq

KAP1 has been extensively studied previously due to its role in development, including cellular proliferation and neoplastic transformation. It is a highly conserved protein known to have many different roles in cellular processes. KAP1 is critical in development; mice that are deficient in KAP1 die before reaching gastrulation (Cammas 2000). It also is important in maintenance of pluripotency, and in control of various types of cell differentiation. For example, KAP1 inhibits red blood cell differentiation, yet promotes the differentiation of U937 cells into macrophages (Iyengar and Farnham 2011). KAP1 has been implicated as a contributor to various forms of cancer, including hepatic, lung, breast and gastric cancers, as its expression is found to be elevated in these cancers. Gastric cancer patients in particular are shown to have a lower survival rate when KAP1 protein levels are elevated (Yokoe 2010), and reduction in KAP1 levels is thought to improve the p53 response and decrease proliferation following chemotherapy (Okamoto 2006).

KAP1 does not itself bind DNA, instead it forms a scaffold for several other proteins involved in chromatin remodelling and is recruited to DNA binding complexes through various partners, many of which are KRAB domain containing zinc finger proteins. The KRAB-ZNF protein family expanded dramatically in primate evolution as segmental duplication events enabled a much wider protein family to develop (Emerson and Thomas 2009, Nowick 2010). ZFP57 is a KRAB-ZNF protein and would be a highly plausible binding partner for KAP1. Of particular relevance when considered in conjunction with ZFP57 is the well reported role of KAP1 in transcriptional silencing and spreading of heterochromatin (Groner 2010). ZFP57 itself, as previously discussed in Chapter 3, has been

implicated in methylation pattern maintenance (Li 2008). It is therefore a reasonable hypothesis that KAP1 interacts with ZFP57 and overlap in binding may be observed giving more confidence that a genuine ZFP57 binding site has been detected.

Most significant reported KAP1 binding sites are in 3' coding exons of ZNF genes and involve interaction of the N-terminal domain of KAP1 (which includes a RING finger type zinc finger motif, two B box zinc finger domains, and a leucine zipper coiled-coil region) interacting with multiple zinc fingers of the C-terminal regions of KRAB-ZNF proteins (Iyengar 2011). Typically expression of KRAB-ZNF genes is not regulated by KAP1 and function of them in combination is instead linked with depositing H3K9me3 and maintenance of heterochromatin although it is unclear as to whether this may then play a role in regulation. The number of repeated zinc finger domains in 3' exons of ZNF genes is correlated with KAP1 binding and the presence of the H3K9me3 mark. It is suggested that the role of the H3K9me3 mark is primarily to prevent inappropriate recombination or DNA repair in the area of the genome (Blahnik 2011). Where KAP1 is involved in a transcriptional regulatory role, it is not thought to include ZNF proteins as binding partners, and is thought to be regulated by post-translational modification (Iyengar and Farnham 2011).

An initial study using ChIP-chip to investigate KAP1 binding in Ntera2 cells (neuronal progenitor like cells) showed that H3K9me3 and KAP1 was found to be associated with 3' exons of ZNF genes only, in the case of other genes KAP1 was seen to bind more to the promoter region (O'Geen 2007). It was therefore assumed that KAP1 with the ZNF protein contributed to auto-regulation of the gene. This seemed probable as reporter assays showed KAP1 could act as a transcriptional repressor however no evidence was found for differential expression associated with the amount of KAP1 or histone modification in the 3' end of the ZNF genes *in vivo* (Blahnik 2011). KAP1 has been shown to associate with a KRAB-ZNF protein ZNF274 and ChIP-seq has been used to show that binding sites of these proteins in K562 (monocyte like), HEK (Human Embryonic Kidney), GM12878 (LCL) and HepG2 (human hepatocellular carcinoma) cells, and additional proteins such as a methyltransferase STEDB1, co-localise over the 3' ends of ZNF genes (Frietze 2010). This study showed specific locations where KAP1 was recruited with ZNF274 and when ZNF274 was knocked down using RNAi, binding of KAP1

was also decreased in these regions. ChIP-seq assays have additionally been performed on U2OS (an osteosarcoma cell line).

In view of its association with KRAB-ZNF proteins as a binding partner, an experimental design was developed in which ChIP-seq would be carried out for KAP1 binding as well as ZFP57 binding in the three LCLs. This would allow the comparison of any binding peaks between KAP1 and ZFP57 as likely binding partners. KAP1 ChIP-seq would also act as an effective positive control, as several studies have already successfully described KAP1 binding in a genome-wide manner, either by ChIP-seq studies or using a ChIP-chip approach as detailed above.

7.2 Aims

The overall objective of this chapter was to define genomic sites of protein-DNA binding by ZFP57. Specific aims were:

1. To determine ZFP57 DNA binding sites by ChIP-seq in high and low expressing LCLs
2. To analyse KAP1 binding in the same LCLs by ChIP-seq as a potential binding partner for ZFP57 and to assess if there were allele-specific binding events between the 3 LCLs used in the analysis.

7.3 Results

7.3.1 Selection of antibodies for ZFP57 and KAP1 ChIP

As previously stated, the only publicly available antibody for human ZFP57 had not been previously used in ChIP or IP experiments in any published literature. As it gave the correct sized band when used in western blot analysis on the LCLs (see Figure 4.21) it was decided to test the functionality of the antibody for immunoprecipitation. To do this, *ZFP57* was amplified from COX cDNA and cloned into a Myc-tagged vector to produce a vector that would express Myc-tagged ZFP57 with the Myc-tag being found as the N-terminus of the protein (Figure 7.2).

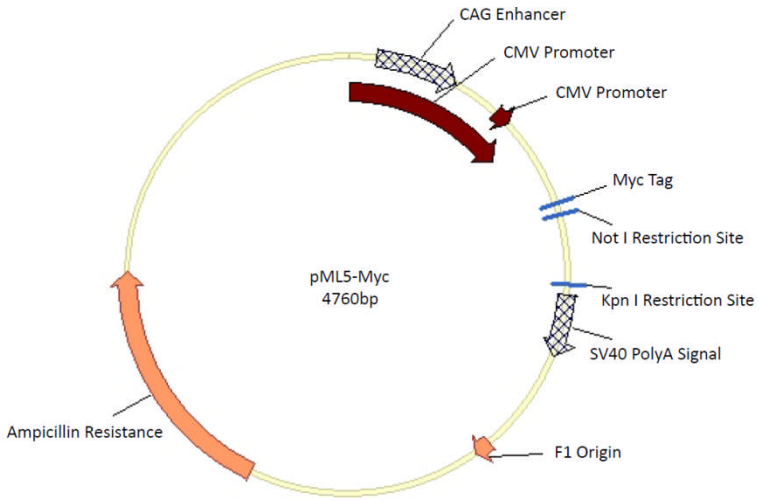


Figure 7.2: Cloning *ZFP57* into pML5-Myc. The pML5-Myc vector was digested using the single cutting restriction enzymes NotI and KpnI. The vector contains an Ampicillin resistance gene to allow selective growth; an F1 origin to allow replication and different sequences such as a CAG enhancer, CMV promoter and SV40 PolyA Signal to ensure expression of the gene that is cloned into the vector. *ZFP57* was amplified using primers with restriction sites for NotI and KpnI added so that it could be introduced into the vector, in frame with the Myc-tag to allow expression of a Myc-tagged ZFP57 fusion protein.

The presence of *ZFP57* in the vector was confirmed by sequencing and no coding mutations were introduced into the *ZFP57* sequence. The ZFP57-Myc fusion protein was then over expressed in the HEK cell line and cell lysates were harvested. An immunoprecipitation using the anti-ZFP57 antibody with subsequent western blot against the Myc-tag showed expression of the fusion protein was successful in the cell lysate; however, the protein was not pulled down in the immunoprecipitation. It is possible that the Myc-tag interferes with the ability of the antibody to bind to ZFP57 or that this reflects the efficacy of the antibody in a cellular context.

ZFP57 has not been extensively studied, as detailed before in Chapter 3, and no additional α-ZFP57 antibodies that detected human ZFP57 were available. Although the immunoprecipitation of the ZFP57-Myc-tag fusion protein was unsuccessful, the previous WB and FACS data suggested that the antibody did specifically bind to ZFP57 and could detect native ZFP57 protein. Therefore, it was thought to be worth proceeding with the ZFP57 ChIP-seq experiment.

The specificity of the KAP1 antibody was first tested using western blot analysis against APD, COX and PGF cell lysates and nuclear lysates (Figure 7.3). This showed that the KAP1 protein was expressed consistently at high levels across all LCLs. The KAP1 antibody was a ChIP grade antibody and had

previously been used in studies where specific sites of KAP1 binding as well as genome-wide binding of KAP1 were assessed (O'Geen 2007, Rambaud 2009).

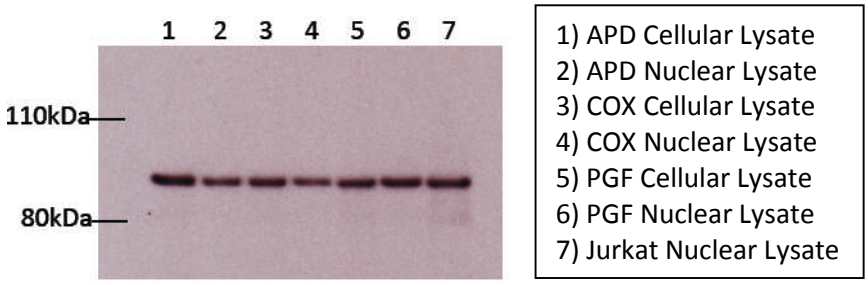


Figure 7.3: KAP1 expression in three MHC homozygous LCLs confirmed by Western Blot. Cellular and nuclear lysates are assayed for KAP1 expression using the ChIP-grade KAP1 antibody. Jurkat nuclear lysate is included as an additional positive control. In all cases, KAP1 expression is clearly seen in the cell lines and one band is seen at approximately 100kDa.

The western blot analysis shows a single band of the correct size for KAP1 (100kDa). This confirmed that the KAP1 antibody was working correctly and as the previously mentioned publications used the antibody successfully for ChIP experiments it was selected to be used as the positive control antibody.

7.3.2 Sample preparation and ChIP

The COX, PGF and APD LCLs were harvested in a resting state in mid log-phase and subject to formaldehyde cross-linking for 15 minutes before cell lysis and extraction of chromatin. Sonication using the Bioruptor was optimised to 18 cycles for 30 seconds on, 30 seconds off at high amplitude to generate fragment lengths of between 100 and 1000bp, at which point crosslinks were reversed and DNA purified. Samples were kept at a constant 4˚C by a water cooling-system. Sonication fragments are shown in Figure A.9.

ChIP was then performed using sonicated chromatin for the three LCLs. KAP1 and ZFP57 antibodies were bound to purified sheep anti-rabbit polyclonal antibody coated magnetic beads and no-antibody control (beads only) was also used. Following overnight incubation, all protein crosslinks were reversed and DNA was purified.

7.3.3 Validation of ChIP using two known KAP1 binding sites

Two previously identified KAP1 binding sites from a study looking at KAP1 and ZNF protein mediation

of heterochromatin spreading (Groner 2010) were selected to use as positive controls for ChIP before submission of samples for sequencing. The study had used qPCR to analyse ChIP experiments. Two separate regions of binding, over ZNF genes, were replicated here using ChIP material generated for ChIP-seq from the three LCLs. The input chromatin and a primer set directed against the 18S DNA sequence were used to normalise the qPCR data in order for enrichment to be detected. The loci chosen were *ZNF554* and *ZNF556*, shown to have KAP1 binding in a study by Groner and colleagues.

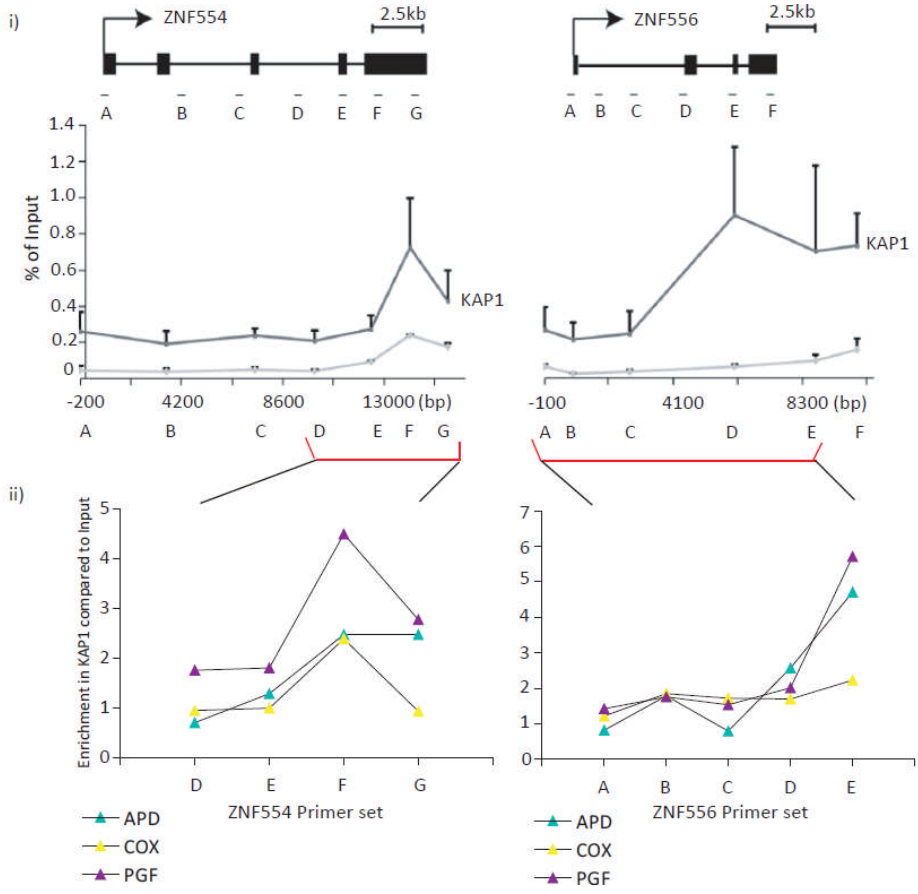


Figure 7.4: Confirmation of success of KAP1 ChIP. i) qPCR showing enrichment in the ChIP material of the sample incubated with KAP1 compared to input at 2 separate ZNF genes on Chromosome 19. The darker line shows KAP1 binding under normal conditions; the lighter line shows KAP1 binding under RNAi against KAP1 which caused an overall reduction of KAP1 in the cells. Redrawn from (Groner 2010). ii) KAP1 binding compared to input in the 3 MHC homozygous LCLs at the same locations. Primer sets used are indicated by the red line highlighting which were chosen from the primers used in the Groner study, and their sequences are detailed in Table A.5.

Figure 7.4 shows the results using ChIP material from the three homozygous LCLs. The same enrichment pattern is seen for all three LCLs consistent with that demonstrated in the original

publication, demonstrating that the ChIP against KAP1 had been successful. Following the confirmation that the ChIP experiment worked as expected with regard to KAP1, the samples for ChIP-seq analysis of both ZFP57 and KAP1 in the three LCLs were submitted for sequencing at the Core Genomics facility (WTCHG).

7.3.4 Sequencing and ChIP-seq data analysis

ChIP material was submitted for sequencing. Libraries were constructed from the fragmented extracted chromatin and size selection was carried out to remove any sequences that were too large to be successfully sequenced. Amplification was performed following the size selection and the library was also quality checked. Sequencing was then performed using the GAIIx Illumina sequencer. All data analysis algorithms were performed by Dr Stephen Sansom and Dr Andreas Heger of the Department of Physiology, Anatomy and Genetics in Oxford as part of a collaborative project with CGAT (Computational Genomics Analysis and Training).

7.3.5 Sequence alignment

The tag sequences were aligned to a reference sequence using Bowtie (http://bowtie-bio.sourceforge.net/index/shtml). Bowtie is a read aligner designed to efficiently map large numbers of short reads to a larger genome sequence (Langmead 2009). Parameters used for alignment with Bowtie meant that only sequences with two or fewer mismatches from the reference sequence would be retained, and that the best possible alignment was picked for each sequence tag. Only reads which were uniquely mapped were reported.

7.3.6 Filtering data

Regions that were abnormal and may be contamination were identified and removed from further analysis. As small amounts of DNA are typically used in ChIP-seq experiments, cross-sample contamination is a well known problem, particularly regarding post-PCR contamination of the pre-PCR ChIP sample (Barski and Zhao 2009). Four of these regions were identified, which had abnormally high copy number in the number of reads present and also showed exonic structure rather than smooth peaks. They were identified as probable contaminants from cloning reactions

carried out in the laboratory. This was confirmed by the presence of a silent point mutation known to have been included in the cloning of *ZFP57* previously in the lab. Other filtered regions included the *TNFRSF1B* promoter and the *PRDM9* gene.

7.3.7 Normalisation

Following filtering, any reads that were duplicated were removed, leaving only unique reads that had been mapped to one genomic location only (see Table 7.1). This also helps to remove contamination by post-PCR products as these will have several identical copies. Normalisation was carried out across the samples so that the overall number of reads considered for each sample was the same (15.8 million). This involved randomly selecting reads to be removed to render all samples equal in terms of read number. The PGF sample for ZFP57 had a much lower read count than either COX or APD (8.7 million), and so it was considered separately to prevent unnecessarily depleting the data for COX and APD ZFP57 ChIP samples.

	Total Reads	Mapped	Duplicates	Uniquely mapped	After normalisation
APD-ZFP57	40,768,692	30,477,027	12,339,602	18,137,425	15,818,488
APD-KAP1	39,240,124	30,799,784	14,981,296	15,818,488	15,818,488
APD-input	38,367,696	31,290,082	2,529,948	28,760,134	15,818,488
COX-ZFP57	37,463,320	28,697,089	12,577,374	16,119,715	15,818,488
COX-KAP1	37,784,931	28,889,693	4,346,803	24,542,890	15,818,488
COX-input	38,406,507	29,394,479	1,396,462	27,998,017	15,818,488
PGF-ZFP57	37,849,947	26,805,018	18,039,513	8,765,506	8,765,506
PGF-KAP1	39,689,058	30,390,379	3,785,090	3,785,090	15,818,488
PGF-input	39,041,305	30,834,061	1,728,380	1,728,380	15,818,488

Table 7.1: Read numbers following normalisation. All samples' read numbers are shown at each stage of the normalisation process. Following removal of duplicates and those reads which do not map uniquely in the reference, the number of reads remaining is normalised to the lowest (in this case APD-KAP1) by removing reads randomly across the sample. PGF-ZFP57 reads are not normalised due to their low number to avoid data loss in the other samples.

7.3.8 Peak calling

There are several different algorithms that can be used to call peaks in ChIP-seq data. They all primarily work in the same way, looking for a build up of unique reads over one location. A peak model is made for each sample by the program, allowing all data to be scanned for peaks which fit

the predicted model. Parameters for each peak calling algorithm can be varied to stringent settings, where confidence is high that when a significant peak is called it will be a real result, and to more relaxed parameters. The more relaxed results can then be further analysed by eye. This helps in situations where low overall binding is predicted, or there is very low protein expression. Peaks in the data are called when there is a significant increase in the sample at a location in comparison to the input control. It is therefore very important that the input is processed as the antibody incubated samples so that it is an adequate control. This has been highlighted in several publications, as the distribution of the input ChIP reads is rarely uniform in nature (Kharchenko 2008).

The MACS peak caller is the most commonly used peak finder in the analysis of ChIP-seq data (http://liulab.dfci.harvard.edu/MACS/). It can be used with or without control data, but the use of an input control reduces the false discovery rate (FDR) to 0.4% typically (Zhang 2008). Settings for peak discovery in this experiment were set to reflect the experimental material: a bandwidth of 500bp was used as the expected size of sonicated fragments was 1000bp; the tag size was set to be 51bp as this was the length of the generated fragments used in sequencing. The p value for a significant peak was set to be $1x10^{-5}$, and fold enrichment for a peak to be called was set between 20 and 100.

The SPP pipeline is designed to be run as an R software package (http://compbio.med.harvard.edu/Supplements/ChIP-seq/). A FDR of 1% was used, and peaks were detected using a series of windows to scan along chromosomes looking for abnormal density in reads, a window tag density (WTD) method (Kharchenko 2008). The window is defined by the characteristics of the data to be analysed, so in this case the window was set to be between 50-1000bp. Anomalous tag sequences were automatically removed by the pipeline.

PeakSeq differs from the first two peak callers in that it uses a two-stage approach to define peaks, filtering out any signal that is the result of open chromatin before comparing the ChIP sample to the input reads to define significantly enriched regions (http://info.gersteinlab.org/PeakSeq). (Rozowsky 2009). The enrichment fragment length used for analysis was 500bp, with a minimum inter-peak distance set at 200bp. A FDR was used at 0.05, with a maximum Q value of 0.05. Although PeakSeq

does allow tag sequences that map to multiple locations to be considered in the analysis pipeline this was not used as multiple mapped reads had been previously discarded from the data.

7.3.9 Peak calls from all algorithms

The peak model determined for all samples was similar showing that the analysis was consistent between the samples and the data did not vary abnormally. The peak model is designated "d" and accounts for both the size of the sonication fragments in the sample and the size of the peak that can be identified. As the peaks are off-set depending on the sense of the strand of DNA two separate peaks are determined, one on each strand, which are then merged to form the peak model, as shown in Figure 7.5. An example of the peak model is shown for the MACS algorithm for all three LCLs:

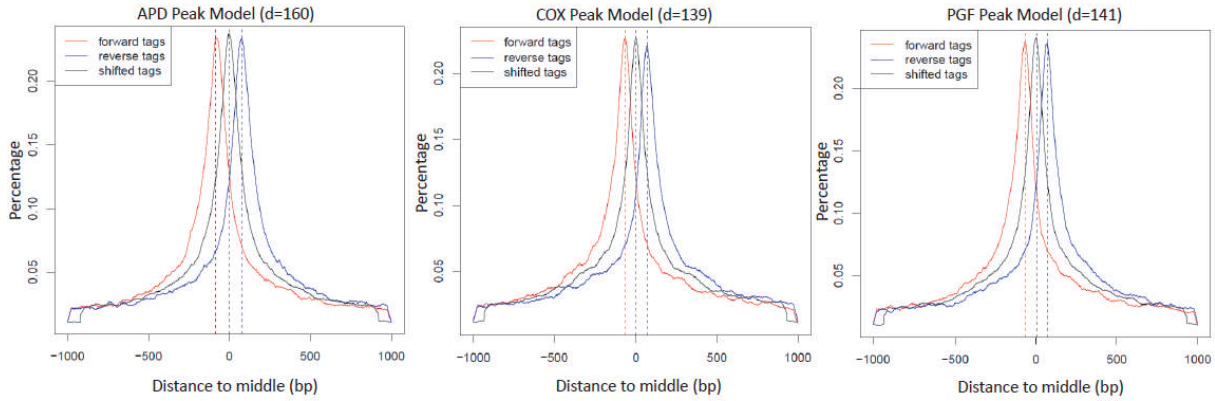


Figure 7.5: Peak model for data analysis using MACS (Figure courtesy of Dr Stephen Sansom). Peaks are shown to be consistent between the three different LCLs ChIP results. The peak model "d" does not vary considerably showing consistent sonication and peak size called in the data between the cell lines.

All algorithms detected significant peaks in the KAP1 ChIP data, but very few in the ZFP57 ChIP data when compared to the input (Table 7.2). Although different total numbers of KAP1 peaks were detected by the 3 peak callers, the results were broadly comparable in terms of trends. Relaxed parameters lead to an increase in peaks detected for ZFP57 however these peaks looked no different than what could reasonably be expected by chance. The PGF ZFP57 sample cannot easily be compared to the ZFP57 ChIP samples for APD and COX, as it had not been normalised so the number

comparisons are not relevant, however binding sites were predicted in the PGF ZFP57 sample, which bearing in mind the low expression of ZFP57 in the cell line makes the results seem untrustworthy.

		APD		COX		PGF	
Algorithm	**FDR**	**KAP1**	**ZFP57**	**KAP1**	**ZFP57**	**KAP1**	**ZFP57**
MACS (default)	1	20,937	1	17,473	1	8,538	20
MACS (relaxed)	20	105,332	2	70,610	1	33,503	21
SPP (default)	1	36,043	0	19,423	0	53	0
SPP (relaxed)	10	138,542	44	79,759	1	40,486	61
PeakSeq (default)	1	18,302	17	8,548	3082	6,031	64
PeakSeq (relaxed)	5	31,917	2761	17,496	8994	14,037	33442

Table 7.2: Number of peaks identified in KAP1 and ZFP57 ChIP samples. For each sample the number of peaks identified using each peak finding algorithm is shown at the default (stringent) FDR and a more relaxed set of parameters.

The KAP1 peaks identified are much more convincing and the experiment appears to have worked well. Peaks can be identified over the locations used as a positive control (see Figure 7.4) as well as in several other locations. MACS KAP1 peaks are taken for analysis as this had been previously identified as a robust algorithm, providing very accurate binding predictions (Wilbanks and Facciotti 2010). Although the SPP pipeline was highlighted as the most accurate algorithm in this publication, it predicted very few PGF KAP1 binding sites, suggesting that its stringency was too high for this analysis. As input material was available for all samples, the accuracy was likely to be high, meaning that MACS was a suitable choice for the final analysis. This analysis suggests that further interrogation of the results of the ZFP57 ChIP-seq data was unlikely to be informative and the remainder of this chapter focuses on exploring in more detail the ChIP-seq data generated for KAP1.

7.3.10 Positive control KAP1 binding sites

ChIP-seq data for the three LCLs is shown in Figure 7.6 for known KAP1 binding sites interrogated by qPCR (Section 7.3.3). MACS detected peaks in all LCLs at these binding sites in *ZNF554* and *ZNF556.* The improved resolution in binding-site detection when using ChIP-seq compared to qPCR is clear, as in both cases the KAP1 binding site is to be located between two qPCR primers sets.

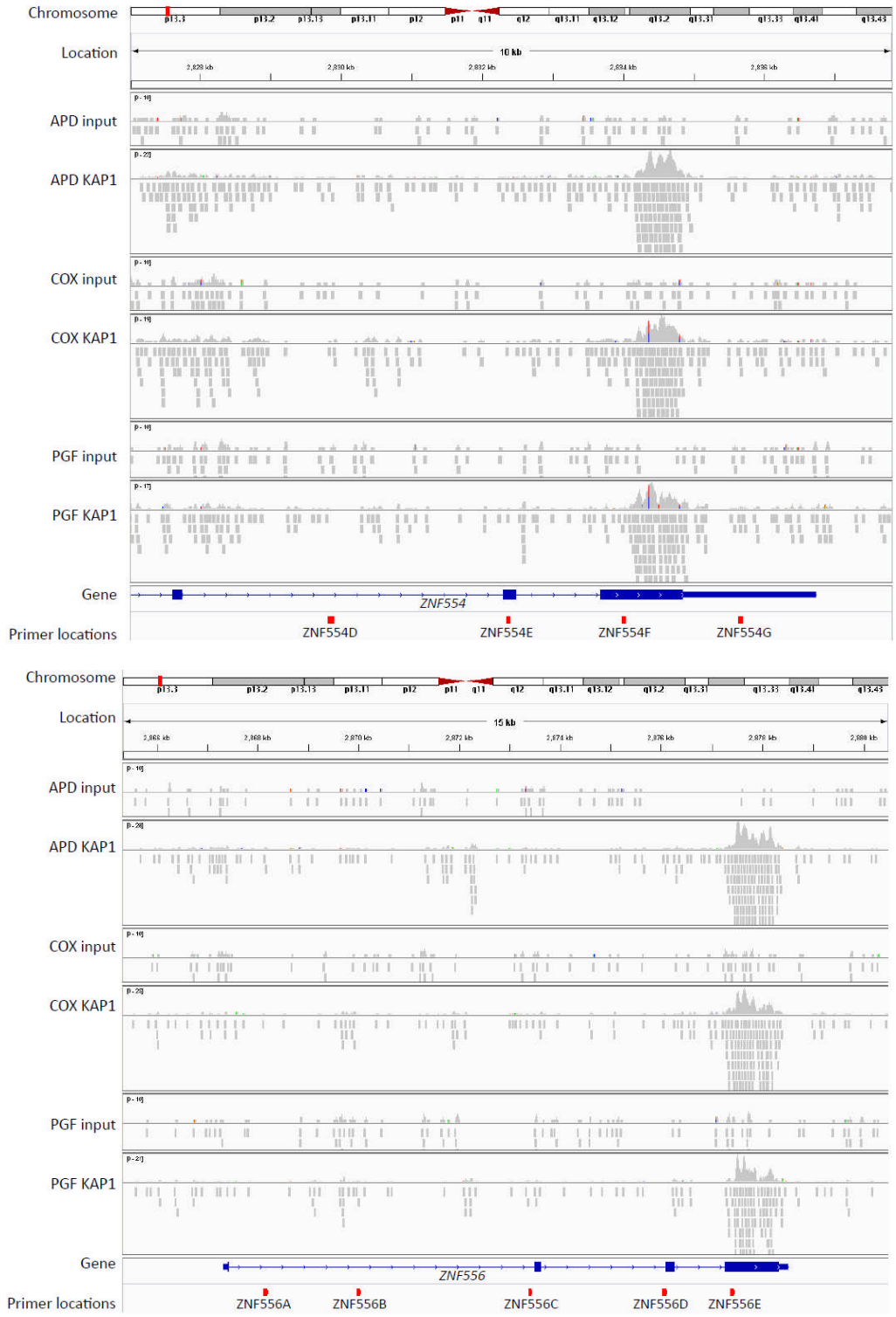


Figure 7.6: KAP1 binding sites determined by ChIP-seq are found over ZNF genes on chromosome 19 previously reported to have KAP1 binding sites. IGV is used to visualise the read frequency of KAP1 immunoprecipitated samples compared to input over ZNF genes on chromosome 19. The positive control genes *ZNF554* and *ZNF556* show clear peaks in positions as expected. These genes were previously selected to confirm the binding of KAP1 at a known binding-site using qPCR prior to submission of the sample for ChIP-seq analysis and the primer locations are shown in red. KAP1 immunoprecipitated samples show significant enrichment of reads compared to the input for the LCLs, and all peaks were detectable using MACS.

This highlights how ChIP-seq can be used to more accurately define binding sites and motifs for TFs and associated proteins. The ability to analyse binding differences between LCLs with known DNA sequence differences could also help to improve models investigating the specificity of particular binding motifs and variation in DNA binding. KAP1 binding was also seen for a further known binding site for KAP1 at *ZNF345* (Figure 7.7).

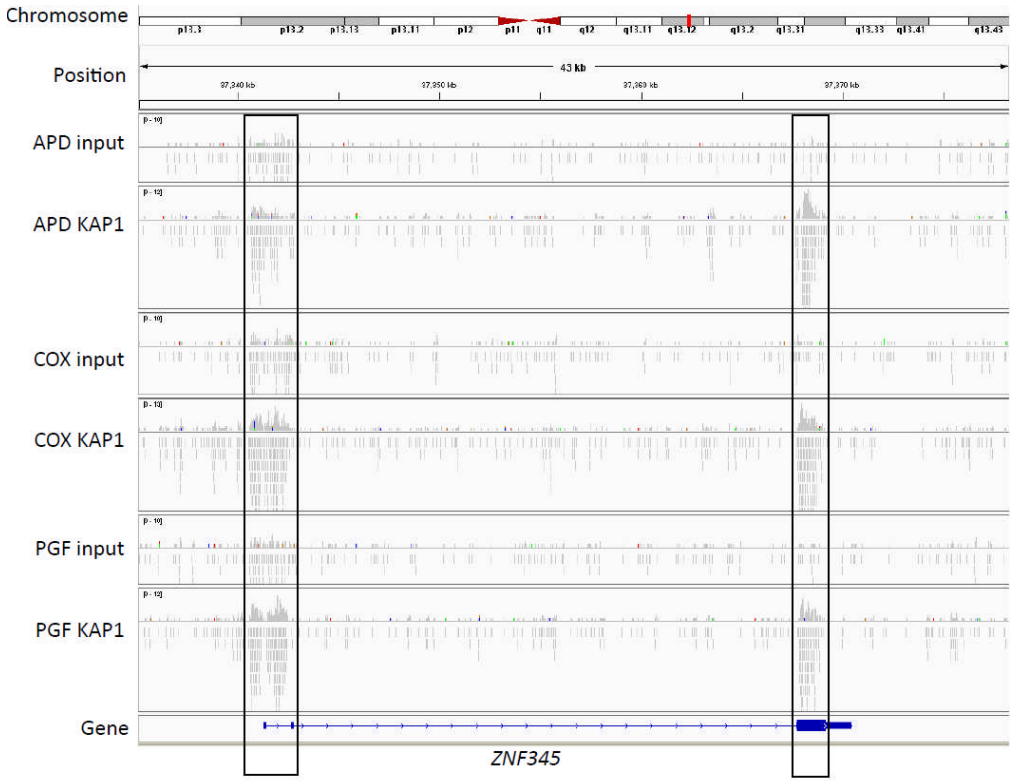


Figure 7.7: Confirmation that KAP1 binding is detected at ZNF345; a region previously identified as a site of KAP1 binding in whole genome analysis of KAP1 binding. IGV is used to visualise the read frequency of KAP1 immunoprecipitated samples compared to the input over *ZNF345*, a gene known to be bound by KAP1 from previous genome-wide analysis (Iyengar 2011).

7.3.11 Genome-wide KAP1 binding

Genome-wide KAP1 binding was assessed using data from analysis with MACS (using the parameters outlined above). Binding sites were required to only be found in one of the three LCLs studied, although many were found in more than one of the LCLs. KAP1 sites were found to have clusters of peaks across the genome; the normalised frequency of KAP1 peaks in all three LCLs is shown below, detailing areas of the genome where KAP1 binding sites are particularly common (Figure 7.8). This

takes account the location of genes in relation to KAP1 binding sites as there is a significant correlation between KAP1 peaks and gene loci ($r=0.65$, $p<2.2\times10^{-16}$). KAP1 binding sites are found to be most common at regions of the genome known to have ZNF gene clusters, for example chromosome 19 (Bellefroid 1993, Eichler 1998). As KAP1 is known to interact with KRAB-ZNF genes and bind over them according to zinc finger number, analysis was carried out to investigate KAP1 binding site location for all genes (Figure 7.8).

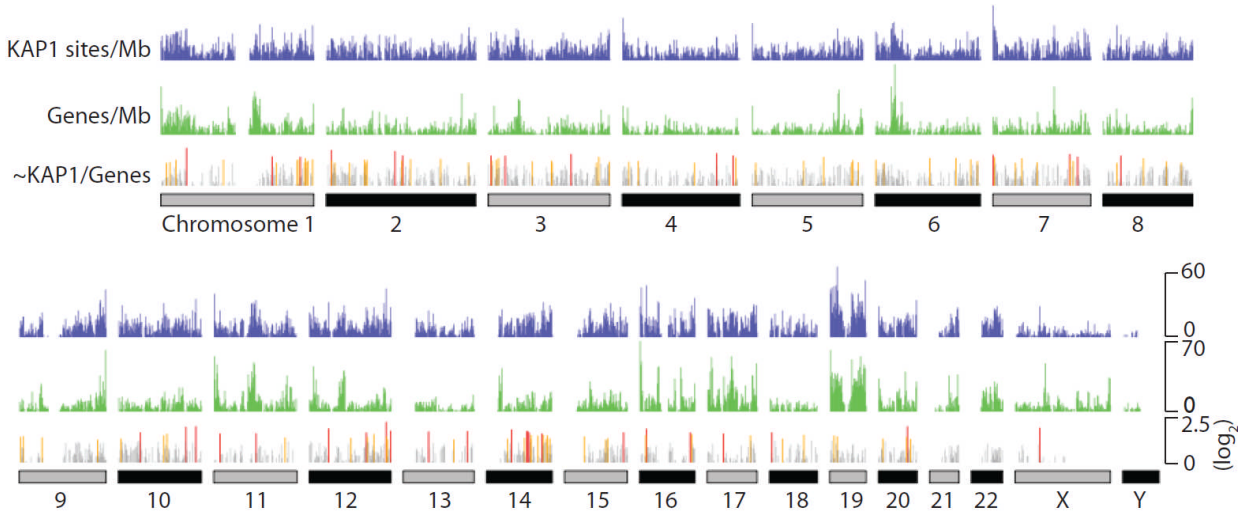


Figure 7.8: Distribution of KAP1 binding sites in mega-base windows over the whole genome (Figure courtesy of Dr Stephen Sansom). KAP1 binding sites (shown in blue) determined in all three LCLs are plotted against genomic location. KAP1 binding sites are significantly correlated ($r=0.65$, $p<2.2\times10^{-16}$) with gene number (shown in green). Regions where the number of KAP1 binding sites exceeded that expected from the number of genes present were identified (~KAP1/Genes) using a linear model. Highly enriched regions are shown in orange (>2 std. dev.) and red (>3 std. dev).

KAP1 binding sites were found at more KRAB-ZNF genes than other types of genes in all three LCLs, and peaks were also found at a higher incidence over KRAB-ZNF genes. The KRAB and ZNF genes were also far more likely to have a KAP1 binding site than any other genes; more than half of all KRAB genes had a KAP1 binding site detected in all three LCLs.

	No. KRAB Genes*	Total No. Sites < 5kb	Average no. KAP1 Sites/KRAB Gene
COX	271	520	1.92
APD	251	420	1.67
PGF	212	287	1.35
	No. ZNF Genes*	**Total No. Sites < 5kb**	
COX	443	805	1.82
APD	441	721	1.63
PGF	324	448	1.38
	No. Other Genes*	**Total No. Sites < 5kb**	
COX	7525	10526	1.40
APD	7478	10126	1.35
PGF	4313	5054	1.17
	***Type of gene**		***Total Gene Number**
	Total No. KRAB genes		354
	Total No. ZNF genes		712
	Total No. other genes (Known protein coding)		19516

Table 7.3: Distribution of KAP1 peaks in relation to all genes and KRAB-ZNF genes. The number of KRAB, ZNF and all other genes that have a KAP1 peak within 5kb of its locus is indicated, with the total number of KAP1 binding peaks found at each type of gene also shown for all three LCLs.

KAP1 sites were shown to be predominantly associated with the 5' end of genes near to the TSS (Figure 7.9). This is unsurprising when considering the general regulatory role KAP1 is predicted to carry out (Iyengar and Farnham 2011). When considering only ZNF genes however, KAP1 binding is found at both the 3' and 5' ends of the gene, which is likely to be a result of the association with the zinc finger motifs at the end of the gene (Iyengar and Farnham 2011).

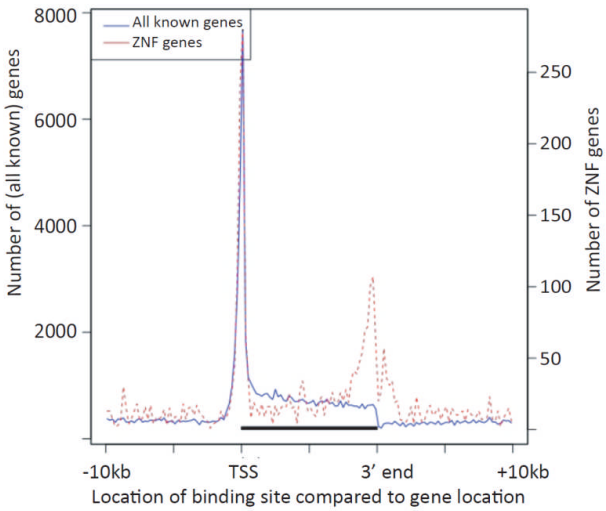


Figure 7.9: Distribution of KAP1 binding sites at loci containing genes (Figure courtesy of Dr Stephen Sansom). For each gene with KAP1 binding sites identified across the genome, the location of the KAP1 binding site relative to the gene was analysed. 10kb upstream and downstream of every gene was investigated, and all gene lengths were normalised to 10kb for the purpose of the graph. For most genes binding sites were located at the 5' end over the TSS, but for ZNF genes the binding sites were also found to be located at the 3' end of the gene.

While the KAP1 sites are enriched at ZNF genes, binding was also seen involving non-ZNF genes. A binding site for KAP1 is shown over the promoter region for the gene *EFCAB7,* which was previously identified as a KAP1 binding site in HEK293 cells (Iyengar 2011). Consistent with the data shown in Figure 7.9, a binding site is only seen at the 5' end of the gene over the TSS (Figure 7.10).

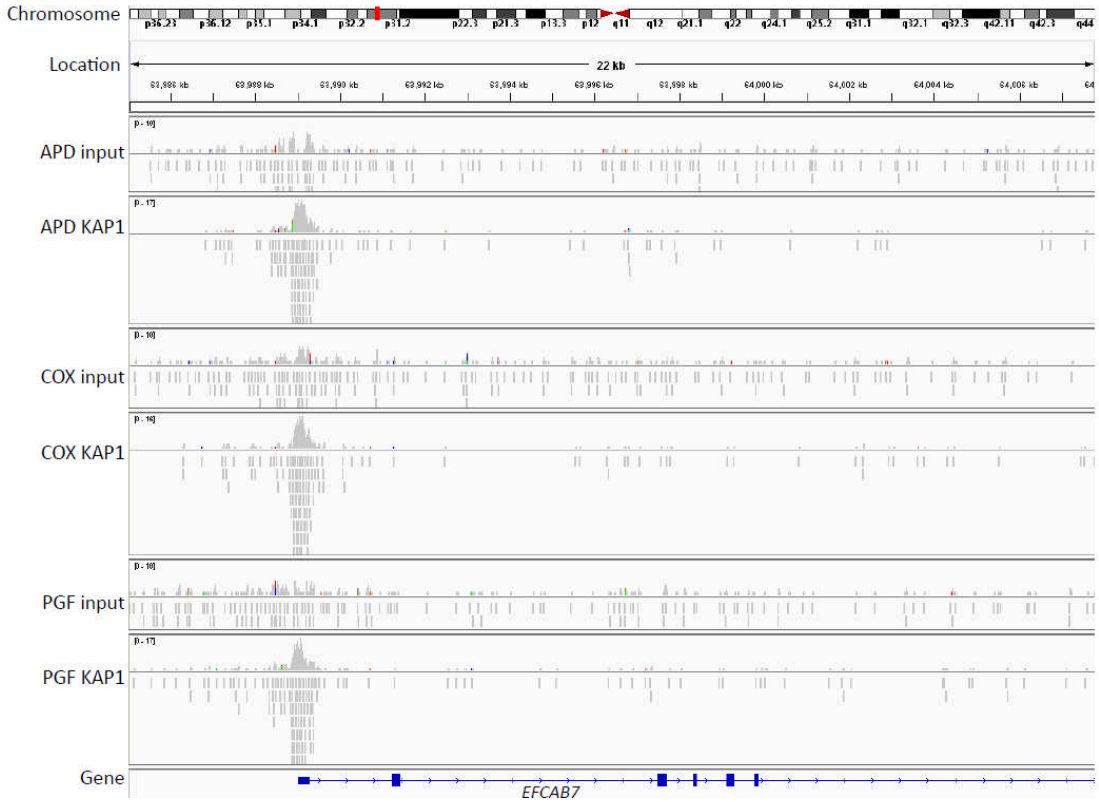


Figure 7.10: KAP1 binding on *EFCAB7* a non-ZNF gene is found predominantly at the 5' end. IGV is used to view the read frequency of KAP1 incubated samples compared to the input over the *EFCAB7* gene. A strong peak in KAP1 incubated samples is found over the 5' end of the gene demonstrating strong KAP1 binding site in the promoter.

7.3.12 Differential KAP1 binding between LCLs

KAP1 binding was investigated for the region spanning the *ZFP57* gene. A novel binding site was found 3kb centromeric of *ZFP57* in the COX LCL, and to a lesser extent the APD LCL (Figure 7.11). This is particularly interesting as the strength of peak signal seen matches the mRNA levels seen in the *ZFP57* gene expression analysis across the three LCLs; highest in COX, intermediate in APD and lowest in PGF (Figure 3.3). While the peak in APD was not detected by the MACS peak caller, it is clear from comparing the input and APD KAP1 ChIP samples that low level binding over a relatively extended area is present.

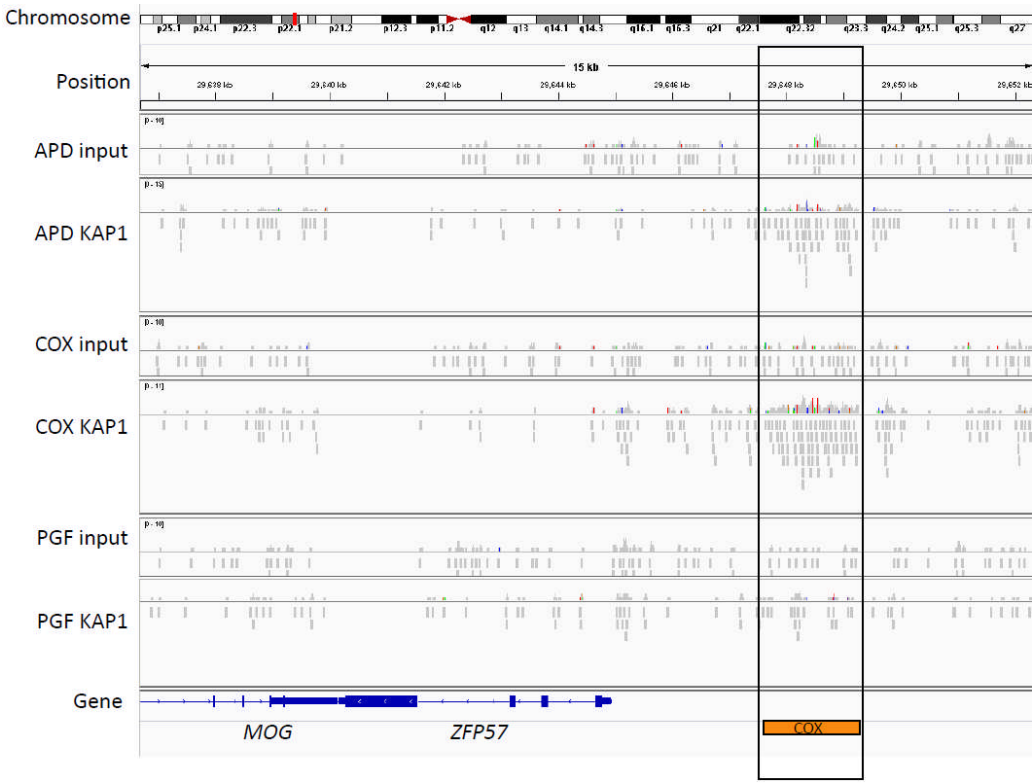


Figure 7.11: KAP1 binding upstream of *ZFP57* in COX and APD, with no binding seen in PGF. IGV is used to visualise the read frequency of KAP1 immunoprecipitated samples compared to the input over the region upstream of *ZFP57*. MACS detected a peak in the COX line at this location, indicated in orange. No peak was called for the other 2 cell lines, although there is some evidence that APD may show KAP1 binding at this location. Due to the likely contamination of amplified *ZFP57* corresponding to coding sequence from cloning reactions in the ChIP-seq samples (Section 7.3.6), no analysis of whether differential KAP1 binding can be seen over the 3' end of the gene can be carried out.

As the sequence is known across the MHC for the APD, COX and PGF LCLs as a result of the MHC Haplotype Project (Horton 2008), this can be used to assess further allele-specific KAP1 binding events in the MHC. While initial analyses did not find large differences in KAP1 binding between points of known variation in the LCLs, there were two notable findings involving *TRIM31* and *PSOR1C1*. Further analysis of this type could be used in future to look for more subtle variation between the LCLs rather than presence or absence of peaks.

TRIM31

The function of TRIM31 is not fully understood, however it is a characteristic RBCC protein like KAP1 (also known as TRIM28) containing a RING finger, B box and coiled-coil domains (Sugiura and Miyamoto 2008). It is thought to be important in cancer development and anti-viral response based

on its up-regulation in gastric cancer; however the mechanisms for its involvement are unclear (Sugiura 2011). Over the 3' end of the gene, a peak of signal for KAP1 binding is found in the COX LCL only, overlaying the final exon, which can be seen in Figure 7.12. This is perhaps unsurprising as the B box domain is a zinc-finger domain and so would expect to be correlated with KAP1 binding at the 3' end of the gene. However, it is interesting that MACS does not call a peak in the other LCLs and the peak corresponds with sequence variation in the COX line when compared with the reference (PGF) sequence. Six SNPs are found between the COX and PGF sequences over the binding site that is identified in the COX LCL (co-ordinates in PGF are chr6:30,070,489-30,071,650).

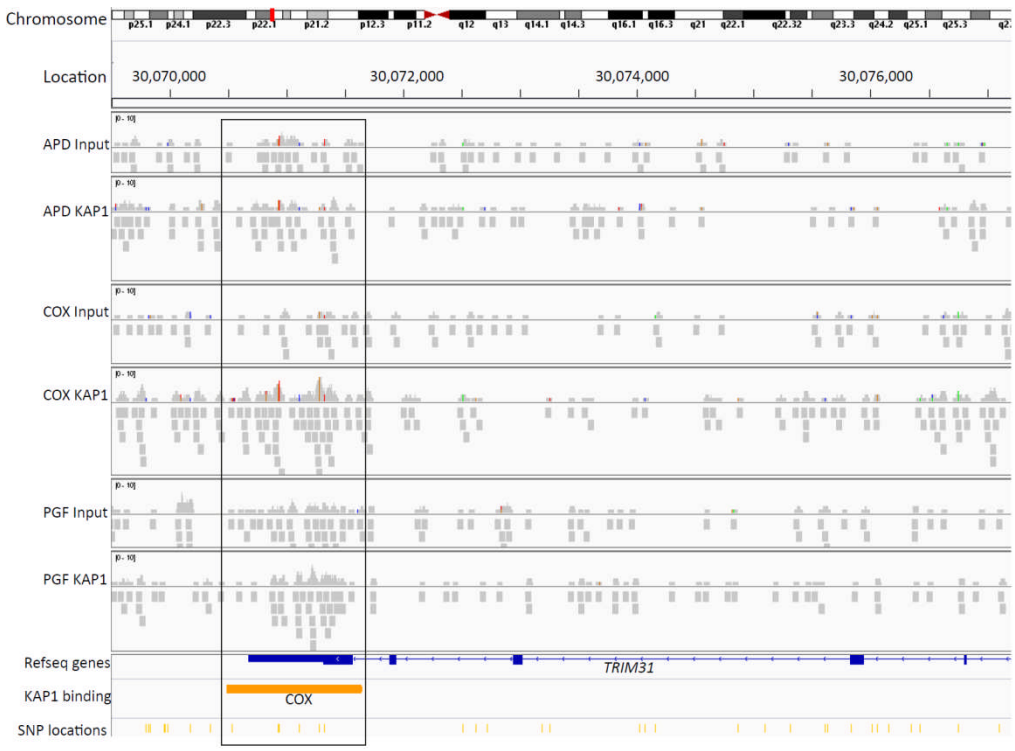


Figure 7.12: KAP1 cell line specific binding in the *TRIM31* gene 3' end. IGV is used to visualise the read frequency of KAP1 incubated samples compared to the input over the *TRIM31* locus. A peak of KAP1 binding detected by MACS is found in the COX cell line only at the 3' end of the gene.

Differential expression is thought to be one of the mechanisms controlling TRIM31 levels, and so the differential binding of KAP1 could be contributing in some way to the overall control of TRIM31 expression. The binding site seen in the COX LCL at this location has also been identified in KAP1 ChIP-seq study, supporting the likelihood that this is a true KAP1 binding site (Iyengar 2011).

PSORS1C1

Another candidate within the MHC showing haplotype specific KAP1 binding between the cell lines is

the gene *PSORS1C1*. *PSORS1C1* has been previously implicated by familial analysis with the autoimmune disease psoriasis (Trembath 1997). Due to LD in the region it is difficult to determine the true locus of association. However the HLA type HLA-C*06 contains SNPs upstream of *HLA-C* in strong LD with the associated SNPs in *PSORS1C1* (Liu 2008). In Chapter 5, increase in *HLA-C* expression was shown to be associated with the HLA-C*06 allele, and interestingly the LCL APD which carries the HLA-C*06 allele had highest expression of HLA-C of all the LCLs analysed (see Figure 3.11). In the analysis of the KAP1 binding ChIP-seq data, MACS called a peak of KAP1 binding towards the 3' end of the gene in the APD LCL only, and visual confirmation of this showed that no evidence of a peak in the other two LCLs could be seen (Figure 7.13).

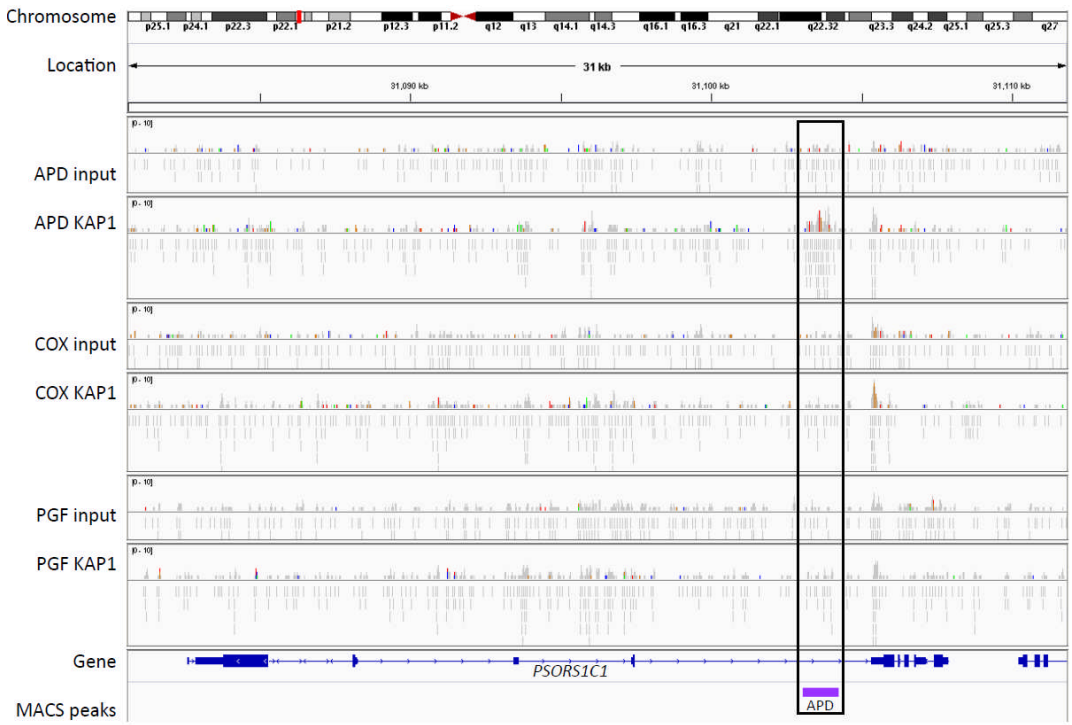


Figure 7.13: KAP1 APD cell line specific binding in the *PSORS1C1* gene 3' end. IGV is used to visualise the read frequency of KAP1 incubated samples compared to the input over the *PSORS1C1* gene. MACS detects a peak of KAP1 binding in the APD cell line only, and no sign of binding is found in the other 2 cell lines.

7.4 Discussion

Data analysis of the ChIP-seq experiment was carried out using three different algorithms to ensure that the correct conclusions were drawn. MACS and SPP have been noted to be reliable at finding binding sites, however they are also predicted to miss many true sites due to their stringency (Wilbanks and Facciotti 2010, Malone 2011). Peakseq has been shown to find the most peaks of several algorithms compared using the same dataset, however it is likely that many of these are false (Malone 2011). By using more than one algorithm to analyse ChIP-seq data, it is likely that a balance can be struck between too-relaxed and too-stringent parameters for analysis.

7.4.1 ZFP57 ChIP-seq

The ChIP-seq experiment could not be used to analyse the binding of ZFP57 in the MHC homozygous LCLs as no true peaks could be determined by any of the three algorithms used in the analysis. Therefore, no new functional information for ZFP57 binding sites across the genome can be determined. This shows that either the antibody against ZFP57 is not suitable for ChIP experiments or the expression of ZFP57 is so low in LCLs that ChIP is insufficiently sensitive to detect with confidence the presence of DNA binding sites. This could also be the case if ZFP57 binding is very weak, and so is not preserved by the formaldehyde cross-linking reaction carried out prior to the extraction of chromatin. There is also no alternative antibody currently available, especially one that has been assayed for its effectiveness in ChIP experiments, so this also makes altering the experiment to try to improve the outcome challenging. An antibody against ZFP57 could be raised; however it is likely that functional protein would need to be used as the immunogenic peptide, to be sure that the antibody raised could recognise the native protein when cross-linked to chromatin. Further analysis of ZFP57 binding would therefore require many more experiments which are beyond the scope of this thesis.

7.4.2 KAP1 ChIP-seq as a positive control in the study of ZFP57

The choice of KAP1 as a positive control was successful and it proved that the ChIP-seq experiment itself was functioning as expected and that the analysis pipeline provided accurate and consistent

results. As KAP1 binding sites were seen at many ZNF gene clusters it reinforces the idea of using KAP1 as a co-factor when studying the binding of ZFP57 and this could be used in further study if this becomes possible for ZFP57. The KAP1 ChIP-seq data has clearly shown peaks at known KAP1 binding sites on chromosome 19, also ZNF genes, and the data is comparable across the different LCLs in many positions showing conservation of the binding sites and reproducibility. Two of these sites were also identified as positive for KAP1 binding in the qPCR analysis carried out on the KAP1 ChIP samples before submission to sequencing showing that there is correlation between the results determined by ChIP analysed by qPCR and by high throughput sequencing.

Intriguingly, a recent study has shown that KAP1 is likely to mediate the function of ZFP57 in recruitment of DNA-methyltransferases in mice embryonic stem cells as mutant ZFP57 which lacked the KAP1 interacting KRAB box was unable to maintain embryonic stem cell methylation patterns (Zuo 2011). In a separate study also working in mouse embryonic stem cells, the maintenance of methylation at imprinting control regions was shown to be reliant on functional ZFP57 as this is lost when ZFP57 is deleted in engineered mouse embryonic stem cells. If KAP1 is deleted, using a drug-induced knock-out in a different mouse embryonic stem cell line, then heterochromatin marks at the imprinting control regions are lost (Quenneville 2011). The Quennville study also finds a six base-pair binding consensus sequence for ZFP57, which would be interesting to compare to future work on ZFP57 binding in human LCLs or primary tissue. The motif-binding section is found at the most conserved part of the ZFP57 protein between mice and humans, and the six base-pair motif can be found in both mice and humans, particularly at imprinting control regions.

7.4.3 KAP1 genome-wide binding

Chromosome 19 contains several large ZNF gene clusters and shows the densest co-incidence of KAP1 binding sites detected across the whole genome. As chromosome 19 has been found to have 13 separate ZNF gene clusters, each containing up to 76 different ZNF genes this is perhaps unsurprising (Tadepally 2008). ZNF clusters are found throughout the genome, with at least one found on each chromosome and totalling a large number of human genes; for example there are 423 known KRAB ZNF genes, 384 of which lie in clusters while the remaining KRAB-ZNF genes are found

alone (Huntley 2006). In total 65 KRAB-ZNF, SCAN-ZNF, and mixed gene clusters are found in the human genome, but while ZNF genes are often found together in clusters, this does not mean that they are co-regulated, in fact many genes within the ZNF clusters are in fact regulated separately (Huntley 2006). This makes it even more likely that ZNF gene clusters would have several KAP1 binding sites, as it is likely that if KAP1 plays a role in their regulation it will be necessary to be able to regulate each separately.

7.4.4 Differential KAP1 binding in the MHC

KAP1 was show to bind upstream of *ZFP57* in the COX and APD lines, consistent with a role in regulating expression of *ZFP57*. This could perhaps be with the eventual outcome of controlling heterochromatin as has been previously suggested (Groner 2010) through ZFP57 expression, which has itself been associated with DNA methylation pattern maintenance (Li 2008). Differential binding at this location could be a result of underlying sequence variation between the lines, or as a feedback mechanism due to the expression of ZFP57 itself affecting the KAP1 protein presence at the binding site. Although sequence variation does occur under the predicted binding site in COX, there is no difference between the APD and PGF lines, so this is unlikely to explain the differential binding completely. It is perhaps most likely to be a combination of sequence variation and regulatory feedback created by expression of ZFP57 itself, tying in with the earlier hypothesis that ZNF genes may be subject to auto-regulation (O'Geen 2007).

Differential binding of KAP1 could be analysed more exhaustively in the three LCLs, particularly in regard to the MHC where detailed sequence information is known for each cell line. This could be particularly insightful in the Zinc finger gene cluster found within the extended MHC. There are 36 separate loci within the cluster, and this further subdivides into different classes of zinc-finger proteins (Horton 2004). Several of these ZNF genes found within the class I of the MHC have sequence information available from the MHC haplotype project in COX, PGF and APD (Horton 2008) and so can be analysed in a sequence-specific manner. An initial analysis has shown differential binding in the *ZFP57* locus as well as *TRIM31* and *PSOR1C1*.

TRIM31 showed differential binding of KAP1, with a detectable peak in the COX LCL only. This was

associated with known variation between the COX sequence and the PGF reference sequence. Expression levels of *TRIM31* using the MHC array data (Figure 3.1) show low expression in the three LCLs analysed (COX, QBL and PGF) though there is a slight increase in expression in the COX LCL compared to the PGF LCL, implying that differential binding of KAP1 in this region could be affecting the gene expression of *TRIM31* (MHC custom array gene expression data for the APD LCL is not available). Both TRIM31 and KAP1 up-regulation have been previously linked with gastric cancer (Sugiura and Miyamoto 2008, Yokoe 2010), so the association between the two genes is interesting. The *PSORS1C1* gene showed differential KAP1 binding specific to the APD LCL. Detailed sequence information was not available for the APD LCL at this location, however APD was the only LCL studied to carry the HLA-C*06 haplotype, which has previously been associated with psoriasis susceptibility (Liu 2008). This could be a possible mechanism of a difference between the HLA-C*06 haplotype compared to possession of other particular haplotypes. Further analysis of expression of *PSORS1C1* and KAP1 binding for individuals of diverse genetic backgrounds may help to determine if this observation is specific to the HLA-C*06 haplotype. Modulation of *KAP1* expression, for example using siRNA, would help determine if KAP1 has a direct effect on expression of PSOR1C1 in the presence of a specific genetic background.

7.4.5 Conclusion

In conclusion, although the immunoprecipitation of ZFP57 was not successful the promise of recent studies in mice makes this an exciting area for further work and validates the choice of KAP1 as a control for ZFP57 binding. ChIP-seq for KAP1 was informative, and identified known KAP1 binding sites, as well as sites that could be specific to individual LCLs, particularly for *TRIM31* where differential binding was seen between the three LCLs analysed at a previously detected KAP1 binding site.

Chapter 8: General discussion

In this chapter, I will assess the overall findings of this thesis in the context of this field of research, and identify areas of further experimentation and research that could be considered in the future.

8.1 MHC and disease

As one of the most variable and gene dense regions in the human genome, the MHC has been the subject of great interest to geneticists. Combined with its long association with disease of different types, this has lead to intensive study since its discovery more than 50 years ago (Dausset 1981). It has been the region most often associated with complex disease traits when studied by GWAS. This underlines previous findings of striking disease association by serological testing and candidate gene association, for example the association of the HLA-B*27 allele with incidence of ankylosing spondylitis (Brewerton 1973). Susceptibility to both autoimmune and infectious disease is modulated by possession of particular MHC variants and haplotypes (Hill 2006, Fernando 2008b). This can be at a structural level, for example modulation of binding sites in the classical MHC genes leading to differential antigen binding or protein-protein interactions (Jones 2006). Another of the ways in which these variants is thought to modulate disease susceptibility and incidence is through modulation of gene expression of particular MHC genes. Therefore the study of gene expression in the MHC and in particular, the role of ASE, in this context was a relevant and interesting research question to address in this thesis.

8.2 ASE and the MHC

ASE has traditionally been difficult to study using array-based technology. Most arrays do not discriminate between different alleles, and SNPs in probes of arrays has led to inaccurate gene expression analysis (Kreil 2006, Alberts 2007). Other methodologies have relied upon the presence of a marker SNP within the exons of the transcript of interest to define allelic origin. One way to avoid this is by use of the HaploChIP technique (Knight 2005a); however this may not be suitable for analysis of low expressed genes as the technique is dependent upon the cross-linking of RNA polymerase to the DNA. The HaploChIP technique is also labour intensive, leading to low throughput

of experiments.

The extreme polymorphism of the MHC has also meant that design of primers for gene expression analysis by quantitative PCR is difficult and, in some genes, impossible to achieve without the presence of confounding variants. A custom MHC array exploring expression variation was therefore developed which included alternate allele probes (Vandiedonck 2011). This was applied to the study of LCLs which were homozygous across the MHC and had been completely sequenced across this region, allowing a direct comparison of ASE across three LCLs. These possessed ancestral haplotypes that were previously associated with common autoimmune and complex disease (Price 1999, Johansson 2003, Larsen and Alper 2004) meaning that analysis of their characteristic gene expression may prove informative of the mechanisms of their disease association. An example of this was the finding that HLA-DQB*06 was strongly associated with *HLA-DQB1* expression, but not with the expression of other genes in the region. As the ancestral extended haplotype HLA-A3-B7-Cw7-DR15 (PGF) carried this allele, it is possible that differential expression of *HLA-DQB1* could play a role in the disease susceptibility or protection seen to be associated with this haplotype. For other diseases contrasting associations were seen, for example T1D. Here, a strongly protective effect against T1D is reported for possession of the HLA-A3-B7-Cw7-DR15 haplotype (Larsen and Alper 2004) but the *HLA-DQB1* expression-associated SNPs include those that are reported to be associated with an increased T1D risk by GWAS (WTCCC 2007). Other variants within the HLA-A3-B7-Cw7-DR15 haplotype may be more significant in determining the overall observed protective effect of this haplotype in T1D.

Differential gene expression seen in this analysis was replicated in cohorts of healthy individuals. Expression variability was therefore proven to be a replicable phenomenon seen in PBMCs in addition to transformed LCLs. Using genotyping information, variation in gene expression was also mapped as a quantitative trait. Classical HLA alleles for the volunteers were inferred from the genotypes of the individuals, meaning that analysis could also focus on the association of classical serologically defined types in addition to the association with individual SNPs.

8.3 Genes and differential expression

While some of the genes that were selected for expression quantitative trait mapping, such as *HCP5* and *HLA-DQB2*, did not show strong genetic association with variable expression levels (see Chapter 4), it was clear that the majority of genes analysed were influenced by genetic variation. In the case of *HLA-DQB2* for example, it should also be noted that the observed levels of expression were too low to accurately measure using qPCR or standard microarray. For the other MHC genes that did show evidence of strong genetic associations such as *HLA-DQB1*, *ZFP57* and *HLA-C* (see Chapters 3 and 4) further fine-mapping of the associations could be carried out. Additional individuals could be recruited to replicate the observed *cis* associations and increase the power to detect *trans* associations. Imputation was used to further resolve the observed associations with genotyped SNPs, for example leading to an additional 683 SNPs in *cis* identified as associated with *ZFP57* expression when imputation was carried out on the locus. To further define associated variants, Sanger sequencing of regions shown to be strongly associated with expression could be used to define all variants in the locus and fine-map those likely to be associated. However, this approach may be less relevant in the MHC, which has been extensively resequenced compared to other genomic regions. A variety of approaches can then be used to help identify the causative SNP including gel-shift assays to detect protein-DNA interactions or reporter-gene assays. The HaploChIP assay could be particularly informative in proving a SNP directly affects the binding of a particular candidate TF or RNA polymerase (Rebbeck 2004).

The ability to carry out the gene expression analysis in the healthy volunteer cohorts, described in Chapters 3, 4 and 5, in primary cells showed that cell lines could be used to pinpoint interesting biological phenomena that were also relevant in primary tissue. Examples of this were the striking genetic associations to differential expression seen with the genes *ZFP57* and *HLA-DQB1* in particular. The volunteer study used primarily a mixed PBMC population, but separation of B cells and monocytes from these individuals by other members of the Knight lab also showed that eQTL analysis may benefit from using populations of cells that are of one type only (Fairfax In Press). This would allow identification of cell-type specific eQTLs and could help to further define disease

pathogenesis mechanisms when these were compared with disease associated SNPs identified by GWAS or other studies. If clinical samples, for example from patients with autoimmune conditions, could be used to map genetic variants associated with gene expression in the context of variants or genes known to be associated with disease it could become a powerful tool for mechanistically determining the pathogenesis of disease. It is likely that many instances of genetic variants that modulate gene expression and impact on disease will be context specific and only found within patient samples.

8.4 Haplotypic analysis

As previously described in Chapters 4 and 5, comparison between known HLA disease susceptibility variants and gene expression associated with these different alleles could be used to pinpoint functional genes that influence disease susceptibility. When comparing the GWAS catalogue of SNPs marking regions associated with disease, or their proxies, to SNPs found to be associated with *HLA-DQB1* expression there were many overlapping significant SNPs. Several of these SNPs marked both an association with disease known to be associated with this haplotype and with *HLA-DQB1* expression. Thus it was hypothesised that for some diseases, the manner in which such susceptibility arises could involve *HLA-DQB1*. The most striking example was the analysis showing that *HLA-DQB1* expression variation was influenced by possession of several SNPs shown to be involved in leprosy susceptibility. These SNPs were also associated with other HLA-DQ genes; the strength of association was far higher with *HLA-DQB1* expression (p values as low as 2.45×10^{-54}) thereby implicating *HLA-DQB1* in leprosy pathogenesis (Chapter 5).

While different MHC alleles have been historically associated with autoimmune disease, more recently in the era of GWAS, high resolution mapping of disease susceptibility loci through SNP markers has been possible (Hirschhorn and Daly 2005). Therefore, is it necessary to consider classical HLA typing when investigating disease association or the mechanisms in which they function? As the type gives information about differences relating to the binding specificities of the HLA molecules, it may be that their association with disease gives a greater insight into function than SNP proxies that are close by or in LD. This is particularly true when considering the high LD found in the MHC along

with the profound structural variation, and the extensive sequence similarities found between different loci in the region (Traherne 2008). Therefore, diseases where the MHC association has been only partially characterised in particular may benefit from the combination of these approaches. An example of this would be the uncertainty over the contribution of different HLA loci in diseases such as psoriasis which has been further characterised by a combined approach considering SNPs and HLA type (Nair 2006). Many autoimmune diseases have been found to share associated loci, and so further integrated analysis may be useful in resolving the function of these loci in different contexts (Cotsapas 2011). In the case of the psoriasis locus, *HLA-C* expression was shown to be associated with disease-associated SNPs, while *HCP5* expression was not found to change according to genotype at these locations (Chapter 5).

Specific examples such as these show how knowledge of SNPs and HLA types can further aid the elucidation of disease mechanism and fine-mapping of functional variants. It is also clear from this type of analysis that GWAS can be effectively combined with functional eQTL studies in order to help define modulated genes and candidates for further analysis. Associations of variants that are found in common between disease and expression traits are particularly interesting, as they can be used to infer how modulation of levels of a particular gene could be deleterious.

8.5 NGS technology and RNA-seq

In the last 10 years, the cost of sequencing a genome has dropped dramatically. The introduction of massively parallel sequencing has allowed the development and widespread adoption of NGS. In the future, the use of NGS for studies of the HLA region is an exciting possibility. RNA-seq will allow more accurate measures of overall and allele-specific gene expression. Less normalisation is required for RNA-seq than for microarray gene-expression analysis, and in theory absolute levels of gene expression could be determined, making it particularly useful in comparison of disease to control samples. As the depth of sequencing is the major limiting factor for determination of expression of lowly expressed genes, high-depth RNA-sequencing will be invaluable in this analysis (Morin 2008). The analysis of three biological replicates of an LCL (GM11830) using mRNA-seq showed that highly expressed genes had highly reproducible detectable expression between replicates. Where a

heterozygous SNP was found in a gene with over 50 reads per allele, the standard deviation of the ratios of the two alleles was very small (0.0076-0.034). However, where read counts were lower, the variability between replicates increased significantly, with the standard deviation between observed ratios increasing to 0.22-0.25. This is unsurprising when considering previous studies that have investigated ASE using RNA-seq. A proposed minimum read count of each alternate allele to detect ASE without false positive results was suggested to be 50 reads (Heap 2010). This seemed to agree with the results seen when the PGF and COX RNA was mixed to give five samples with different mixed ratios of the two RNA samples. When a SNP was created by mixing these lines then analysed across the five samples, the observed ratios between the two alleles were found to be significantly associated with the expected ratios when the read count for both alleles was greater than 50. While the initial results of the mRNA-seq experiment were encouraging, there were significant analytical problems with genes in the MHC. Mapping using Stampy led to low coverage, as many transcripts could not be unambiguously mapped and so were removed, particularly when the COX single haplotype was used in mapping. It is unclear how accurate traditional mapping using either one or all available reference sequences over the MHC will be, when samples that are dissimilar to the available references are sequenced. Reference-free mapping is a promising technique for future analysis of the MHC in RNA-seq data sets for this reason. The test example over an exon in *HLA-B,* showed the correct trend in observed ratios between alleles of an raSNP in the five mixed RNA samples. Further analysis will be required to determine the most appropriate method for mapping and defining variants from RNA-seq data, especially in the MHC. This is likely to involve a combination of both traditional mapping and reference-free mapping to ensure maximum detection of variants and to avoid bias in the results.

8.6 ChIP-seq

A further important application of NGS has been the development of ChIP-seq, allowing higher resolution of genome-wide TF binding or histone modifications compared with other techniques. Although binding sites for ZFP57 could not be determined in our experiment, the power of the approach was illustrated by the analysis of KAP1. It is likely that not only the quality of the ZFP57

antibody, but also the extremely low expression of the gene itself contributed to the failure to find binding sites. Generation of a new antibody may help to improve this, particularly if ZFP57 in its native state were to be isolated and used as the antigen. Optimisation of the antibody and incubation times, temperature, buffer and cross-linking could also be carried out to ensure optimum experimental conditions. Additionally, use of material that has higher expression of ZFP57 could help to ensure that enough material remains following the incubation with the antibody for successful isolation of the DNA and amplification. Over-expression of cloned ZFP57 protein and then use of this to analyse ZFP57 binding is a possibility, especially as a tag could be included for which there is already a ChIP-optimised antibody. However, it is unclear whether the resulting data would have biological relevance as *ZFP57* expression is generally low, so such a strategy may introduce false positives. As NGS technology continues to improve, it is likely that less starting material could lead to successful analysis of TF binding as higher depth of sequencing can be achieved. In our analysis of KAP1 binding we found no binding sites to be "saturated" and so it is likely that sequencing to a higher depth would have uncovered further binding sites that did not reach a level of significance in our study (Park 2009). As depth of sequencing improves it is likely that genes with very low expression will become more viable targets for study.

The use of biological replicates in ChIP-seq studies has already been shown to be useful in minimising the incidence of false positive prediction of binding sites and maximising the number of sites that can be determined (Tuteja 2009). Multiple replicates will likely help to improve the resolution of the binding sites, particularly in the case genes with low expression. Even if only small numbers of peaks can be detected, if they are replicated in new samples (and preferably with a different antibody used against the same TF) then there is a high chance that these are true binding sites. The ability to compare ChIP-seq results with known binding partners of TFs will also hopefully lead to more accurate findings when looking at binding patterns of low expressed genes. With the increasing amount of data from the ENCODE project freely available (Rosenbloom 2011) it may be possible to use the binding sites found in ChIP-seq studies to filter likely binding sites for known partners of these TFs. This would rely on binding being consistent between different cell lines unless the same

cells were used for the study. In the case of the use of primary cells in ChIP-seq, the pre-existence of cell line data would be useful for comparison, and could be used to compare known strong binding sites as a positive control.

The increase of ChIP-seq studies in comparison to ChIP-chip or other ChIP experiments allows not only the detection of binding over the whole genome, but also allows interrogation of factors previously difficult to determine on a large scale. These include allele-specific binding events and binding of TFs in regions that are highly repetitive and tend to be masked in arrays (Park 2009). We have shown that knowledge of the underlying genetic sequence can lead to identification of sequence-specific binding events found between the MHC homozygous LCL lines studied. This showed how use of the LCLs for which genetic information is becoming available in the 1000 genomes project (Pennisi 2010) could become extremely informative when applied with experiments such as ChIP-seq.

Allele-specific ChIP-seq is beginning to be used, where sequence variation is taken into account. Pipelines for the analysis have been set up, and so future experiments for either ZFP57 or KAP1 would not have to use homozygous LCLs to determine differential binding (Rozowsky 2011). Further analysis of the data, or of new replicates of the ChIP-seq samples, could determine binding in a more continuous manner, rather than purely looking for the presence or absence of a binding site. This would help to more precisely define a binding motif, as the sequence variation under the peak could be filtered by peak strength to find the optimum binding motif. This could also be achieved with positions of known heterozygous SNPs under a binding site, looking for over-representation of one of the alleles over the other.

8.7 Epigenetic analysis and chromatin profiling

Although not addressed in this thesis, epigenetic mechanisms have a significant impact on gene expression (Pastinen 2004) and can be influenced by underlying genetic variation (Hellman and Chess 2010). DNA methylation and variable histone marks are known to lead to dramatic changes in DNA structure and gene expression (Heintzman 2009). This is likely to affect the binding of TFs and RNA polymerase, therefore it may also be important to take this into account when analysing ChIP-seq

binding results and other regulatory analysis. Future experiments could analyse these differences in DNA modification and structure, with a view to combining epigenetic analysis with genome-wide sequence variation in the volunteer cohort. Alternatively, recruitment of a new cohort of related samples, for example parent and child trios, could allow assessment of inherited epigenetic marks that are not affected by sequence variation.

A number of techniques can be used to study chromatin and help focus efforts to resolve regulatory genetic variants. Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) allows the analysis of nucleosome depleted chromatin, which reflects the chromatin found in a more open state. In order to achieve this DNA is cross-linked, sonicated, and then extracted using phenol-chloroform (Giresi 2007). The aqueous phase contains the nucleosome depleted DNA and this can then be sequenced for genome-wide analysis. Likewise, DNA susceptible to digestion by DNase I is thought to be nucleosome depleted and thus contain more active regulatory elements such as TF binding sites and active promoters (Song and Crawford 2010). DNase I digested DNA can be analysed by sequencing to define these hypersensitive sites. These techniques have been successfully used in 7 cell lines, showing shared FAIRE or DHSs to be close to promoters of CTCF binding sites while sites found in only 1 cell line tended to be found away from these features. They were also found to contain DNA binding motifs for TFs associated with cell type identity (Song 2011).

Global methylation patterns are likely to be different in patients suffering from a disease, whether due to environmental or genetic reasons (Bell and Beck 2010), and this could be important to analyse not only for the functional information it is likely to give, but also may be specific to prognosis, and so could be used as a clinical tool. A genome-wide assessment of DNA methylation is currently possible using whole-genome bisulphite sequencing (BS-seq), methylated DNA immunoprecipitation sequencing (MeDIP-seq) or methyl-binding protein-sequencing (MBP-seq). However, although BS-seq is the most accurate way of assessing global methylation, it is inefficient in terms of both time and cost as it is difficult to align the bisulphite converted sequences to a reference. MeDIP- and MBP-seq are both enrichment based, meaning that although they can clearly define highly methylated sequences, they are biased and so cannot be used quantitatively (Li 2010).

Ideally epigenetic changes such as chromatin modifications and DNA methylation, genetic variation, gene expression and relevant TF binding would be assessed in the same samples to give the most complete picture of function. This could then be used to discern the most likely mechanisms by which a heritable variant may be acting. While this has not been possible in this work, it is clear that as technology continues to improve and costs continue to drop, it may well soon be possible to begin to assess more of these features together. In the case of ZFP57 for example, the ability to analyse the differential methylation patterns over the whole genome would prove particularly informative when combined with *ZFP57* expression levels. Given ZFP57's role in maintenance of methylation patterns in development (Li 2008), it would also be interesting to compare how this changes in line with different expression beyond development. FAIRE or methylation-seq could be extremely informative in experiments such as these.

The combination of eQTL analysis, GWAS and other functional approaches such as DHS mapping and ChIP-seq could be used in the future to define more precisely the genes of interest in disease and their function. GWAS-identified regions of disease association can be compared to eQTL results, particularly focussing on varied expression of TFs. ChIP-seq can then be carried out on the implicated TFs in patient samples, and compared to results from healthy controls to detect binding that is specifically found in the disease context alone.

8.8 Concluding remarks

The aim of this thesis was to analyse ASE in the context of the MHC, and to relate allele-specific differences with the many autoimmune and other complex traits associated with genetic variation in this region of the genome. Validation of differential expression determined by a custom array over the MHC was carried out using qPCR for 10 genes in the MHC: *ZFP57, HLA-DQA2, HLA-DQB2, HLA-DQA1, HLA-DQB1, HLA-DPA1, HLA-DPB1, TNF* and *HLA-C*. These genes all showed variable expression in primary PBMCs enabling eQTL analysis; in the case of *HLA-DQB2* expression was so low that often it could not be quantified accurately, and therefore no eQTL analysis was performed. Many other classical MHC genes would doubtless prove to also be differentially expressed and are known to be associated with disease. However, as they were too polymorphic to be able to confidently assess

using qPCR in the healthy volunteer cohorts they were not investigated in this study.

In the analysis of *ZFP57*, two new isoforms of the gene were described, and differential expression was found in primary cells as well as the LCLs initially analysed. This is particularly interesting when considering the generally accepted role of ZFP57 in maintenance of the methylation at imprinting control centres in development. Expression of the *ZFP57* gene in human immune-related cells may have profound biological relevance to health and disease. A novel eQTL was defined for *HLA-DQB1*, which helps to resolve previously reported associations of variants with leprosy. This also highlights the manner in which eQTL studies can impact upon the findings of GWAS and other studies in terms of discovering functional information. Analysis of *HLA-C* expression in conjunction with *HCP5* expression has shown that associations found by GWAS investigating psoriasis susceptibility impact on only *HLA-C* expression, clarifying the likely functional variation. *HLA-C* expression has also been associated with HIV-1 progression, and eQTL analysis of *HLA-C* has described several SNPs that are associated with differential expression, that have also been implicated in control of HIV-1 by GWAS, thus supporting this observation.

A genome-wide analysis of KAP1 binding in LCLs has been presented together with an analysis of how this may be modulated by underlying genetic variation in the MHC. In particular differential binding of KAP1 upstream of *ZFP57* was correlated to the gene expression level of *ZFP57* in the LCLs analysed. KAP1 was shown to have a particular association with ZNF genes, supporting its role as a co-factor to many zinc finger DNA binding proteins. Initial findings of an experiment to investigate the possibilities and limitations of mRNA-seq for analysing ASE have been presented. These preliminary results show the promise of RNA-seq, especially when combined with genetic sequence information, when considering several non-MHC genes such as *GAPDH*, *NADK* and *NOTCH1*. However, low abundance transcripts are not so well defined, and the MHC presented several problems with data mapping that could not simply be overcome by using a different mapping technique of reference-free mapping. This highlights the technical difficulties of using NGS in the MHC, due to the short reads that must be mapped to a reference sequence and the highly variable nature of the MHC.

While there are almost limitless possibilities for further work related to gene expression and the MHC, trying to integrate NGS with the study of disease susceptibility in a disease context is perhaps the most exciting. As the MHC has been so frequently associated with autoimmune disease, a systemic autoimmune condition in which circulating immune cells play an important role would be an ideal scenario for such work. Different immune-related cell types would be analysed with the aim to define gene expression by RNA-seq according to genotype in affected patients thus avoiding any SNP bias found in other methods for analysis of gene expression. This could then hopefully be used to reveal both the genes and genetic variants involved in autoimmune disease susceptibility and the cells that may be the most important in disease pathogenesis.

Appendix

A.1 Primer sequences:

Primer	Amplicon (bp)	Sequence	Tm ˚C	location	Gene
ZFP57-A	83	GTCAGGTTTCTAGAAGCAGGGTAG	59.85	Exon 1-2	*ZFP57*
ZFP57-B		TCACCAGCTGCCATTGTTTA	60.26	Exon 2	*ZFP57*
ZFP57-C	128	GGTAGATAGCCCCAGGAAGG	59.92	Exon 2	*ZFP57*
ZFP57-D		AGCAATCTCTAGGTTCGATTGG	59.75	Exon 2-3	*ZFP57*
ZFP57-E	96	GAAGAAGAAGCCAGTCACCTTT	59.03	Exon 3-4	*ZFP57*
ZFP57-F		TCCTGGTAAAGGACCCTCTG	59.13	Exon 4	*ZFP57*
ZFP57-G	99	ACATCTGTGGCCAGAATCTTTC	60.5	Exon 4-5	*ZFP57*
ZFP57-H		TGTGTTTGGGAGATGGACAA	59.94	Exon 5	*ZFP57*
ZFP57-I	62	GCCTTTCAGAAGGCAAGAAG	59.19	Exon 5-6	*ZFP57*
ZFP57-J		CCCCTCATCTCTCAGACTGG	59.79	Exon 6	*ZFP57*
ZFP57-K	129	AGACCCTGTGCCCCTAAATC	60.33	Exon 1	*ZFP57*
ZFP57-F	or 615	TCCTGGTAAAGGACCCTCTG	59.13	Exon 4	*ZFP57*
ZFP57-L	95	CGTCAGAAGCCAGTCACCTT	60.44	Exon 1-4	*ZFP57*
GAPDHF	87	TCGACAGTCAGCCGCATCT	63.19	Exon 1	*GAPDH*
GAPDHR		CCGTTGACTCCGACCTTCA	62.24	Exon 2	*GAPDH*
B-Act F	106	CCAGCCTTCCTTCCTGGGC	66.5	Exon 4-5	*B-ACTIN*
B-Act R		TGTGTTGGCGTACAGGTCTTTGC	66.6	Exon 5	*B-ACTIN*
MOG-A	70	CCAGCTATGCAGGGCAGT	59.97	Exon 1	*MOG*
MOG-B		ACTTCATCCCCGACCAGAG	60.06	Exon 1-2	*MOG*
MOG-C	108	CTGGAGTGCTGGTTCTCCTC	59.99	Exon 4	*MOG*
MOG-D		CTGCTCGAAGTTTTCCTCTCA	59.74	Exon 4-5	*MOG*
TNF -A	104	AGCCCATGTTGTAGCAAACC	60	Exon 3-4	*TNF*
TNF-B		GCTGGTTATCTCTCAGCTCCA	59.59	Exon 4	*TNF*
HLA-DPB1-A	100	CTGTCACCGTGGAGTGGAA	60.73	Exon 3-4	*HLA-DPB1*
HLA-DPB1-B		AGATGATGAGCCCCAGCAC	61.22	Exon 4	*HLA-DPB1*
HLA-DQB2-A	107	AAGGACTTCTTGGAGCAGGA	59.01	Exon 2	*HLA-DQB2*
HLA-DQB2-B		GTCACTGTGGGCTCCACTT	58.66	Exon 2-3	*HLA-DQB2*
HLA-DQA2-A	81	CCAATGAGGTTCCTGAGGTC	59.51	Exon 2-3	*HLA-DQA2*
HLA-DQA2-B		CCACAAGACAGATGAGGGTGT	60.02	Exon 3	*HLA-DQA2*
HLA-DQA1-A	118	GCTGCTACCAATGAGGTTCC	59.7	Exon 2	*HLA-DQA1*
HLA-DQA1-B		TGATGTTGACCACAGGAGGA	60.09	Exon 2-3	*HLA-DQA1*
HLA-DPA1-A	109	CGCCCTGAAGACAGAATGTT	60.25	Exon 1	*HLA-DPA1*
HLA-DPA1-B		ACACATGGTCCGCCTTGAT	61.37	Exon 1-2	*HLA-DPA1*
HLA-DQB1-A	127	AGGCGCTGGCTGTGAC	59.62	Exon 5	*HLA-DQB1*
HLA-DQB1-B		TGCTTCTCTTGAGCAGTCTGA	59.02	Exon 5	*HLA-DQB1*
HLA-C-A	69	GAGGAAGAGCTCAGGTGGAA	59.53	Exon 8	*HLA-C*
HLA-C-B		GAGCCCTGGGCACTGTT	59.77	Exon 6-7	*HLA-C*
BTN3A2-A	76	CAGGACCCCTTCTTCAGGA	60.18	Exon 5-6	*BTN3A2*
BTN3A2-B		GAAGCAGCAGCAAGATAGGC	60.26	Exon 6	*BTN3A2*
HCP5-A	86	GTTGCGGGTCATGGAGTC	60.05	Exon 1	*HCP5*
HCP5-B		TGTAATTGTAATCTGCCCAGGTC	60.25	Exon 1-2	*HCP5*

Table A.1 Primer sequences used in real-time qPCR gene expression analysis

Primer	Amplicon (bp)	Sequence	Tm ˚C	location	Gene
ZFP57-5'-RACE	NA	ATGCATGCGTCTGTGATAGC	60.24	Exon 6	*ZFP57*
ZFP57-3'-RACE	NA	GTCAGGTTTCTAGAAGCAGGGTAG	59.85	Exon1-2	*ZFP57*

Table A.2 Primers sequences for RACE PCR

Primer	Amplicon (bp)	Sequence	Tm ˚C	location	Gene
ZFP57-NotI-F	1577	GCGGCCGCGAGAAGATGTTTGAACAGCTGAAG	80.1	Exon 2	*ZFP57*
ZFP57 KpnI-R		GGTACCGGCCATTTATTTATGTTTCAAG	66.3	Exon 6	*ZFP57*

Table A.3 Primer sequences used for cloning

Primer	Amplicon (bp)	Sequence	Tm ˚C	location	Gene
rs2269423-F	146	CGCCAATTGTAGAGCAGTCA	60.01	Exon 1	*AGPAT1*
rs2269423-R		CTTGACGGAAAATGGTGGTT	59.83	Intron 1-2	*AGPAT1*
rs9264942-F	280	ATTCAGTTTGGGGCCTGG	60.86	35kb Upstream	*HLA-C*
rs9264942-R		GATAGGAGGAAGGGGACCTG	59.89	35kb Upstream	*HLA-C*
M13-F	variable	TGTAAAACGACGGCCAGT	53.6	NA	NA
M13-R		CAGGAAACAGCTATGACC	53.8	NA	NA

Table A.4 Primer sequences for Sanger sequencing

Primer	Amplicon (bp)	Sequence	Tm ˚C	location	Gene
ZNF556-AF	73	CCCGTGTGGTTAAGTTTCTCACA	63.4	Intron 1-2	*ZNF556*
ZNF556-AR		GGGCTAAGCCCTGGAATCC	63.1	Intron 1-2	*ZNF556*
ZNF556-BF	60	CAAGACCCGCTCCTTCCA	62.3	Intron 1-2	*ZNF556*
ZNF556-BR		AAATTGTTCTGGGTGGGAGAAA	61.9	Intron 1-2	*ZNF556*
ZNF556-CF	61	GGCAGACTGCAGGATCTTGTG	63.3	Intron 1-2	*ZNF556*
ZNF556-CR		CACAGCACTGCGGAACTGA	62.3	Intron 1-2	*ZNF556*
ZNF556-DF	88	CTTTCTTTGCAGATGCCTTCCT	62.1	Intron 2-3/Exon 3	*ZNF556*
ZNF556-DR		GAAATAGACCCACTGGCTTTAAGC	61.6	Intron 2-3/Exon 3	*ZNF556*
ZNF556-EF	68	AAGACGGTACGAATGCAGTCAGT	62.3	Exon 4	*ZNF556*
ZNF556-ER		TTTTGTGCCTTATCAGGGATGA	61.7	Exon 4	*ZNF556*
ZNF554-DF	97	CCAAAACTCTGTACCCGTTAAATTC	61.2	Intron 4-5	*ZNF554*
ZNF554-DR		AGGGTAGGCAAATCCATAGAGATG	61.8	Intron 4-5	*ZNF554*
ZNF554-EF	65	CACTTTCCCCAGCACAAACC	62.8	Exon 5	*ZNF554*
ZNF554-ER		AAAAAGTAAATACCCCGTCCATGA	61.8	Exon 5	*ZNF554*
ZNF554-FF	62	AAGCTTGGCATCCGATTCAG	62.6	Exon 6	*ZNF554*
ZNF554-FR		AAGGTCTCCGATCTGCATATCC	61.7	Exon 6	*ZNF554*
ZNF554-GF	68	CACTACGTCCAGTAAGCTCCAGACT	62.3	Exon 6	*ZNF554*
ZNF554-GR		ATCCCATCCATGACAGCATTC	62	Exon 6	*ZNF554*
18S-F	201	GGGAATCAGGGTTCGATTCC	62.8	Unknown	*Unknown*
18S-R		CCAGACTTGCCCTCCAATGG	65.1	Unknown	*Unknown*

Table A.5 Primer sequences for KAP1 ChIP positive control (Groner 2010) and the 18S normalising primer set (De Gobbi 2006).

SNP	forward primer	reverse primer	extension primer
rs17843603	ACGTTGGATGAITGTTAATCAGGAAATGGG	ACGTTGGATGGCTTGCCAAGGAACTGCTIT	AGGTCCTGIATTAAATTCAGTTT
rs9273410	ACGTTGGATGGGIAAGTCACATCIATCAAG	ACGTTGGATGGCTTCCTTTGIACCATGAAC	CCTATTTGGGITTCTATCATTTAGA
rs9273440	ACGTTGGATGCTTGAGGTGACCCAGCIAAC	ACGTTGGATGGAACCATGAATGATGCIAC	AGTGTCTTTTCAAGGTACAA
rs3134970	ACGTTGGATGTCCTTTGTTGGGCATCTTTT	ACGTTGGATGGATATTCAAAACATCTGAGG	TGGGCATCTTTTAATTTTTATGAAG
rs3135006	ACGTTGGATGCTGCACAGAAICIGATCCAA	ACGTTGGATGGTACAATCIITGGCTTCAGG	AGATCIAAATTGTATTTATAGTAGTC

Table A.6: Primers for sequenom MassArray genotyping of *HLA-DQB1* associated SNPs in a leprosy cohort

A.2 Cycling conditions

qPCR

Cell line gene expression

95˚C for 3 minutes and then 30 seconds at 95˚C, 30 seconds at 61.5˚C, 30 seconds at 72˚C, for 50 cycles, and then 1 minute at 95˚C, 1 minute at 55˚C and 10 minutes at 80˚C.

Volunteer cohort gene expression

95˚C for 3 minutes and then 10 seconds at 95˚C, 15 seconds at 61.5˚C, 20 seconds at 72˚C, for 50 cycles, and then 1 minute at 95˚C, 1 minute at 55˚C and 10 minutes at 80˚C.

RACE PCR (touchdown PCR)

94˚C for 2 minutes and then; (94˚C for 30 seconds then 72˚C for 2 minutes) repeated for 5 cycles followed by (94˚C for 30 seconds then 70˚C, 2 minutes) repeated for 5 cycles followed by (94˚C for 30 seconds, then 65˚C for 30 seconds, then 68˚C for 2 minutes) repeated for 25 cycles, reducing the annealing temperature by 1˚C every 5 cycles, finally 68˚C for 10 minutes.

gDNA amplification for sequencing

94˚C for 2 minutes and then; (94˚C for 30 seconds, then 60˚C for 30 seconds, then 72˚C for 1 minute) repeated for 30 cycles finishing with 10 minutes at 72˚C.

Sanger sequencing PCR

96˚C for 10 seconds followed by 50˚C for 5 seconds followed by 60˚C for 4 minutes, all repeated for 25 cycles.

A.3 Supplementary data: Chapter 4

A.3.1 *BTN3A2* and *ZFP57* expression associated SNPs

SNP	*BTN3A2* p value	Rank in BTN3A2 association	*ZFP57* p value	rank in *ZFP57* association
rs2517911	$1.04x10^{-7}$	1	$2.32x10^{-25}$	25
rs1736922	$1.04x10^{-7}$	2	$2.32x10^{-25}$	26
rs1628578	$1.04x10^{-7}$	3	$2.32x10^{-25}$	27
rs1632962	$1.75x10^{-7}$	4	$8.21x10^{-28}$	24
rs13198716	$4.02x10^{-7}$	5	$1.04x10^{-7}$	426
rs9258122	$6.02x10^{-7}$	6	$3.72x10^{-31}$	15
rs13207082	$6.91x10^{-7}$	7	$1.16x10^{-7}$	421
rs3129066	$7.51x10^{-7}$	8	$5.76x10^{-31}$	16
rs2107203	$7.51x10^{-7}$	9	$5.76x10^{-31}$	17
rs3129055	$7.51x10^{-7}$	10	$5.76x10^{-31}$	18
rs375984	$8.09x10^{-6}$		$1.08x10^{-41}$	1
rs2535238	$8.09x10^{-6}$		$1.08x10^{-41}$	2
rs2747421	$8.09x10^{-6}$		$1.08x10^{-41}$	3
rs387642	$6.56x10^{-6}$		$2.05x10^{-41}$	4
rs2747429	$6.24x10^{-6}$		$6.57x10^{-41}$	5
rs3117289	$7.21x10^{-6}$		$3.36x10^{-38}$	6
rs3129073	$7.21x10^{-6}$		$3.36x10^{-38}$	7
rs29228	$1.29x10^{-5}$		$1.37x10^{-37}$	8

Table A.7: SNPs associated with both *ZFP57* and *BTN3A2* expression. The table shows the most associated SNPs with *BTN3A2* expression and their p values of association with both *BTN3A2* and *ZFP57* expression, followed by the most associated SNPs with *ZFP57* expression and their p values of association with both *BTN3A2* and *ZFP57* expression.

A.3.2 SNPs predicted to affect splicing in *ZFP57*

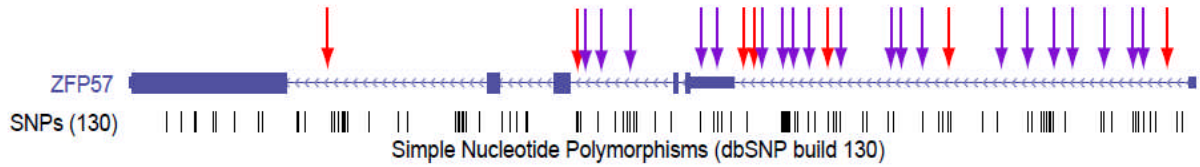


Figure A.1: Schematic showing positions of splice factor binding sites associated with sequence variation in the COX haplotype. Red arrows indicate deletions in QBL/PGF compared with the splice factor binding site seen in COX, while purple arrows show the splice factor binding sites that contain a SNP when comparing COX to QBL/ PGF.

Splice Factor	Number of new binding sites	Number at unique location
SC35	13	13
SF2/ASF	10	9
SF2/ASF (IgM-BRCA1	28	16
SRp40	9	8
SRp55	23	4

Table A.8: Splice factors with binding sites at SNPs in *ZFP57*. The numbers of binding sites for each splice factor that are created by a SNP are shown. Where the same SNP creates more than one binding site for a particular splice factor, the result is pooled to give the number at a unique location.

A.3.3 Volunteer cohort validation and QC data

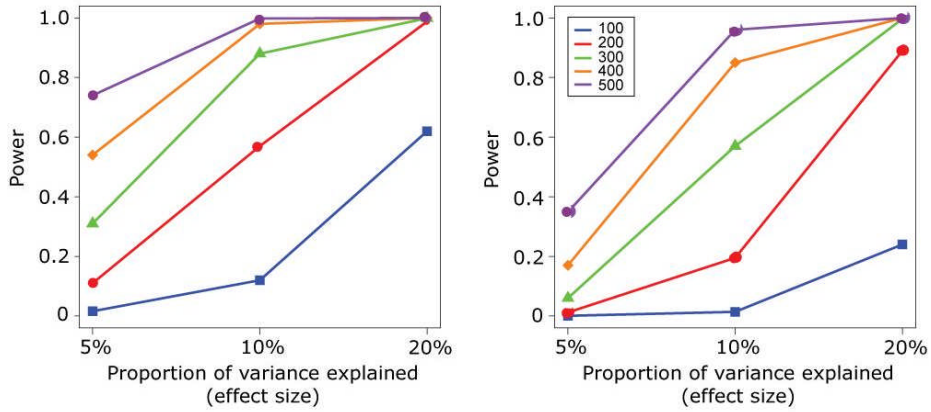


Figure A.2: Power calculations for eQTL analysis based on an additive model that assumes the marker and causal SNP are in perfect LD. i) To detect cis-eQTL associations, a sample size of 300 individuals is estimated to have more than 80% power to detect an expression-associated SNP explaining 10% of the variance in mRNA levels of a given gene. This assumes a significance threshold of $1x10^{-5}$, calculated from empirical thresholds based on published datasets (Stranger 2007). This threshold is the mean of all per-gene significance thresholds corresponding to the 0.001 tail of the permutation p-value distribution derived from 10K permutations for a given gene. i) To detect trans-eQTL associations, a genome-wide significance threshold of $5x10^{-8}$ is selected based on effective number of independent tests that would be performed using a genome-wide SNP panel. Power to detect trans-eQTLs is lower than for cis-acting variants, but a sample size of 300 will have approximately 80% power to detect an expression-associated SNP explaining 15% of the variance in mRNA levels of a given gene (Personal communication, Barbara Stranger).

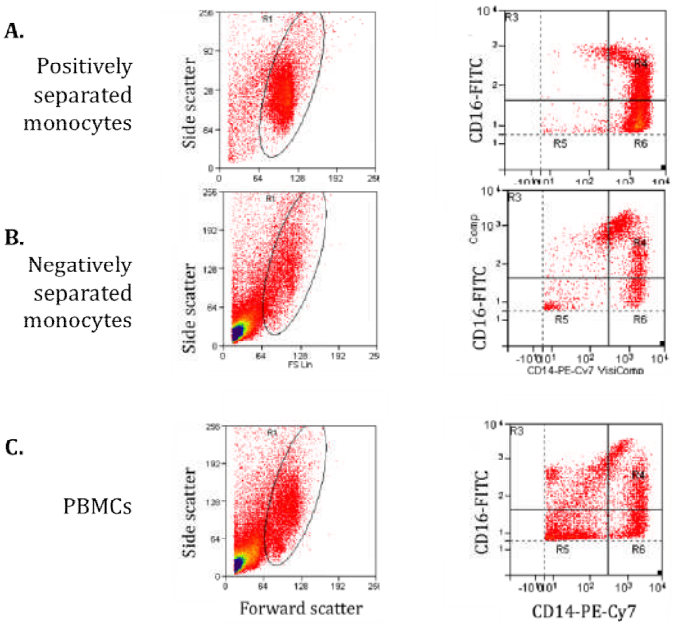


Figure A.3: Flow cytometry analysis of monocytes isolated from PBMCs through either positive or negative selection with a non-sorted PBMC mixture control (Courtesy of Seiko Makino). A) The rationale of using positive selection to sort particular cell types from PBMCs can be seen with the marked increase in cell purity demonstrated by positive selection. B) Negative selection of monocytes results in a more heterogeneous population and preserves dead cells and debris. C) PBMC control of unsorted cells.

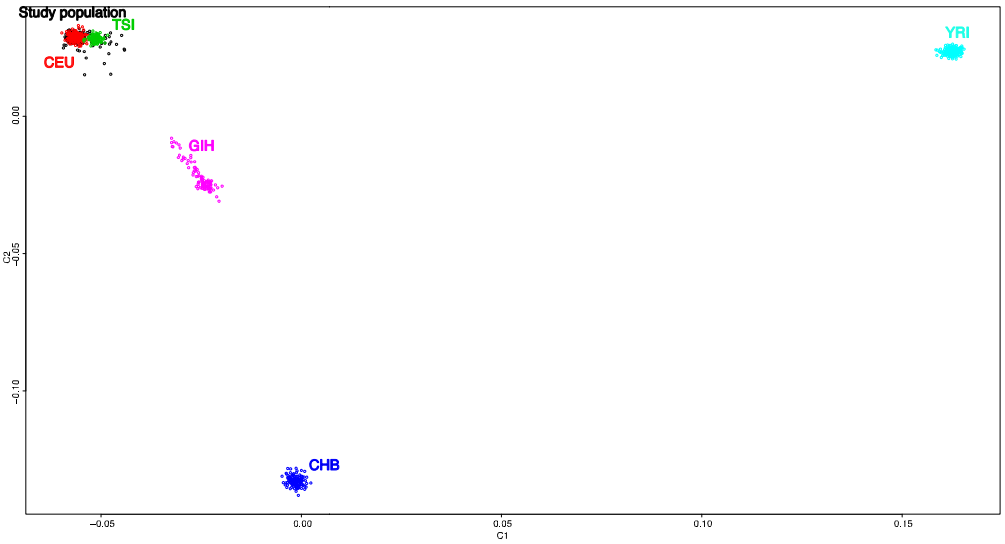


Figure A.4: MDS analysis of the volunteer cohort population (courtesy Seiko Makino). Genetic stratification in the volunteer cohort was analysed using MDS analysis comparing with the published HapMap populations. CEU (Utah residents with Northern and Western European ancestry from the CEPH collection), TSI (Toscani in Italia), CHB (Han Chinese in Beijing), YRI (Yoruba in Ibadan, Nigeria), and GIH (Gujarati Indians in Houston, Texas) were analysed using pairwise identity by state (IBS) distance and autosomal SNPs that are not on same LD (R^2<0.2). The volunteer cohort is shown in black, CEU in red, TSI in green, CHB in blue, YRI in light blue, and GIH in pink.

A.3.4 HLA Type imputation QC

As described in Chapter 3, HLA type imputation was performed by Stephen Leslie and Alexander Dilthey following a previously published method (Leslie 2008).

The 2-digit HLA types called were validated using 4 criteria; sensitivity, i.e. the proportion of true calls for each allele, specificity, i.e. is an allele called even when not present (false positive), R2, the correlation between the known and imputed alleles and PPV, proportion of accuracy of the imputed allele call. Generally specificity is good, as there are many different alleles. The HLA type was said to be called confidently when R2 was greater than 0.7, showing high correlation. This was the case for all imputed 2-digit HLA types. PPV is the best indicator of accuracy, and this was visualised as shown below in the example plot for HLA-DR 2-digit and 4-digit type. As the 4-digit types were less accurate, particularly for HLA-DRB imputation, the 2-digit types were used to analyse HLA type effect on gene expression in the new healthy volunteer cohort. The frequency of all the HLA alleles at each locus was compared to the training data, and in general frequencies remained similar between the healthy volunteer cohort and the training data. QC plots for HLA-DRB and HLA-C are shown below, confirming the performance of the imputation in relation to the number of available training alleles. A plot for HLA-DRB also confirms the greater sensitivity and performance of the imputation of 2-digit alleles, compared to 4-digit alleles, especially in the case of HLA-DRB. This supports the decision to use 2-digit HLA-type information to assess effects of alleles in the MHC on gene expression.

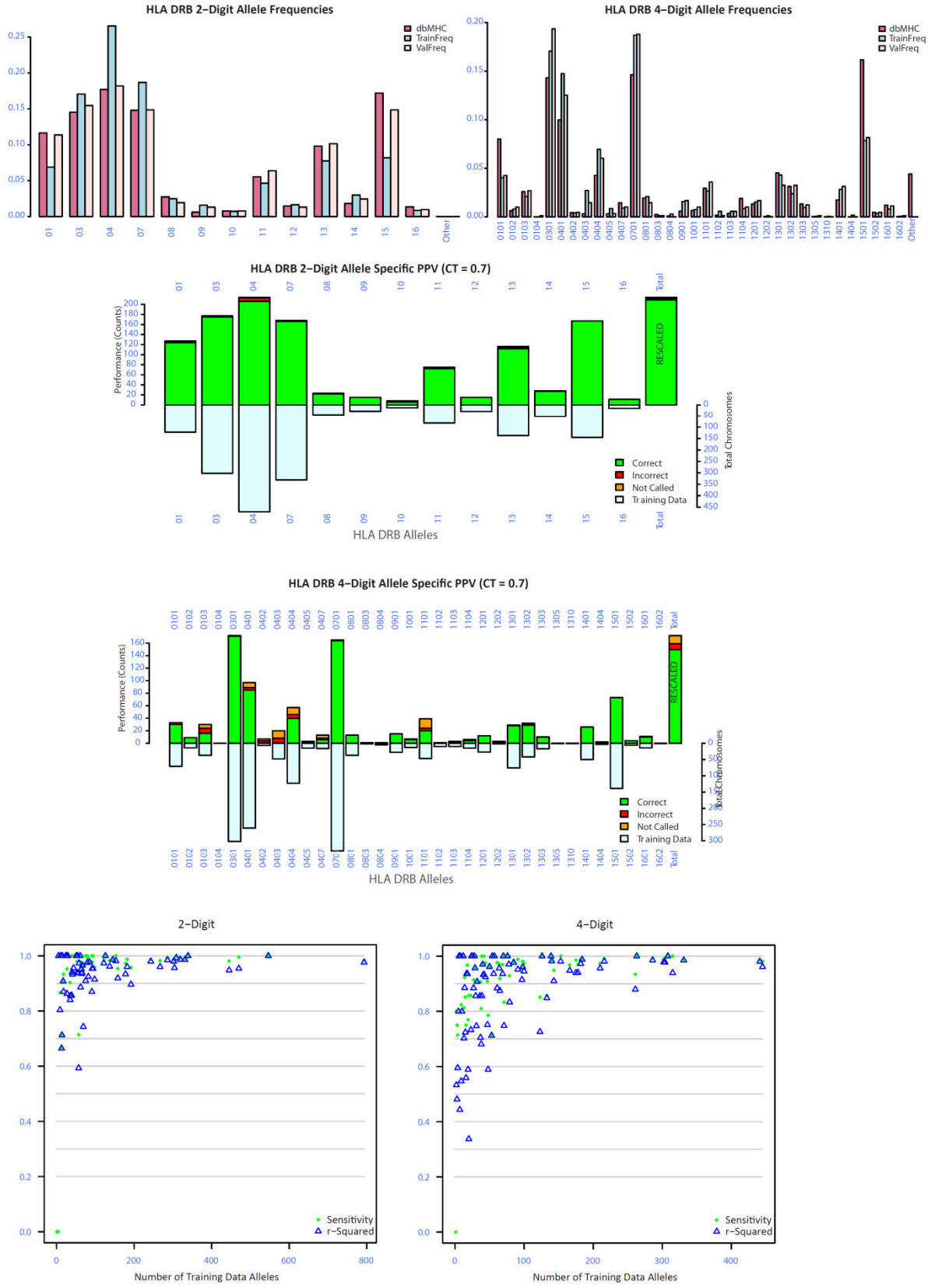


Figure A.5: HLA-DRB allele imputation QC. Frequencies of each HLA-DRB allele are shown for the training data set and the healthy volunteers. The proportion of accurate calls made when carrying out imputation of HLA-DRB 2-digit and 4-digit types is shown (PPV plots). Green bars indicate correct calls, while red bars indicate incorrect calls. Orange bars show where an allele was not called. HLA-DRB 4-digit types show the highest levels of incorrect or uncalled alleles across the 6 loci, however when the HLA-DRB 2-digit type is imputed the alleles are predicted with high accuracy.

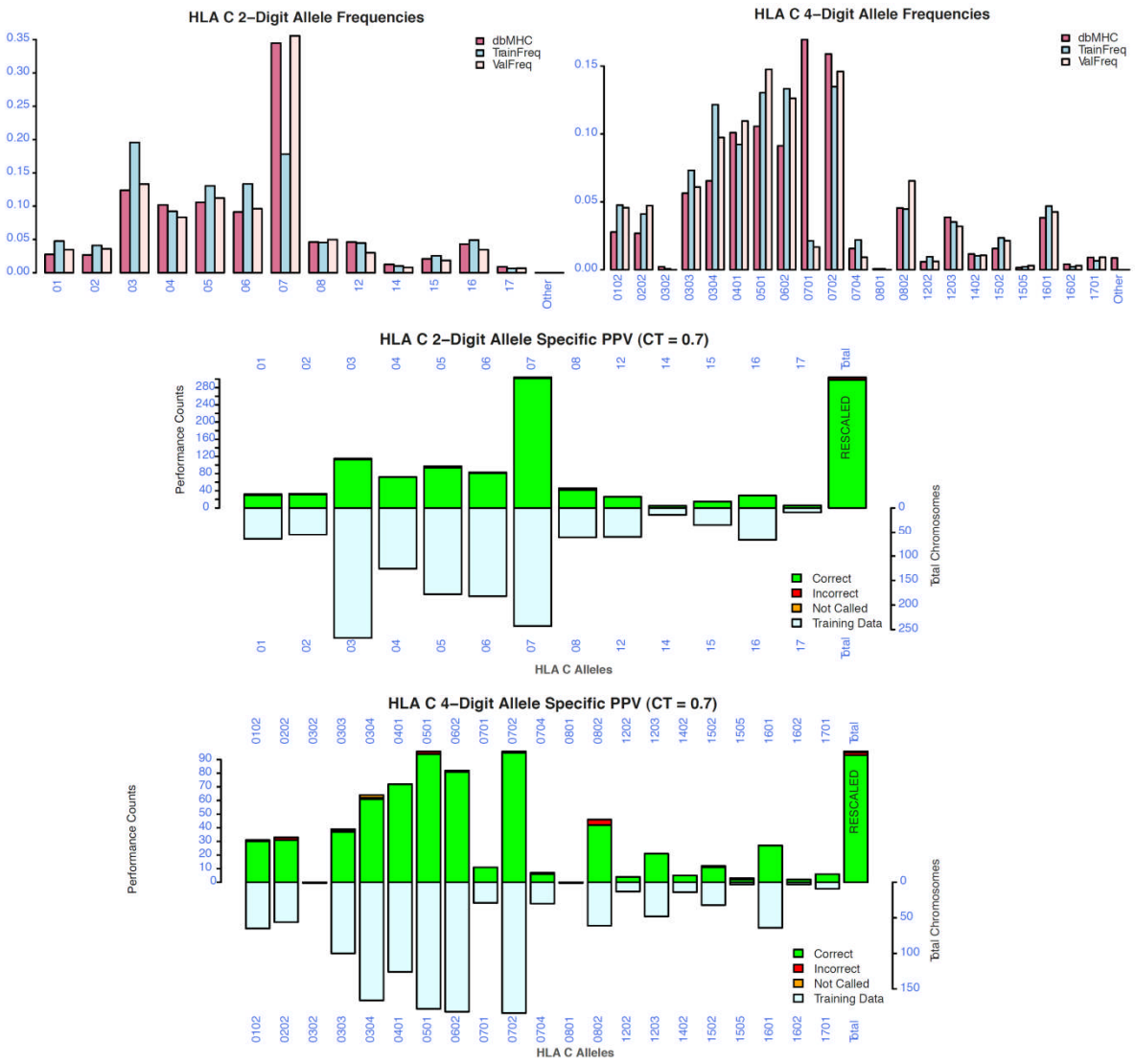


Figure A.6: HLA-C allele imputation QC. Frequencies of each HLA-C allele are shown for the training data set and the healthy volunteers. The proportion of accurate calls made when carrying out imputation of HLA-C 2-digit and 4-digit types is shown (PPV plots). Green bars indicate correct calls, while red bars indicate incorrect calls. Orange bars show where an allele was not called. In contrast to the HLA-DRB 4-digit calls, HLA-C imputation tends to be quite accurate, with only the HLA-C*0802 allele incorrectly called several times. 2-digit imputation for HLA-C shows high accuracy across all alleles.

A.4 Supplementary data: Chapter 5

A.4.1 Haplotype reconstruction over HLA-DQB1 associated SNPs

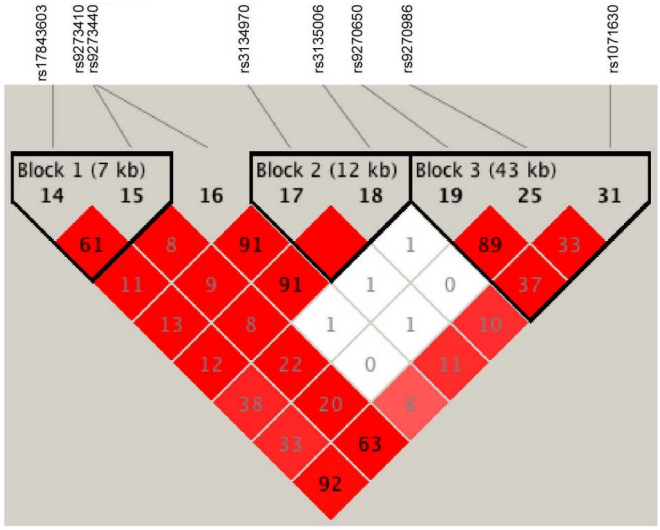


Figure A.7: Haplotype reconstruction in an Indian control cohort (Courtesy of Dr Tom Parks). LD is calculated using the Confidence Intervals method, colours of the squares indicate D'/LOD relationships where bright red is D' = 1 and LOD ≥ 2, and numbers indicated within then are the r^2 value x 100. The GCC haplotype of SNPs rs17843603, rs9273410 and rs9273440 can be seen on the left-hand side. rs107843603 was found to be in near direct linkage with rs11071630 (r^2 = 0.92, D' = 0.98).

A.4.2 *HCP5* expression and eQTL analysis

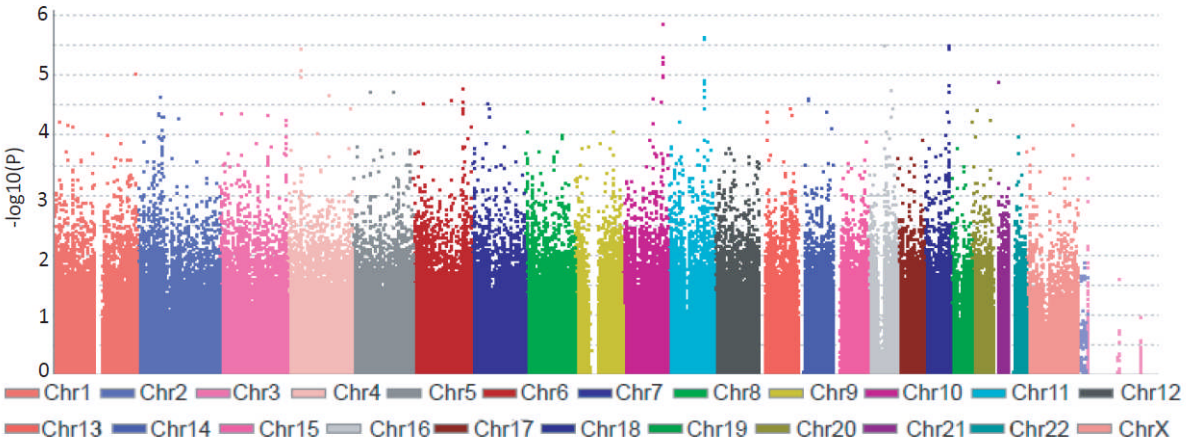


Figure A.8: Manhattan Plot showing the association of SNPs with *HCP5* expression. *HCP5* expression in 288 healthy volunteer PBMC samples was determined using qPCR and normalised to *ACTB*. Genotyping of gDNA was performed using the Illumina Infinium high-density genotyping bead arrays and association analysis was performed using PLINK.

No association was seen at a genome-wide significance threshold between genetic variation and *HCP5* expression. *HCP5* may either effect the HIV-1 and psoriasis susceptibility at the protein level or may be an artefact of the HLA-B*57 haplotype or there may be a further unidentified possibility; however, it is hard to conclusively determine the functional mechanism for this association.

A.5 Supplementary data: Chapter 7

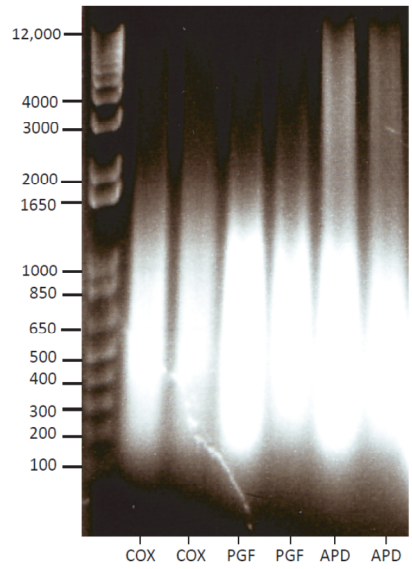


Figure A.9: Sonication of LCL chromatin. LCLs PGF, COX and APD are sonicated for 18 cycles (30 seconds on, 30 seconds off) using the Bioruptor sonication system. A sample of chromatin is taken and the cross-links reversed before the DNA is phenol: chloroform extracted and analysed on an agarose gel to assess the fragment size. Lane 1 shows 1kb+ ladder and size of fragments is indicated (bp); 6 lanes of sonicated LCL chromatin show 2 biological replicates of each LCL.

Bibliography

Akagi, Usuda, Matsuda, Ko, Niwa, Asano, Koide and Yokota (2005). Identification of Zfp-57 as a downstream molecule of STAT3 and Oct-3/4 in embryonic stem cells. **Biochem Biophys Res Commun** 331, 23-30

Alarcon-Segovia, Alarcon-Riquelme, Cardiel, Caeiro, Massardo, Villa and Pons-Estel (2005). Familial aggregation of systemic lupus erythematosus, rheumatoid arthritis, and other autoimmune diseases in 1,177 lupus patients from the GLADEL cohort. **Arthritis Rheum** 52, 1138-1147

Alberts, Terpstra, Li, Breitling, Nap and Jansen (2007). Sequence polymorphisms cause many false cis eQTLs. **PLoS One** 2, e622

Allamargot and Gardinier (2007). Alternative isoforms of myelin/oligodendrocyte glycoprotein with variable cytoplasmic domains are expressed in human brain. **J Neurochem** 101, 298-312

Allanore, Saad, Dieude, Avouac, Distler, Amouyel, Matucci-Cerinic, Riemekasten, Airo, Melchers*, et al.* (2011). Genome-wide scan identifies TNIP1, PSORS1C1, and RHOB as novel risk loci for systemic sclerosis. **PLoS Genet** 7, e1002091

Alonso, Zoidl, Taveggia, Bosse, Zoidl, Rahman, Parmantier, Dean, Harris, Wrabetz*, et al.* (2004). Identification and Characterization of ZFP-57, a Novel Zinc Finger Transcription Factor in the Mammalian Peripheral Nervous System. **J. Biol. Chem.** 279, 25653-25664

Altshuler, Daly and Lander (2008). Genetic Mapping in Human Disease. **Science** 322, 881-888

Altshuler, Gibbs, Peltonen, Dermitzakis, Schaffner, Yu, Bonnen, de Bakker, Deloukas, Gabriel*, et al.* (2010). Integrating common and rare genetic variation in diverse human populations. **Nature** 467, 52-58

Altshuler, TIHMP. and TIHMP (2005). A haplotype map of the human genome. **Nature** 437, 1299-1320

Aly, Ide, Jahromi, Barker, Fernando, Babu, Yu, Miao, Erlich, Fain*, et al.* (2006). Extreme genetic risk for type 1A diabetes. **Proc Natl Acad Sci U S A** 103, 14074-14079

Anderson, Boucher, Lees, Franke, D'Amato, Taylor, Lee, Goyette, Imielinski, Latiano*, et al.* (2011). Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. **Nat Genet** 43, 246-252

Anderson, Pettersson, Barrett, Zhuang, Ragoussis, Cardon and Morris (2008). Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. **Am J Hum Genet** 83, 112-119

Asano, Matsushita, Umeno, Hosono, Takahashi, Kawaguchi, Matsumoto, Matsui, Kakuta, Kinouchi*, et al.* (2009). A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. **Nat Genet** 41, 1325-1329

Auer, Srivastava and Doerge (2011). Differential expression - the next generation and beyond. **Briefings in Functional Genomics**

Bahlo M (2009). Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. **Nat Genet** 41, 824-828

Ballana, Senserrich, Pauls, Faner, Mercader, Uyttebroeck, Palou, Mena, Grau, Clotet*, et al.* (2010). ZNRD1 (zinc ribbon domain-containing 1) is a host cellular factor that influences HIV-1 replication and disease progression. **Clin Infect Dis** 50, 1022-1032

Barcellos, Oksenberg, Begovich, Martin, Schmidt, Vittinghoff, Goodin, Pelletier, Lincoln, Bucher*, et al.* (2003). HLA-DR2 dose effect on susceptibility to multiple sclerosis and influence on disease course. **Am J Hum Genet** 72, 710-716

Barrett and Cardon (2006). Evaluating coverage of genome-wide association studies. **Nat Genet** 38, 659-662

Barrett, Clayton, Concannon, Akolkar, Cooper, Erlich, Julier, Morahan, Nerup, Nierras*, et al.* (2009a). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. **Nat Genet** 41, 703-707

Barrett, Fry, Maller and Daly (2005). Haploview: analysis and visualization of LD and haplotype maps. **Bioinformatics** 21, 263-265

Barrett, Hansoul, Nicolae, Cho, Duerr, Rioux, Brant, Silverberg, Taylor, Barmada*, et al.* (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. **Nat Genet** 40, 955-962

Barrett, Lee, Lees, Prescott, Anderson, Phillips, Wesley, Parnell, Zhang, Drummond*, et al.* (2009b). Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. **Nat Genet** 41, 1330-1334

Barski and Zhao (2009). Genomic location analysis by ChIP-Seq. **Journal of Cellular Biochemistry** 107, 11-18

Bayley, Ottenhoff and Verweij (2004). Is there a future for TNF promoter polymorphisms? **Genes Immun** 5, 315-329

Beck (1999). Complete sequence and gene map of a human major histocompatibility complex. The MHC sequencing consortium. **Nature** 401, 921-923

Bei, Li, Jia, Feng, Zhou, Chen, Feng, Low, Zhang, He*, et al.* (2010). A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci. **Nat Genet** 42, 599-603

Bell and Beck (2009). Advances in the identification and analysis of allele-specific expression. **Genome Med** 1, 56

Bell and Beck (2010). The epigenomic interface between genome and environment in common complex diseases. **Brief Funct Genomics** 9, 477-485

Bellefroid, Marine, Ried, Lecocq, Riviere, Amemiya, Poncelet, Coulie, de Jong, Szpirer*, et al.* (1993). Clustered organization of homologous KRAB zinc-finger genes with enhanced expression in human T lymphoid cells. **EMBO J** 12, 1363-1374

Benacerraf (1981). Role of MHC gene products in immune regulation. **Science** 212, 1229-1238

Bertone, Stolc, Royce, Rozowsky, Urban, Zhu, Rinn, Tongprasit, Samanta, Weissman*, et al.* (2004). Global Identification of Human Transcribed Sequences with Genome Tiling Arrays. **Science** 306, 2242-2246

Bjornsson, Albert, Ladd-Acosta, Green, Rongione, Middle, Irizarry, Broman and Feinberg (2008). SNP-specific array-based allele-specific expression analysis. **Genome Res** 18, 771-779

Blahnik, Dou, Echipare, Iyengar, O'Geen, Sanchez, Zhao, Marra, Hirst, Costello*, et al.* (2011). Characterization of the Contradictory Chromatin Signatures at the 3' Exons of Zinc Finger Genes. **PLoS ONE** 6, e17121

Blais, Dong and Rowland-Jones (2011). HLA-C as a mediator of natural killer and T-cell activation: spectator or key player? **Immunology** 133, 1-7

Blencowe (2006). Alternative splicing: new insights from global analyses. **Cell** 126, 37-47

Boonen, Hahnemann, Mackay, Tommerup, Brondum-Nielsen, Tumer and Gronskov (2011). No evidence for pathogenic variants or maternal effect of ZFP57 as the cause of Beckwith-Wiedemann Syndrome. **Eur J Hum Genet** In Print

Borghans, Beltman and De Boer (2004). MHC polymorphism under host-pathogen coevolution. **Immunogenetics** 55, 732-739

Braciale, Morrison, Sweetser, Sambrook, Gething and Braciale (1987). Antigen presentation pathways to class I and class II MHC-restricted T lymphocytes. **Immunol Rev** 98, 95-114

Bradfield, Qu, Wang, Zhang, Sleiman, Kim, Mentch, Qiu, Glessner, Thomas*, et al.* (2011). A Genome-Wide Meta-Analysis of Six Type 1 Diabetes Cohorts Identifies Multiple Associated Loci. **PLoS Genet** 7, e1002293

Bray, Buckland, Owen and O'Donovan (2003). Cis-acting variation in the expression of a high proportion of genes in human brain. **Hum Genet** 113, 149-153

Brem, Yvert, Clinton and Kruglyak (2002). Genetic dissection of transcriptional regulation in budding yeast. **Science** 296, 752-755

Brewerton, Hart, Nicholls, Caffrey, James and Sturrock (1973). Ankylosing spondylitis and HL-A 27. **Lancet** 1, 904-907

Britten, Mijovic, Barnett and Kelly (2009). Differential expression of HLA-DQ alleles in peripheral blood mononuclear cells: alleles associated with susceptibility to and protection from autoimmune type 1 diabetes. **Int J Immunogenet** 36, 47-57

Broderick, Wang, Vijayakrishnan, Matakidou, Spitz, Eisen, Amos and Houlston (2009). Deciphering the impact of common genetic variation on lung cancer risk: a genome-wide association study. **Cancer Res** 69, 6633-6641

Bryne, Valen, Tang, Marstrand, Winther, da Piedade, Krogh, Lenhard and Sandelin (2008). JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. **Nucleic Acids Res** 36, D102-106

Buckland (2004). Allele-specific gene expression differences in humans. **Hum Mol Genet** 13 Spec No 2, R255-260

Buiting, Saitoh, Gross, Dittrich, Schwartz, Nicholls and Horsthemke (1995). Inherited microdeletions in the Angelman and Prader-Willi syndromes define an imprinting centre on human chromosome 15. **Nat Genet** 9, 395-400

Bullard, Purdom, Hansen and Dudoit (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. **BMC Bioinformatics** 11, 94

Cammas, Mark, Dolle, Dierich, Chambon and Losson (2000). Mice lacking the transcriptional corepressor TIF1beta are defective in early postimplantation development. **Development** 127, 2955-2963

Carrington, Nelson, Martin, Kissner, Vlahov, Goedert, Kaslow, Buchbinder, Hoots and O'Brien (1999). HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. **Science** 283, 1748-1752

Cassinotti, Birindelli, Clerici, Trabattoni, Lazzaroni, Ardizzone, Colombo, Rossi and Porro (2009). HLA and Autoimmune Digestive Disease: A Clinically Oriented Review for Gastroenterologists. **Am J Gastroenterol** 104, 195-217

Catano, Kulkarni, He, Marconi, Agan, Landrum, Anderson, Delmar, Telles, Song*, et al.* (2008). HIV-1 disease-influencing effects associated with ZNRD1, HCP5 and HLA-C alleles are attributable mainly to either HLA-A10 or HLA-B*57 alleles. **PLoS One** 3, e3636

Chakravarti (2011). Widespread Promiscuous Genetic Information Transfer From DNA to RNA. **Circulation Research** 109, 1202-1203

Chamary, Parmley and Hurst (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. **Nat Rev Genet** 7, 98-108

Chanda, Zhang, Sucheston and Ramanathan (2009). A two-stage search strategy for detecting multiple loci associated with rheumatoid arthritis. **BMC Proc** 3 Suppl 7, S72

Chang, Juan, Yin, Chi and Tsay (1998). Up-regulation of beta-actin, cyclophilin and GAPDH in N1S1 rat hepatoma. **Oncol Rep** 5, 469-471

Cheng, Geng, Cheng, Liang, Bai, Li, Srivastava, Ng, Fukagawa, Wu*, et al.* (2010). KRAB zinc finger protein ZNF382 is a proapoptotic tumor suppressor that represses multiple oncogenes and is commonly silenced in multiple carcinomas. **Cancer Res** 70, 6516-6526

Cheung, Conlin, Weber, Arcaro, Jen, Morley and Spielman (2003). Natural variation in human gene expression assessed in lymphoblastoid cells. **Nat Genet** 33, 422-425

Cheung, Nayak, Wang, Elwyn, Cousins, Morley and Spielman (2010). Polymorphic Cis- and Trans-Regulation of Human Gene Expression. **PLoS Biol** 8, e1000480

Cheung and Spielman (2002). The genetics of variation in gene expression. **Nat Genet** 32 Suppl, 522-525

Cheung and Spielman (2009). Genetics of human gene expression: mapping DNA variants that influence gene expression. **Nat Rev Genet** 10, 595-604

Cheung, Spielman, Ewens, Weber, Morley and Burdick (2005). Mapping determinants of human gene expression by regional and genome-wide association. **Nature** 437, 1365-1369

Chung, Taylor, Graham, Nititham, Lee, Ortmann, Jacob, Alarcon-Riquelme, Tsao, Harley*, et al.* (2011). Differential genetic associations for systemic lupus erythematosus based on anti-dsDNA autoantibody production. **PLoS Genet** 7, e1001323

Clancy, Marion, Kaufman, Ramos, Adler, Harley, Langefeld and Buyon (2010). Identification of candidate loci at 6p21 and 21q22 in a genome-wide association study of cardiac manifestations of neonatal lupus. **Arthritis Rheum** 62, 3415-3424

Comabella, Craig, Camina-Tato, Morcillo, Lopez, Navarro, Rio, Montalban and Martin (2008). Identification of a novel risk locus for multiple sclerosis at 13q31.3 by a pooled genome-wide scan of 500,000 single nucleotide polymorphisms. **PLoS One** 3, e3490

Conde, Halperin, Akers, Brown, Smedby, Rothman, Nieters, Slager, Brooks-Wilson, Agana*, et al.* (2010). Genome-wide association study of follicular lymphoma identifies a risk locus at 6p21.32. **Nat Genet** 42, 661-664

Conrad, Pinto, Redon, Feuk, Gokcumen, Zhang, Aerts, Andrews, Barnes, Campbell*, et al.* (2009). Origins and functional impact of copy number variation in the human genome. **Nature** 464, 704-712

Cookson, Liang, Abecasis, Moffatt and Lathrop (2009). Mapping complex disease traits with global gene expression. **Nat Rev Genet** 10, 184-194

Coon, Myers, Craig, Webster, Pearson, Lince, Zismann, Beach, Leung, Bryden*, et al.* (2007). A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease. **J Clin Psychiatry** 68, 613-618

Cooper, Smyth, Smiles, Plagnol, Walker, Allen, Downes, Barrett, Healy, Mychaleckyj*, et al.* (2008). Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. **Nat Genet** 40, 1399-1401

Corder, Saunders, Strittmatter, Schmechel, Gaskell, Small, Roses, Haines and Pericak-Vance (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. **Science** 261, 921-923

Costa, Angelini, De Feis and Ciccodicola (2010). Uncovering the complexity of transcriptomes with RNA-Seq. **J Biomed Biotechnol** 2010, 853916

Cotsapas, Voight, Rossin, Lage, Neale, Wallace, Abecasis, Barrett, Behrens, Cho*, et al.* (2011). Pervasive Sharing of Genetic Effects in Autoimmune Disease. **PLoS Genet** 7, e1002254

Cowles, Hirschhorn, Altshuler and Lander (2002). Detection of regulatory variation in mouse genes. **Nat Genet** 32, 432-437

Cullen, Noble, Erlich, Thorpe, Beck, Klitz, Trowsdale and Carrington (1997). Characterization of recombination in the HLA class II region. **Am J Hum Genet** 60, 397-407

D'Alfonso, Bolognesi, Guerini, Barizzone, Bocca, Ferrante, Castelli, Bergamaschi, Agliardi, Ferrante*, et al.* (2008). A sequence variation in the MOG gene is involved in multiple sclerosis susceptibility in Italy. **Genes Immun** 9, 7-15

da Silva, Mazini, Reis, Sell, Tsuneto, Peixoto and Visentainer (2009). HLA-DR and HLA-DQ alleles in patients from the south of Brazil: markers for leprosy susceptibility and resistance. **BMC Infectious Diseases** 9, 134

Daly, Rioux, Schaffner, Hudson and Lander (2001). High-resolution haplotype structure in the human genome. **Nat Genet** 29, 229-232

Darvasi (2003). Genomics: Gene expression meets genetics. **Nature** 422, 269-270

Dausset (1981). The major histocompatibility complex in man. **Science** 213, 1469-1474

Dausset, Le Brun and Sasportes (1972). [Lymphocyte mixed culture reaction (LMC) between parents and children with the same HL-A phenotype. Hypothesis of a genetic recognition system]. **C R Acad Sci Hebd Seances Acad Sci D** 275, 2279-2282

de Bakker, McVean, Sabeti, Miretti, Green, Marchini, Ke, Monsuur, Whittaker, Delgado*, et al.* (2006). A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. **Nat Genet** 38, 1166-1172

De Gobbi, Viprakasit, Hughes, Fisher, Buckle, Ayyub, Gibbons, Vernimmen, Yoshinaga, de Jong*, et al.* (2006). A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. **Science** 312, 1215-1217

De Jager, Jia, Wang, de Bakker, Ottoboni, Aggarwal, Piccio, Raychaudhuri, Tran, Aubin*, et al.* (2009). Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. **Nat Genet** 41, 776-782

Degli-Esposti, Leaver, Christiansen, Witt, Abraham and Dawkins (1992). Ancestral haplotypes: conserved population MHC haplotypes. **Human Immunology** 34, 242-252

Degner, Marioni, Pai, Pickrell, Nkadori, Gilad and Pritchard (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. **Bioinformatics** 25, 3207-3212

Di Sabatino and Corazza (2009). Coeliac disease. **The Lancet** 373, 1480-1493

Dimas, Deutsch, Stranger, Montgomery, Borel, Attar-Cohen, Ingle, Beazley, Arcelus, Sekowska*, et al.* (2009). Common Regulatory Variation Impacts Gene Expression in a Cell Type-Dependent Manner. **Science** 325, 1246-1250

Dimas, Stranger, Beazley, Finn, Ingle, Forrest, Ritchie, Deloukas, Tavare and Dermitzakis (2008). Modifier effects between regulatory and protein-coding variation. **PLoS Genet** 4, e1000244

Dixon, Liang, Moffatt, Chen, Heath, Wong, Taylor, Burnett, Gut, Farrall*, et al.* (2007). A genome-wide association study of global gene expression. **Nat Genet** 39, 1202-1207

Dubois, Trynka, Franke, Hunt, Romanos, Curtotti, Zhernakova, Heap, Adany, Aromaa*, et al.* (2010). Multiple common variants for celiac disease influencing immune gene expression. **Nat Genet** 42, 295-302

Durbin, Abecasis, Altshuler, Auton, Brooks, Durbin, Gibbs, Hurles and McVean (2010). A map of human genome variation from population-scale sequencing. **Nature** 467, 1061-1073

Eichler, Hoffman, Adamson, Gordon, McCready, Lamerdin and Mohrenweiser (1998). Complex beta-satellite repeat structures and the expansion of the zinc finger gene cluster in 19p12. **Genome Res** 8, 791-808

Eleftherohorinou, Wright, Hoggart, Hartikainen, Jarvelin, Balding, Coin and Levin (2009). Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases. **PLoS One** 4, e8068

Elgin (1988). The formation and function of DNase I hypersensitive sites in the process of gene activation. **J Biol Chem** 263, 19259-19262

Emerson and Thomas (2009). Adaptive Evolution in Zinc Finger Transcription Factors. **PLoS Genet** 5, e1000325

Emilsson, Thorleifsson, Zhang, Leonardson, Zink, Zhu, Carlson, Helgason, Walters, Gunnarsdottir*, et al.* (2008). Genetics of gene expression and its effect on disease. **Nature** 452, 423-428

Enciso-Mora, Broderick, Ma, Jarrett, Hjalgrim, Hemminki, van den Berg, Olver, Lloyd, Dobbins*, et al.* (2010). A genome-wide association study of Hodgkin's lymphoma identifies new susceptibility loci at 2p16.1 (REL), 8q24.21 and 10p14 (GATA3). **Nat Genet** 42, 1126-1130

Epstein (2009). Cis-regulatory mutations in human disease. **Brief Funct Genomic Proteomic** 8, 310-316

Evans, Spencer, Pointon, Su, Harvey, Kochan, Oppermann, Dilthey, Pirinen, Stone*, et al.* (2011). Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. **Nat Genet** 43, 761-767

Fairfax (In Press). Genetics of Gene Expression in Primary Immune Cells Defines Cell-Specific Master Regulators And Role of HLA Alleles. **Nat Genet**

Fairfax, Vannberg, Radhakrishnan, Hakonarson, Keating, Hill and Knight (2009). An integrated expression phenotype mapping approach defines common variants in LEP, ALOX15 and CAPNS1 associated with induction of IL-6. **Hum. Mol. Genet.** ddp530

Fang and Cui (2011). Design and validation issues in RNA-seq experiments. **Brief Bioinform** 12, 280-287

Fang, Lev-Lehman, Tsai, Matsuura, Benton, Sutcliffe, Christian, Kubota, Halley, Meijers-Heijboer*, et al.* (1999). The spectrum of mutations in UBE3A causing Angelman syndrome. **Hum Mol Genet** 8, 129-135

Fellay, Ge, Shianna, Colombo, Ledergerber, Cirulli, Urban, Zhang, Gumbs, Smith*, et al.* (2009). Common genetic variation and the control of HIV-1 in humans. **PLoS Genet** 5, e1000791

Fellay, Shianna, Ge, Colombo, Ledergerber, Weale, Zhang, Gumbs, Castagna, Cossarizza*, et al.* (2007). A Whole-Genome Association Study of Major Determinants for Host Control of HIV-1. **Science** 317, 944-947

Fernando, Stevens, Walsh, De Jager, Goyette, Plenge, Vyse and Rioux (2008a). Defining the Role of the MHC in Autoimmunity: A Review and Pooled Analysis. **PLoS Genet** 4, e1000024

Fernando, Stevens, Walsh, De Jager, Goyette, Plenge, Vyse and Rioux (2008b). Defining the role of the MHC in autoimmunity: a review and pooled analysis. **PLoS Genet** 4, e1000024

Ferreira, Pan-Hammarstrom, Graham, Gateva, Fontan, Lee, Ortmann, Urcelay, Fernandez-Arquero, Nunez*, et al.* (2010). Association of IFIH1 and other autoimmunity risk alleles with selective IgA deficiency. **Nat Genet** 42, 777-780

Fickett (1996). Quantitative discrimination of MEF2 sites. **Mol Cell Biol** 16, 437-441

Fitness, Tosh and Hill (2002). Genetics of susceptibility to leprosy. **Genes Immun** 3, 441-453

Franke, Balschun, Karlsen, Sventoraityte, Nikolaus, Mayr, Domingues, Albrecht, Nothnagel, Ellinghaus*, et al.* (2008). Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility. **Nat Genet** 40, 1319-1323

Franke and Jansen (2009). eQTL analysis in humans. **Methods Mol Biol** 573, 311-328

Franke, McGovern, Barrett, Wang, Radford-Smith, Ahmad, Lees, Balschun, Lee, Roberts*, et al.* (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. **Nat Genet** 42, 1118-1125

Frazer, Ballinger, Cox, Hinds, Stuve, Gibbs, Belmont, Boudreau, Hardenbol, Leal*, et al.* (2007). A second generation human haplotype map of over 3.1 million SNPs. **Nature** 449, 851-861

Frazer, Murray, Schork and Topol (2009). Human genetic variation and its contribution to complex traits. **Nat Rev Genet** 10, 241-251

Frietze, O'Geen, Blahnik, Jin and Farnham (2010). ZNF274 recruits the histone methyltransferase SETDB1 to the 3' ends of ZNF genes. **PLoS One** 5, e15082

Fussell, Thomas, Street and Darke (1996). HLA-A9 antibodies and epitopes. **Tissue Antigens** 47, 307-312

Gabriel, Schaffner, Nguyen, Moore, Roy, Blumenstiel, Higgins, DeFelice, Lochner, Faggart*, et al.* (2002a). The Structure of Haplotype Blocks in the Human Genome. **Science** 296, 2225-2229

Gabriel, Schaffner, Nguyen, Moore, Roy, Blumenstiel, Higgins, DeFelice, Lochner, Faggart*, et al.* (2002b). The structure of haplotype blocks in the human genome. **Science** 296, 2225-2229

Gaidatzis, Jacobeit, Oakeley and Stadler (2009). Overestimation of alternative splicing caused by variable probe characteristics in exon arrays. **Nucleic Acids Res** 37, e107

Ge, Pokholok, Kwan, Grundberg, Morcos, Verlaan, Le, Koka, Lam, Gagne*, et al.* (2009). Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. **Nat Genet** 41, 1216-1222

Genin, Schumacher, Roujeau, Naldi, Liss, Kazma, Sekula, Hovnanian and Mockenhaupt (2011). Genome-wide association study of Stevens-Johnson Syndrome and Toxic Epidermal Necrolysis in Europe. **Orphanet J Rare Dis** 6, 52

Gharavi, Kiryluk, Choi, Li, Hou, Xie, Sanna-Cherchi, Men, Julian, Wyatt*, et al.* (2011). Genome-wide association study identifies susceptibility loci for IgA nephropathy. **Nat Genet** 43, 321-327

Gibson and Weir (2005). The quantitative genetics of transcription. **Trends in Genetics** 21, 616-623

Giresi, Kim, McDaniell, Iyer and Lieb (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. **Genome Res** 17, 877-885

Glazier, Nadeau and Aitman (2002). Finding Genes That Underlie Complex Traits. **Science** 298, 2345-2349

Goldberg, Arnett, Bias and Shulman (1976). Histocompatibility antigens in systemic lupus erythematosus. **Arthritis Rheum** 19, 129-132

Goldstein (2009). Common genetic variation and human traits. **N Engl J Med** 360, 1696-1698

Goring, Curran, Johnson, Dyer, Charlesworth, Cole, Jowett, Abraham, Rainwater, Comuzzie*, et al.* (2007). Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. **Nat Genet** 39, 1208-1216

Gorlova, Martin, Rueda, Koeleman, Ying, Teruel, Diaz-Gallo, Broen, Vonk, Simeon*, et al.* (2011). Identification of novel genetic markers associated with clinical phenotypes of systemic sclerosis through a genome-wide association strategy. **PLoS Genet** 7, e1002178

Graveley (2008). The haplo-spliceo-transcriptome: common variations in alternative splicing in the human population. **Trends Genet** 24, 5-7

Green and Jabri (2006). Celiac disease. **Annu Rev Med** 57, 207-221

Greenwood and Kelsoe (2003). Promoter and intronic variants affect the transcriptional regulation of the human dopamine transporter gene. **Genomics** 82, 511-520

Gregersen, Silver and Winchester (1987). The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. **Arthritis Rheum** 30, 1205-1213

Groner, Meylan, Ciuffi, Zangger, Ambrosini, DÃ©nervaud, Bucher and Trono (2010). KRAB Zinc Finger Proteins and KAP1 Can Mediate Long-Range Transcriptional Repression through Heterochromatin Spreading. **PLoS Genet** 6, e1000869

Grundberg, Adoue, Kwan, Ge, Duan, Lam, Koka, Kindmark, Weiss, Tantisira*, et al.* (2011). Global Analysis of the Impact of Environmental Perturbation on *cis*-Regulation of Gene Expression. **PLoS Genet** 7, e1001279

Hafler, Compston, Sawcer, Lander, Daly, De Jager, de Bakker, Gabriel, Mirel, Ivinson*, et al.* (2007). Risk alleles for multiple sclerosis identified by a genomewide study. **N Engl J Med** 357, 851-862

Hakonarson, Grant, Bradfield, Marchand, Kim, Glessner, Grabs, Casalunovo, Taback, Frackelton*, et al.* (2007). A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. **Nature** 448, 591-594

Hall (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. **Nucleic Acids Symposium Series** 41, 95-98

Han, Zheng, Cui, Sun, Ye, Hu, Xu, Cai, Huang, Zhao*, et al.* (2009). Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. **Nat Genet** 41, 1234-1237

Hanifi Moghaddam, de Knijf, Roep, Van der Auwera, Naipal, Gorus, Schuit and Giphart (1998). Genetic structure of IDDM1: two separate regions in the major histocompatibility complex contribute to susceptibility or protection. Belgian Diabetes Registry. **Diabetes** 47, 263-269

Hansen, Brenner and Dudoit (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. **Nucleic Acids Res** 38, e131

Hao, Chudin, McElwee and Schadt (2009). Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. **BMC Genet** 10, 27

Harley, Alarcon-Riquelme, Criswell, Jacob, Kimberly, Moser, Tsao, Vyse, Langefeld, Nath*, et al.* (2008). Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. **Nat Genet** 40, 204-210

Heap and van Heel (2009). The genetics of chronic inflammatory diseases. **Hum Mol Genet** 18, R101-106

Heap, Yang, Downes, Healy, Hunt, Bockett, Franke, Dubois, Mein, Dobson*, et al.* (2010). Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. **Hum Mol Genet** 19, 122-134

Heintzman, Hon, Hawkins, Kheradpour, Stark, Harp, Ye, Lee, Stuart, Ching*, et al.* (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. **Nature** 459, 108-112

Heinzen, Ge, Cronin, Maia, Shianna, Gabriel, Welsh-Bohmer, Hulette, Denny and Goldstein (2008). Tissue-specific genetic control of splicing: implications for the study of complex traits. **PLoS Biol** 6, e1

Hellman and Chess (2010). Extensive sequence-influenced DNA methylation polymorphism in the human genome. **Epigenetics Chromatin** 3, 11

Hill (2006). Aspects of genetic susceptibility to human infectious diseases. **Annu Rev Genet** 40, 469-486

Hirasawa and Feil (2008). A KRAB domain zinc finger protein in imprinting and disease. **Dev Cell** 15, 487-488

Hirota, Takahashi, Kubo, Tsunoda, Tomita, Doi, Fujita, Miyatake, Enomoto, Miyagawa*, et al.* (2011). Genome-wide association study identifies three new susceptibility loci for adult asthma in the Japanese population. **Nat Genet** 43, 893-896

Hirschhorn and Daly (2005). Genome-wide association studies for common diseases and complex traits. **Nat Rev Genet** 6, 95-108

Holcomb, Hoglund, Anderson, Blake, Bohme, Egholm, Ferriola, Gabriel, Gelber, Goodridge*, et al.* (2011). A multi-site study using high-resolution HLA genotyping by next generation sequencing. **Tissue Antigens** 77, 206-217

Hom, Graham, Modrek, Taylor, Ortmann, Garnier, Lee, Chung, Ferreira, Pant*, et al.* (2008). Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX. **N Engl J Med** 358, 900-909

Horton, Gibson, Coggill, Miretti, Allcock, Almeida, Forbes, Gilbert, Halls, Harrow*, et al.* (2008). Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. **Immunogenetics** 60, 1-18

Horton, Wilming, Rand, Lovering, Bruford, Khodiyar, Lush, Povey, Talbot, Wright*, et al.* (2004). Gene map of the extended human MHC. **Nat Rev Genet** 5, 889-899

Howie, Donnelly and Marchini (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. **PLoS Genet** 5, e1000529

Hudson (2003). Wanted: regulatory SNPs. **Nat Genet** 33, 439-440

Hull, Campino, Rowlands, Chan, Copley, Taylor, Rockett, Elvidge, Keating, Knight*, et al.* (2007). Identification of Common Genetic Variation That Modulates Alternative Splicing. **PLoS Genet** 3, e99

Hunt, Zhernakova, Turner, Heap, Franke, Bruinenberg, Romanos, Dinesen, Ryan, Panesar*, et al.* (2008). Newly identified genetic risk variants for celiac disease related to the immune response. **Nat Genet** 40, 395-402

Huntley, Baggott, Hamilton, Tran-Gyamfi, Yang, Kim, Gordon, Branscomb and Stubbs (2006). A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. **Genome Res** 16, 669-677

Hutchison, Stallings, McGeary and Bryan (2004). Population stratification in the candidate gene study: fatal threat or red herring? **Psychol Bull** 130, 66-79

Ihle, Fleckenstein, Terreaux, Beck, Albert and Dannecker (2003). Differential peptide binding motif for three juvenile arthritis associated HLA-DQ molecules. **Clin Exp Rheumatol** 21, 257-262

Ionita-Laza, Lange and N (2009). Estimating the number of unseen variants in the human genome. **Proc Natl Acad Sci U S A** 106, 5008-5013

IPDGC and Vincent Plagnol (2011). A two-stage meta-analysis identifies several new loci for Parkinson's disease. **PLoS Genet** 7, e1002142

Iqbal, Caccamo, Turner, Flicek and McVean (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. **Nat Genet** advance online publication,

Iyengar and Farnham (2011). KAP1 Protein: An Enigmatic Master Regulator of the Genome. **Journal of Biological Chemistry** 286, 26267-26276

Iyengar, Ivanov, Jin, Rauscher and Farnham (2011). Functional Analysis of KAP1 Genomic Recruitment. **Mol. Cell. Biol.** 31, 1833-1847

Jeffreys, Kauppi and Neumann (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. **Nat Genet** 29, 217-222

Jiang, Dong and Dai (2009). Genome-wide association study of rheumatoid arthritis by a score test based on wavelet transformation. **BMC Proc** 3 Suppl 7, S8

Jin, Birlea, Fain, Gowan, Riccardi, Holland, Mailloux, Sufit, Hutton, Amadi-Myers*, et al.* (2010). Variant of TYR and autoimmunity susceptibility loci in generalized vitiligo. **N Engl J Med** 362, 1686-1697

Johansen and Prywes (1995). Serum response factor: transcriptional regulation of genes induced by growth factors and differentiation. **Biochimica et Biophysica Acta (BBA) - Reviews on Cancer** 1242, 1-10

Johansson, Lie, Todd, Pociot, Nerup, Cambon-Thomsen, Kockum, Akselsen, Thorsby and Undlien (2003). Evidence of at least two type 1 diabetes susceptibility genes in the HLA complex distinct from HLA-DQB1, -DQA1 and -DRB1. **Genes Immun** 4, 46-53

Johnson, Castle, Garrett-Engele, Kan, Loerch, Armour, Santos, Schadt, Stoughton and Shoemaker (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. **Science** 302, 2141-2144

Johnson, Handsaker, Pulit, Nizzari, O'Donnell and de Bakker (2008). SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. **Bioinformatics** 24, 2938-2939

Jones, Fugger, Strominger and Siebold (2006). MHC class II proteins and disease: a structural perspective. **Nat Rev Immunol** 6, 271-282

Julia, Ballina, Canete, Balsa, Tornero-Molina, Naranjo, Alperi-Lopez, Erra, Pascual-Salcedo, Barcelo*, et al.* (2008). Genome-wide association study of rheumatoid arthritis in the Spanish population: KLF12 as a risk locus for rheumatoid arthritis susceptibility. **Arthritis Rheum** 58, 2275-2286

Jurinke, Denissenko, Oeth, Ehrich, van den Boom and Cantor (2005). A single nucleotide polymorphism based approach for the identification and characterization of gene expression modulation using MassARRAY. **Mutat Res** 573, 83-95

Jurinke, van den Boom, Cantor and Koster (2001). Automated genotyping using the DNA MassArray technology. **Methods Mol Biol** 170, 103-116

Kacem and Feil (2009). Chromatin mechanisms in genomic imprinting. **Mamm Genome** 20, 544-556

Kaslow, Duquesnoy, VanRaden, Kingsley, Marrari, Friedman, Su, Saah, Detels, Phair*, et al.* (1990). A1, Cw7, B8, DR3 HLA antigen combination associated with rapid decline of T-helper lymphocytes in HIV-1 infection. A report from the Multicenter AIDS Cohort Study. **Lancet** 335, 927-930

Kauppi, Sajantila and Jeffreys (2003). Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. **Hum Mol Genet** 12, 33-40

Keating, Tischfield, Murray, Bhangale, Price, Glessner, Galver, Barrett, Grant, Farlow*, et al.* (2008). Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. **PLoS One** 3, e3583

Kent, Sugnet, Furey, Roskin, Pringle, Zahler, Haussler and David (2002). The Human Genome Browser at UCSC. **Genome Research** 12, 996-1006

Kharchenko, Tolstorukov and Park (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. **Nat Biotech** 26, 1351-1359

Kim, Basak and Holtzman (2009). The role of apolipoprotein E in Alzheimer's disease. **Neuron** 63, 287-303

Kim, Goren and Ast (2008). Alternative splicing: current perspectives. **Bioessays** 30, 38-47

Knight (2004). Allele-specific gene expression uncovered. **Trends in Genetics** 20, 113-116

Knight (2005a). HaploChIP: an in vivo assay. **Methods Mol Biol** 311, 49-60

Knight (2005b). Regulatory polymorphisms underlying complex disease traits. **Journal of Molecular Medicine** 83, 97-109

Knight (2006). Analysis of allele-specific gene expression. **Methods Mol Biol** 338, 153-165

Knight (2009a). Genetics and the general physician: insights, applications and future challenges. **QJM** 102, 757-772

Knight (2009b). *Human genetic diversity : functional consequences for health and disease*. Oxford, Oxford University Press

Knight (In Press). Resolving the variable genome and epigenome in human disease. **Journal of Internal Medicine**

Knight, Keating, Rockett and Kwiatkowski (2003). In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. **Nat Genet** 33, 469-475

Korman, Boss, Spies, Sorrentino, Okada and Strominger (1985). Genetic complexity and expression of human class II histocompatibility antigens. **Immunol Rev** 85, 45-86

Kreil, Russell and Russell (2006). Microarray oligonucleotide probes. **Methods Enzymol** 410, 73-98

Kruglyak (2008). The road to genome-wide association studies. **Nat Rev Genet** 9, 314-318

Kugathasan, Baldassano, Bradfield, Sleiman, Imielinski, Guthery, Cucchiara, Kim, Frackelton, Annaiah*, et al.* (2008). Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. **Nat Genet** 40, 1211-1215

Kulpa and Collins (2011). The emerging role of HLA-C in HIV-1 infection. **Immunology** 134, 116-122

Kumar, Kato, Urabe, Takahashi, Muroyama, Hosono, Otsuka, Tateishi, Omata, Nakagawa*, et al.* (2011). Genome-wide association study identifies a susceptibility locus for HCV-induced hepatocellular carcinoma. **Nat Genet** 43, 455-458

Kwan, Benovoy, Dias, Gurd, Provencher, Beaulieu, Hudson, Sladek and Majewski (2008). Genome-wide analysis of transcript isoform variation in humans. **Nat Genet** 40, 225-231

Kwan, Benovoy, Dias, Gurd, Serre, Zuzan, Clark, Schweitzer, Staples, Wang*, et al.* (2007). Heritability of alternative splicing in the human genome. **Genome Res** 17, 1210-1218

Lander, Linton, Birren, Nusbaum, Zody, Baldwin, Devon, Dewar, Doyle, FitzHugh*, et al.* (2001). Initial sequencing and analysis of the human genome. **Nature** 409, 860-921

Landi, Chatterjee, Yu, Goldin, Goldstein, Rotunno, Mirabello, Jacobs, Wheeler, Yeager*, et al.* (2009). A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. **Am J Hum Genet** 85, 679-691

Langmead, Trapnell, Pop and Salzberg (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. **Genome Biology** 10, R25

Larsen and Alper (2004). The genetics of HLA-associated disease. **Curr Opin Immunol** 16, 660-667

Ledbetter, Riccardi, Airhart, Strobel, Keenan and Crawford (1981). Deletions of chromosome 15 as a cause of the Prader-Willi syndrome. **N Engl J Med** 304, 325-329

Leek and Storey (2007). Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. **PLoS Genet** 3, e161

Leslie, Donnelly and McVean (2008). A statistical method for predicting classical HLA alleles from SNP data. **Am J Hum Genet** 82, 48-56

Lettre and Rioux (2008). Autoimmune diseases: insights from genome-wide association studies. **Hum Mol Genet** 17, R116-121

Levsky, Shenoy, Pezo and Singer (2002). Single-Cell Gene Expression Profiling. **Science** 297, 836-840

Lewis and Reik (2006). How imprinting centres work. **Cytogenet Genome Res** 113, 81-89

Li, Ito, Zhou, Youngson, Zuo, Leder and Ferguson-Smith (2008). A maternal-zygotic effect gene, Zfp57, maintains both maternal and paternal imprints. **Dev Cell** 15, 547-557

Li and Leder (2007). Identifying genes preferentially expressed in undifferentiated embryonic stem cells. **BMC Cell Biol** 8, 37

Li, Wang, Li, Bruzel, Richards, Toung and Cheung (2011). Widespread RNA and DNA Sequence Differences in the Human Transcriptome. **Science** 333, 53-58

Li, Ye, Li, Yan, Butcher, Sun, Han, Chen, Zhang and Wang (2010). Whole genome DNA methylation analysis based on high throughput sequencing technology. **Methods** 52, 203-212

Limou, Le Clerc, Coulonges, Carpentier, Dina, Delaneau, Labib, Taing, Sladek, Deveau*, et al.* (2009). Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02). **J Infect Dis** 199, 419-426

Liu, Helms, Liao, Zaba, Duan, Gardner, Wise, Miner, Malloy, Pullinger*, et al.* (2008). A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci. **PLoS Genet** 4, e1000041

Lo, Wang, Hu, Yang, Gere, Buetow and Lee (2003). Allelic variation in gene expression is common in the human genome. **Genome Res** 13, 1855-1862

Looman, Abrink, Mark and Hellman (2002). KRAB zinc finger proteins: an analysis of the molecular mechanisms governing their increase in numbers and complexity during evolution. **Mol Biol Evol** 19, 2118-2130

Lunter and Goodson (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. **Genome Res** 21, 936-939

Mack, Johnson, Roberts, Roberts, Estes, David, Grumet and McLeod (1999). HLA-class II genes modify outcome of Toxoplasma gondii infection. **Int J Parasitol** 29, 1351-1358

Mackay, Boonen, Clayton-Smith, Goodship, Hahnemann, Kant, Njolstad, Robin, Robinson, Siebert*, et al.* (2006). A maternal hypomethylation syndrome presenting as transient neonatal diabetes mellitus. **Hum Genet** 120, 262-269

Mackay, Callaway, Marks, White, Acerini, Boonen, Dayanikli, Firth, Goodship, Haemers*, et al.* (2008). Hypomethylation of multiple imprinted loci in individuals with transient neonatal diabetes is associated with mutations in ZFP57. **Nat Genet** 40, 949-951

Mackay and Temple (2010). Transient neonatal diabetes mellitus type 1. **Am J Med Genet C Semin Med Genet** 154C, 335-342

Maher (2008). Personal genomes: The case of the missing heritability. **Nature** 456, 18-21

Majewski and Pastinen (2011). The study of eQTL variations by RNA-seq: from SNPs to phenotypes. **Trends in Genetics** 27, 72-79

Majumder, Gomez, Chadwick and Boss (2008). The insulator factor CTCF controls MHC class II gene expression and is required for the formation of long-distance chromatin interactions. **J Exp Med** 205, 785-798

Malhotra, Darvishi, Sood, Sharma, Grover, Relhan, Reddy and Bamezai (2005). IL-10 promoter single nucleotide polymorphisms are significantly associated with resistance to leprosy. **Hum Genet** 118, 295-300

Malone, Tan, Bridges and Peng (2011). Comparison of Four ChIP-Seq Analytical Algorithms Using Rice Endosperm H3K27 Trimethylation Profiling Data. **PLoS ONE** 6, e25260

Manolio (2010). Genomewide association studies and assessment of the risk of disease. **N Engl J Med** 363, 166-176

Manolio, Collins, Cox, Goldstein, Hindorff, Hunter, McCarthy, Ramos, Cardon, Chakravarti*, et al.* (2009). Finding the missing heritability of complex diseases. **Nature** 461, 747-753

Marchini, Cardon, Phillips and Donnelly (2004). The effects of human population structure on large genetic association studies. **Nat Genet** 36, 512-517

Marioni, Mason, Mane, Stephens and Gilad (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. **Genome Res** 18, 1509-1517

Massie and Mills (2009). Chromatin immunoprecipitation (ChIP) methodology and readouts. **Methods Mol Biol** 505, 123-137

Maston, Evans and Green (2006). Transcriptional regulatory elements in the human genome. **Annu Rev Genomics Hum Genet** 7, 29-59

Mathew (2008). New links to the pathogenesis of Crohn disease provided by genome-wide association scans. **Nat Rev Genet** 9, 9-14

Mathew, Xu and George (2009). Simultaneous analysis of all single-nucleotide polymorphisms in genome-wide association study of rheumatoid arthritis. **BMC Proc** 3 Suppl 7, S11

Maynard, Chen, Stuart, Fan and Ren (2008). Genome-wide mapping of allele-specific protein-DNA interactions in human cells. **Nat Meth** 5, 307-309

Mbarek, Ochi, Urabe, Kumar, Kubo, Hosono, Takahashi, Kamatani, Miki, Abe*, et al.* (2011). A genome-wide association study of chronic hepatitis B identified novel risk locus in a Japanese population. **Hum Mol Genet** 20, 3884-3892

McCarthy, Abecasis, Cardon, Goldstein, Little, Ioannidis and Hirschhorn (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. **Nat Rev Genet** 9, 356-369

McEwen and Ferguson-Smith (2010). Distinguishing epigenetic marks of developmental and imprinting regulation. **Epigenetics Chromatin** 3, 2

McManus, Coolon, Duff, Eipper-Mains, Graveley and Wittkopp (2011). Regulatory divergence in Drosophila revealed by mRNA-seq. **Genome Res** 20, 816-825

Mill, Tang, Kaminsky, Khare, Yazdanpanah, Bouchard, Jia, Assadzadeh, Flanagan, Schumacher*, et al.* (2008). Epigenomic profiling reveals DNA-methylation changes associated with major psychosis. **Am J Hum Genet** 82, 696-711

Milner and Campbell (1990). Structure and expression of the three MHC-linked HSP70 genes. **Immunogenetics** 32, 242-251

Moffatt, Gut, Demenais, Strachan, Bouzigon, Heath, von Mutius, Farrall, Lathrop and Cookson (2010). A large-scale, consortium-based genomewide association study of asthma. **N Engl J Med** 363, 1211-1221

Monks, Leonardson, Zhu, Cundiff, Pietrusiak, Edwards, Phillips, Sachs and Schadt (2004). Genetic Inheritance of Gene Expression in Human Cell Lines. **American journal of human genetics** 75, 1094-1105

Montgomery, Lappalainen, Gutierrez-Arcelus and Dermitzakis (2011). Rare and common regulatory variation in population-scale sequenced human genomes. **PLoS Genet** 7, e1002144

Montgomery, Sammeth, Gutierrez-Arcelus, Lach, Ingle, Nisbett, Guigo and Dermitzakis (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. **Nature** 464, 773-777

Moore and Silver (2008). Global analysis of mRNA splicing. **RNA** 14, 197-203

Morin, Bainbridge, Fejes, Hirst, Krzywinski, Pugh, McDonald, Varhol, Jones and Marra (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. **Biotechniques** 45, 81-94

Morison, Ramsay and Spencer (2005). A census of mammalian imprinting. **Trends Genet** 21, 457-465

Morley, Molony, Weber, Devlin, Ewens, Spielman and Cheung (2004). Genetic analysis of genome-wide variation in human gene expression. **Nature** 430, 743-747

Mortazavi, Williams, McCue, Schaeffer and Wold (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. **Nat Meth** 5, 621-628

Mungall, Palmer, Sims, Edwards, Ashurst, Wilming, Jones, Horton, Hunt, Scott*, et al.* (2003). The DNA sequence and analysis of human chromosome 6. **Nature** 425, 805-811

Murray, Moore, Van Dyke, Lahr, Dierkhising, Zinsmeister, Melton, Kroning, El-Yousseff and Czaja (2007). HLA DQ gene dosage and risk and severity of celiac disease. **Clin Gastroenterol Hepatol** 5, 1406-1412

Myers, Stamatoyannopoulos, Snyder, Dunham, Hardison, Bernstein, Gingeras, Kent, Birney, Wold*, et al.* (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). **PLoS Biol** 9, e1001046

Nagalakshmi, Wang, Waern, Shou, Raha, Gerstein and Snyder (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. **Science** 320, 1344-1349

Nair, Stuart, Nistor, Hiremagalore, Chia, Jenisch, Weichenthal, Abecasis, Lim, Christophers*, et al.* (2006). Sequence and haplotype analysis supports HLA-C as the psoriasis susceptibility 1 gene. **Am J Hum Genet** 78, 827-851

Neefjes and Ploegh (1988). Allele and locus-specific differences in cell surface expression and the association of HLA class I heavy chain with β2-microglobulin: differential effects of inhibition of glycosylation on class I subunit association. **European Journal of Immunology** 18, 801-810

Nejentsev, Howson, Walker, Szeszko, Field, Stevens, Reynolds, Hardy, King, Masters*, et al.* (2007). Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. **Nature** 450, 887-892

Nembaware, Lupindo, Schouest, Spillane, Scheffler and Seoighe (2008). Genome-wide survey of allele-specific splicing in humans. **BMC Genomics** 9, 265

Nembaware, Wolfe, Bettoni, Kelso and Seoighe (2004). Allele-specific transcript isoforms in human. **FEBS Letters** 577, 233-238

Nica, Montgomery, Dimas, Stranger, Beazley, Barroso and Dermitzakis (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. **PLoS Genet** 6, e1000895

Nica, Parts, Glass, Nisbet, Barrett, Sekowska, Travers, Potter, Grundberg, Small*, et al.* (2011). The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study. **PLoS Genet** 7, e1002003

Nicholls (1993). Genomic imprinting and uniparental disomy in Angelman and Prader-Willi syndromes: a review. **Am J Med Genet** 46, 16-25

Nielsen (2010). Genomics: In search of rare human variants. **Nature** 467, 1050-1051

Nischwitz, Cepok, Kroner, Wolf, Knop, Muller-Sarnowski, Pfister, Roeske, Rieckmann, Hemmer*, et al.* (2010). Evidence for VAV2 and ZNF433 as susceptibility genes for multiple sclerosis. **J Neuroimmunol** 227, 162-166

Nowick, Hamilton, Zhang and Stubbs (2010). Rapid Sequence and Expression Divergence Suggest Selection for Novel Function in Primate-Specific KRAB-ZNF Genes. **Molecular Biology and Evolution** 27, 2606-2617

O'Geen, Squazzo, Iyengar, Blahnik, Rinn, Chang, Green and Farnham (2007). Genome-Wide Analysis of KAP1 Binding Suggests Autoregulation of KRAB-ZNFs. **PLoS Genet** 3, e89

Okada, Yamazaki, Umeno, Takahashi, Kumasaka, Ashikawa, Aoi, Takazoe, Matsui, Hirano*, et al.* (2011). HLA-Cw*1202-B*5201-DRB1*1502 haplotype increases risk for ulcerative colitis but reduces risk for Crohn's disease. **Gastroenterology** 141, 864-871 e861-865

Okamoto, Kitabayashi and Taya (2006). KAP1 dictates p53 response induced by chemotherapeutic agents via Mdm2 interaction. **Biochem Biophys Res Commun** 351, 216-222

Okazaki, Tanase, Choudhury, Setoyama, Miura, Ogawa and Setoyama (1994). A novel nuclear protein with zinc fingers down-regulated during early mammalian cell differentiation. **J. Biol. Chem.** 269, 6900-6907

Oshlack and Wakefield (2009). Transcript length bias in RNA-seq data confounds systems biology. **Biol Direct** 4, 14

Pant, Tao, Beilharz, Ballinger, Cox and Frazer (2006). Analysis of allelic differential expression in human white blood cells. **Genome Res** 16, 331-339

Park (2009). ChIP-seq: advantages and challenges of a maturing technology. **Nat Rev Genet** 10, 669-680

Pastinen (2010). Genome-wide allele-specific analysis: insights into regulatory variation. **Nat Rev Genet** 11, 533-538

Pastinen, Ge, Gurd, Gaudin, Dore, Lemire, Lepage, Harmsen and Hudson (2005). Mapping common regulatory variants to human haplotypes. **Hum Mol Genet** 14, 3963-3971

Pastinen and Hudson (2004). Cis-Acting Regulatory Variation in the Human Genome. **Science** 306, 647-650

Pastinen, Sladek, Gurd, Sammak, Ge, Lepage, Lavergne, Villeneuve, Gaudin, Brandstrom*, et al.* (2004). A survey of genetic and epigenetic variation affecting human gene expression. **Physiol Genomics** 16, 184-193

Pennisi (2010). Genomics. 1000 Genomes Project gives new map of genetic diversity. **Science** 330, 574-575

Pereyra, Jia, McLaren, Telenti, de Bakker, Walker, Ripke, Brumme, Pulit, Carrington*, et al.* (2010). The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. **Science** 330, 1551-1557

Pericak-Vance, Bebout, Gaskell, Yamaoka, Hung, Alberts, Walker, Bartlett, Haynes, Welsh*, et al.* (1991). Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. **Am J Hum Genet** 48, 1034-1050

Perricone, Ceccarelli and Valesini (2011). An overview on the genetic of rheumatoid arthritis: A never-ending story. **Autoimmun Rev** 10, 599-608

Petukhova, Duvic, Hordinsky, Norris, Price, Shimomura, Kim, Singh, Lee, Chen*, et al.* (2010). Genome-wide association study in alopecia areata implicates both innate and adaptive immunity. **Nature** 466, 113-117

Phillips and Corces (2009). CTCF: master weaver of the genome. **Cell** 137, 1194-1211

Pickrell, Marioni, Pai, Degner, Engelhardt, Nkadori, Veyrieras, Stephens, Gilad and Pritchard (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. **Nature** 464, 768-772

Plath, Mlynarczyk-Evans, Nusinow and Panning (2002). Xist RNA and the mechanism of X chromosome inactivation. **Annu Rev Genet** 36, 233-278

Plomin, Haworth and Davis (2009). Common disorders are quantitative traits. **Nat Rev Genet** 10, 872-878

Price, Witt, Allcock, Sayer, Garlepp, Kok, French, Mallal and Christiansen (1999). The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases. **Immunol Rev** 167, 257-274

Prugnolle, Manica, Charpentier, Guegan, Guernier and Balloux (2005). Pathogen-driven selection and worldwide HLA class I diversity. **Curr Biol** 15, 1022-1027

Pugliese, Zeller, Fernandez, Zalcberg, Bartlett, Ricordi, Pietropaolo, Eisenbarth, Bennett and Patel (1997). The insulin gene is transcribed in the human thymus and transcription levels correlated with allelic variation at the INS VNTR-IDDM2 susceptibility locus for type 1 diabetes. **Nat Genet** 15, 293-297

Purcell, Consortium and Wray NR (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. **Nature** 460, 748-752

Purcell, Neale, Todd-Brown, Thomas, Ferreira, Bender, Maller, Sklar, de Bakker, Daly*, et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. **Am J Hum Genet** 81, 559-575

Qiao, Sollid and Blumberg (2009). Antigen presentation in celiac disease. **Curr Opin Immunol** 21, 111-117

Quenneville, Verde, Corsinotti, Kapopoulou, Jakobsson, Offner, Baglivo, Pedone, Grimaldi, Riccio*, et al.* (2011). In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. **Mol Cell** 44, 361-372

Radstake, Gorlova, Rueda, Martin, Alizadeh, Palomino-Morales, Coenen, Vonk, Voskuyl, Schuerwegh*, et al.* (2010). Genome-wide association study of systemic sclerosis identifies CD247 as a new susceptibility locus. **Nat Genet** 42, 426-429

Rambaud, Desroches, Balsalobre and Drouin (2009). TIF1beta/KAP-1 is a coactivator of the orphan nuclear receptor NGFI-B/Nur77. **J Biol Chem** 284, 14147-14156

Rebbeck, Spitz and Wu (2004). Assessing the function of genetic variants in candidate gene association studies. **Nat Rev Genet** 5, 589-597

Redon, Ishikawa, Fitch, Feuk, Perry, Andrews, Fiegler, Shapero, Carson, Chen*, et al.* (2006). Global variation in copy number in the human genome. **Nature** 444, 444-454

Redondo, Fain and Eisenbarth (2001). Genetics of type 1A diabetes. **Recent Prog Horm Res** 56, 69-89

Reich and Lander (2001). On the allelic spectrum of human disease. **Trends in Genetics** 17, 502-510

Reik and Walter (2001). Genomic imprinting: parental influence on the genome. **Nat Rev Genet** 2, 21-32

Rhodes and Vyse (2008). The genetics of SLE: an update in the light of genome-wide association studies. **Rheumatology** 47, 1603-1611

Rhodes and Vyse (2010). Using genetics to deliver personalized SLE therapy[mdash]a realistic prospect? **Nat Rev Rheumatol** 6, 373-377

Richards (2006). Inherited epigenetic variation--revisiting soft inheritance. **Nat Rev Genet** 7, 395-401

Rioux, Daly, Silverberg, Lindblad, Steinhart, Cohen, Delmonte, Kocher, Miller, Guschwan*, et al.* (2001). Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. **Nat Genet** 29, 223-228

Rioux, Goyette, Vyse, Hammarstrom, Fernando, Green, De Jager, Foisy, Wang, de Bakker*, et al.* (2009). Mapping of multiple susceptibility variants within the MHC region for 7 immune-mediated diseases. **Proc Natl Acad Sci U S A** 106, 18680-18685

Risch and Merikangas (1996). The future of genetic studies of complex human diseases. **Science** 273, 1516-1517

Roach, Glusman, Smit, Huff, Hubley, Shannon, Rowen, Pant, Goodman, Bamshad*, et al.* (2010). Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. **Science** 328, 636-639

Rosenbloom, Dreszer, Long, Malladi, Sloan, Raney, Cline, Karolchik, Barber, Clawson*, et al.* (2011). ENCODE whole-genome data in the UCSC Genome Browser: update 2012. **Nucleic Acids Res** In Print

Roy (2003). Immunology: Professional secrets. **Nature** 425, 351-352

Roze, Bonnet, Betuing and Caboche (2010). Huntington's disease. **Adv Exp Med Biol** 685, 45-63

Rozowsky, Abyzov, Wang, Alves, Raha, Harmanci, Leng, Bjornson, Kong, Kitabayashi*, et al.* (2011). AlleleSeq: analysis of allele-specific expression and binding in a network framework. **Mol Syst Biol** 7, 522

Rozowsky, Euskirchen, Auerbach, Zhang, Gibson, Bjornson, Carriero, Snyder and Gerstein (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. **Nat Biotechnol** 27, 66-75

Rudy and Lew (1997). The nonpolymorphic MHC class II isotype, HLA-DQA2, is expressed on the surface of B lymphoblastoid cells. **J Immunol** 158, 2116-2125

Sankaran, Menne, Xu, Akie, Lettre, Van Handel, Mikkola, Hirschhorn, Cantor and Orkin (2008). Human Fetal Hemoglobin Expression Is Regulated by the Developmental Stage-Specific Repressor BCL11A. **Science** 322, 1839-1842

Saunders, Strittmatter, Schmechel, George-Hyslop, Pericak-Vance, Joo, Rosi, Gusella, Crapper-MacLachlan, Alberts*, et al.* (1993). Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. **Neurology** 43, 1467-1472

Sawcer, Maranian, Setakis, Curwen, Akesson, Hensiek, Coraddu, Roxburgh, Sawcer, Gray*, et al.* (2002). A whole genome screen for linkage disequilibrium in multiple sclerosis confirms disease associations with regions previously linked to susceptibility. **Brain** 125, 1337-1347

Saxena, Voight, Lyssenko, Burtt, de Bakker, Chen, Roix, Kathiresan, Hirschhorn, Daly*, et al.* (2007). Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels. **Science** 316, 1331-1336

Schadt, Monks, Drake, Lusis, Che, Colinayo, Ruff, Milligan, Lamb, Cavet*, et al.* (2003). Genetics of gene expression surveyed in maize, mouse and man. **Nature** 422, 297-302

Schön and Boehncke (2005). Psoriasis. **New England Journal of Medicine** 352, 1899-1912

Schork, Murray, Frazer and Topol (2009). Common vs. rare allele hypotheses for complex diseases. **Curr Opin Genet Dev** 19, 212-219

Schreiber, Matthias, Muller and Schaffner (1989). Rapid detection of octamer binding proteins with 'mini-extracts', prepared from a small number of cells. **Nucleic Acids Res** 17, 6419

Serre, Gurd, Ge, Sladek, Sinnett, Harmsen, Bibikova, Chudin, Barker, Dickinson*, et al.* (2008). Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. **PLoS Genet** 4, e1000006

Shao, Huang, Yu, Huang, Wu, Xia, Wang, Feng, Ren, Ernberg*, et al.* (2001). Loss of heterozygosity and its correlation with clinical outcome and Epstein-Barr virus infection in nasopharyngeal carcinoma. **Anticancer Res** 21, 3021-3029

Sheehy, Scharf, Rowe, Neme de Gimenez, Meske, Erlich and Nepom (1989). A diabetes-susceptible HLA haplotype is best defined by a combination of HLA-DR and -DQ alleles. **J Clin Invest** 83, 830-835

Shi, Levinson, Duan, Sanders, Zheng, Pe'er, Dudbridge, Holmans, Whittemore, Mowry*, et al.* (2009). Common variants on chromosome 6p22.1 are associated with schizophrenia. **Nature** 460, 753-757

Silva, Hamamoto, Furukawa and Nakamura (2006). TIPUH1 encodes a novel KRAB zinc-finger protein highly expressed in human hepatocellular carcinomas. **Oncogene** 25, 5063-5070

Silverberg, Cho, Rioux, McGovern, Wu, Annese, Achkar, Goyette, Scott, Xu*, et al.* (2009). Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. **Nat Genet** 41, 216-220

Siva (2008). 1000 Genomes project. **Nat Biotechnol** 26, 256

Skibola, Bracci, Halperin, Conde, Craig, Agana, Iyadurai, Becker, Brooks-Wilson, Curry*, et al.* (2009). Genetic variants at 6p21.33 are associated with susceptibility to follicular lymphoma. **Nat Genet** 41, 873-875

Smedby, Foo, Skibola, Darabi, Conde, Hjalgrim, Kumar, Chang, Rothman, Cerhan*, et al.* (2011). GWAS of follicular lymphoma reveals allelic heterogeneity at 6p21.32 and suggests shared genetic susceptibility with diffuse large B-cell lymphoma. **PLoS Genet** 7, e1001378

Song and Crawford (2010). DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. **Cold Spring Harb Protoc** 2, 5384

Song, Zhang, Grasfeder, Boyle, Giresi, Lee, Sheffield, Graf, Huss, Keefe*, et al.* (2011). Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. **Genome Res** 21, 1757-1767

Sorek, Shamir and Ast (2004). How prevalent is functional alternative splicing in the human genome? **Trends Genet** 20, 68-71

Spielman, Bastone, Burdick, Morley, Ewens and Cheung (2007). Common genetic variants account for differences in gene expression among ethnic groups. **Nat Genet** 39, 226-231

Stastny (1976). Mixed lymphocyte cultures in rheumatoid arthritis. **J Clin Invest** 57, 1148-1157

Stefansson, Ophoff, Steinberg, Andreassen, Cichon, Rujescu, Werge, Pietilainen, Mors, Mortensen*, et al.* (2009). Common variants conferring risk of schizophrenia. **Nature** 460, 744-747

Stephens, Horton, Humphray, Rowen, Trowsdale and Beck (1999). Gene organisation, sequence variation and isochore structure at the centromeric boundary of the human MHC. **J Mol Biol** 291, 789-799

Stewart, Horton, Allcock, Ashurst, Atrazhev, Coggill, Dunham, Forbes, Halls, Howson*, et al.* (2004). Complete MHC haplotype sequencing for common disease gene mapping. **Genome Res** 14, 1176-1187

Strange, Capon, Spencer, Knight, Weale, Allen, Barton, Band, Bellenguez, Bergboer*, et al.* (2010). A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. **Nat Genet** 42, 985-990

Stranger and Dermitzakis (2005). The genetics of regulatory variation in the human genome. **Hum Genomics** 2, 126-131

Stranger, Forrest, Clark, Minichiello, Deutsch, Lyle, Hunt, Kahl, Antonarakis, Tavare*, et al.* (2005). Genome-wide associations of gene expression variation in humans. **PLoS Genet** 1, e78

Stranger, Forrest, Dunning, Ingle, Beazley, Thorne, Redon, Bird, de Grassi, Lee*, et al.* (2007). Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes. **Science** 315, 848-853

Strittmatter, Saunders, Schmechel, Pericak-Vance, Enghild, Salvesen and Roses (1993a). Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. **Proc Natl Acad Sci U S A** 90, 1977-1981

Strittmatter, Weisgraber, Huang, Dong, Salvesen, Pericak-Vance, Schmechel, Saunders, Goldgaber and Roses (1993b). Binding of human apolipoprotein E to synthetic amyloid beta peptide: isoform-specific effects and implications for late-onset Alzheimer disease. **Proc Natl Acad Sci U S A** 90, 8098-8102

Sugiura (2011). The cellular level of TRIM31, an RBCC protein overexpressed in gastric cancer, is regulated by multiple mechanisms including the ubiquitin-proteasome system. **Cell Biol Int** 35, 657-661

Sugiura and Miyamoto (2008). Characterization of TRIM31, upregulated in gastric adenocarcinoma, as a novel RBCC protein. **J Cell Biochem** 105, 1081-1091

Sultan, Schulz, Richard, Magen, Klingenhoff, Scherf, Seifert, Borodina, Soldatov, Parkhomchuk*, et al.* (2008). A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. **Science** 321, 956-960

Tadepally, Burger and Aubry (2008). Evolution of C2H2-zinc finger genes and subfamilies in mammals: species-specific duplication and loss of clusters, genes and effector domains. **BMC Evol Biol** 8, 176

Tang, Barbacioru, Nordman, Bao, Lee, Wang, Tuch, Heard, Lao and Surani (2011). Deterministic and stochastic allele specific gene expression in single mouse blastomeres. **PLoS One** 6, e21208

Tarazona, GarcÃa-Alcalde, Dopazo, Ferrer and Conesa (2011). Differential expression in RNA-seq: A matter of depth. **Genome Research** 21, 2213-2223

Tazi, Bakkour and Stamm (2009). Alternative splicing and disease. **Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease** 1792, 14-26

Temple and Shield (2002). Transient neonatal diabetes, a disorder of imprinting. **J Med Genet** 39, 872-875

Temple and Shield (2010). 6q24 transient neonatal diabetes. **Rev Endocr Metab Disord** 11, 199-204

Templeton (2002). Out of Africa again and again. **Nature** 416, 45-51

Thellin, Zorzi, Lakaye, De Borman, Coumans, Hennen, Grisar, Igout and Heinen (1999). Housekeeping genes as internal standards: use and limits. **J Biotechnol** 75, 291-295

Thomas, Apps, Qi, Gao, Male, O'HUigin, O'Connor, Ge, Fellay, Martin*, et al.* (2009). HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. **Nat Genet** 41, 1290-1294

TIHMP (2003). The International HapMap Project. **Nature** 426, 789-796

Todd (1990). Genetic control of autoimmunity in type 1 diabetes. **Immunol Today** 11, 122-129

Todd, Acha-Orbea, Bell, Chao, Fronek, Jacob, McDermott, Sinha, Timmerman, Steinman*, et al.* (1988). A molecular basis for MHC class II--associated autoimmunity. **Science** 240, 1003-1009

Trachtenberg, Bhattacharya, Ladner, Phair, Erlich and Wolinsky (2009). The HLA-B/-C haplotype block contains major determinants for host control of HIV. **Genes Immun** 10, 673-677

Trachtenberg, Korber, Sollars, Kepler, Hraber, Hayes, Funkhouser, Fugate, Theiler, Hsu*, et al.* (2003). Advantage of rare HLA supertype in HIV disease progression. **Nat Med** 9, 928-935

Traherne (2008). Human MHC architecture and evolution: implications for disease association studies. **Int J Immunogenet** 35, 179-192

Traherne, Horton, Roberts, Miretti, Hurles, Stewart, Ashurst, Atrazhev, Coggill, Palmer*, et al.* (2006). Genetic analysis of completely sequenced disease-associated MHC haplotypes identifies shuffling of segments in recent human history. **PLoS Genet** 2, e9

Trembath, Lee Clough, Rosbotham, Jones, Camp, Frodsham, Browne, Barber, Terwilliger, Mark Lathrop*, et al.* (1997). Identification of a Major Susceptibility Locus on Chromosome 6p and Evidence for Further Disease Loci Revealed by a Two Stage Genome-Wide Search in Psoriasis. **Human Molecular Genetics** 6, 813-820

Tse, Su, Chang, Tsang, Yu, Tang, See, Hsueh, Yang, Hao*, et al.* (2009). Genome-wide association study reveals multiple nasopharyngeal carcinoma-associated loci within the HLA region at chromosome 6p21.3. **Am J Hum Genet** 85, 194-203

Tuteja, White, Schug and Kaestner (2009). Extracting transcription factor targets from ChIP-Seq data. **Nucleic Acids Research** 37, e113

Valdes, Erlich and Noble (2005). Human leukocyte antigen class I B and C loci contribute to Type 1 Diabetes (T1D) susceptibility and age at T1D onset. **Hum Immunol** 66, 301-313

van Heel, Franke, Hunt, Gwilliam, Zhernakova, Inouye, Wapenaar, Barnardo, Bethel, Holmes*, et al.* (2007). A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. **Nat Genet** 39, 827-829

Vandiedonck, Beaurain, Giraud, Hue-Beauvais, Eymard, Tranchant, Gajdos, Dausset and Garchon (2004). Pleiotropic effects of the 8.1 HLA haplotype in patients with autoimmune myasthenia gravis and thymus hyperplasia. **Proc Natl Acad Sci U S A** 101, 15464-15469

Vandiedonck and Knight (2009). The human Major Histocompatibility Complex as a paradigm in genomics research. **Brief Funct Genomic Proteomic** 8, 379-394

Vandiedonck, Taylor, Lockstone, Plant, Taylor, Durrant, Broxholme, Fairfax and Knight (2011). Pervasive haplotypic variation in the spliceo-transcriptome of the human major histocompatibility complex. **Genome Research** 21, 1042-1054

Venter, Adams, Myers, Li, Mural, Sutton, Smith, Yandell, Evans, Holt*, et al.* (2001). The sequence of the human genome. **Science** 291, 1304-1351

Verlaan, Ge, Grundberg, Hoberman, Lam, Koka, Dias, Gurd, Martin, Mallmin*, et al.* (2009). Targeted screening of cis-regulatory variation in human haplotypes. **Genome Research** 19, 118-127

Via, Gignoux and Burchard (2010). The 1000 Genomes Project: new opportunities for research and social challenges. **Genome Med** 2, 3

Viken, Blomhoff, Olsson, Akselsen, Pociot, Nerup, Kockum, Cambon-Thomsen, Thorsby, Undlien*, et al.* (2009). Reproducible association with type 1 diabetes in the extended class I region of the major histocompatibility complex. **Genes Immun** 10, 323-333

Wain, Armour and Tobin (2009). Genomic copy number variation, human health, and disease. **Lancet** 374, 340-350

Walter, McWeeney, Peters, Belknap, Hitzemann and Buck (2007). SNPs matter: impact on detection of differential expression. **Nat Methods** 4, 679-680

Wang, Barratt, Clayton and Todd (2005). Genome-wide association studies: theoretical and practical concerns. **Nat Rev Genet** 6, 109-118

Wang, Broderick, Webb, Wu, Vijayakrishnan, Matakidou, Qureshi, Dong, Gu, Chen*, et al.* (2008a). Common 5p15.33 and 6p21.33 variants influence lung cancer risk. **Nat Genet** 40, 1407-1409

Wang and Cooper (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. **Nat Rev Genet** 8, 749-761

Wang, Gerstein and Snyder (2009). RNA-Seq: a revolutionary tool for transcriptomics. **Nat Rev Genet** 10, 57-63

Wang, Sandberg, Luo, Khrebtukova, Zhang, Mayr, Kingsmore, Schroth and Burge (2008b). Alternative isoform regulation in human tissue transcriptomes. **Nature** 456, 470-476

Waring and Rosenberg (2008). Genome-wide association studies in Alzheimer disease. **Arch Neurol** 65, 329-334

Wasserman and Sandelin (2004). Applied bioinformatics for the identification of regulatory elements. **Nat Rev Genet** 5, 276-287

Weintraub and Groudine (1976). Chromosomal subunits in active genes have an altered conformation. **Science** 193, 848-856

Wilbanks and Facciotti (2010). Evaluation of Algorithm Performance in ChIP-Seq Peak Detection. **PLoS ONE** 5, e11471

Williams and Barclay (1988). The immunoglobulin superfamily--domains for cell surface recognition. **Annu Rev Immunol** 6, 381-405

Williams, Spilianakis and Flavell (2010). Interchromosomal association and gene regulation in trans. **Trends Genet** 26, 188-197

Wong, Gochhait, Malhotra, Pettersson, Teo, Khor, Rautanen, Chapman, Mills, Srivastava*, et al.* (2010). Leprosy and the Adaptation of Human Toll-Like Receptor 1. **PLoS Pathog** 6, e1000979

WTCCC (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. **Nature** 447, 661-678

Xie, Rowen, Aguado, Ahearn, Madan, Qin, Campbell and Hood (2003). Analysis of the gene-dense major histocompatibility complex class III region and its comparison to mouse. **Genome Res** 13, 2621-2636

Yan, Yuan, Velculescu, Vogelstein and Kinzler (2002). Allelic variation in human gene expression. **Science** 297, 1143

Yao, Kimura, Hartung, Haas, Volgger, Brunnler, Bonisch and Albert (1993). Polymorphism of the DQA1 promoter region (QAP) and DRB1, QAP, DQA1, DQB1 haplotypes in systemic lupus erythematosus. SLE Study Group members. **Immunogenetics** 38, 421-429

Yeo, De Jager, Gregory, Barcellos, Walton, Goris, Fenoglio, Ban, Taylor, Goodman*, et al.* (2007). A second major histocompatibility complex susceptibility locus for multiple sclerosis. **Ann Neurol** 61, 228-236

Yokoe, Toiyama, Okugawa, Tanaka, Ohi, Inoue, Mohri, Miki and Kusunoki (2010). KAP1 is associated with peritoneal carcinomatosis in gastric cancer. **Ann Surg Oncol** 17, 821-828

Yunis, Larsen, Fernandez-Vina, Awdeh, Romero, Hansen and Alper (2003). Inheritable variable sizes of DNA stretches in the human MHC: conserved extended haplotypes and their fragments or blocks. **Tissue Antigens** 62, 1-20

Zhang, Huang, Chen, Sun, Liu, Li, Cui, Yan, Yang, Rong-De*, et al.* (2009). Genomewide Association Study of Leprosy. **New England Journal of Medicine** 361, 2609-2618

Zhang, Liu, Meyer, Eeckhoute, Johnson, Bernstein, Nusbaum, Myers, Brown, Li*, et al.* (2008). Model-based analysis of ChIP-Seq (MACS). **Genome Biol** 9, R137

Zheng, Chung and Zhao (2011). Bias detection and correction in RNA-Sequencing data. **BMC Bioinformatics** 12, 290

Zielenski (2000). Genotype and phenotype in cystic fibrosis. **Respiration** 67, 117-133

Zuo, Sheng, Lau, McDonald, Andrade, Cullen, Bell, Iacovino, Kyba, Xu*, et al.* (2011). The zinc finger protein ZFP57 requires its cofactor to recruit DNA methyltransferases and maintains the DNA methylation imprint in embryonic stem cells via its transcriptional repression domain. **J Biol Chem** Advance online publication. doi:10.1074/jbc.M1111.322644