CONSERVED EXTENDED HAPLOTYPES IN MHC OF SINGAPOREAN CHINESE

LAM TZE HAU

B.Eng. (Hons.)

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF MICROBIOLOGY

NATIONAL UNIVERSITY OF SINGAPORE

2014

DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety.

I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Lam Tze Hau

6th Jan 2014

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Assoc Prof Ren Ee Chee, for the invaluable guidance and mentorship over the years. Thank you for all the opportunities and support you have given me, it is a pleasure to work under your leadership. I would also like thank my TAC members, Prof Wong Lim Soon and Assoc Prof Sung Win-kin for the discussions and advice on the project.

The success of this project would not be possible without the help and support of each and every member of the REC lab. In particular, I would like to thanks Dr Xiao Ziwei, Dr Shen Meixin and Dr Wang Bei for their guidance in the experimental work. I also wish to thank Mr Matthew Tay for his help and advice on this project as well as all the research staffs for the technical and administrative support throughout my PhD candidature.

I am grateful to Immunology Center, NUS and Prof Chan Soh Ha for providing some of the cell lines used in this work.

Last but not least, I would like to thank my family for their support and encouragement.

Table of Contents

List of Figures

List of Tables

Abbreviations

B-LCL	B-Lymphoblastoid cell line
CEH	Conserved extended haplotype
CEU	HapMap European population
CG	Complete genomics
CHSG	Singapore Chinese
EHH	Extended haplotype homozygosity
eQTL	Expression quantitative trait loci
FPKM	Fragments per kilobase of exon per million mapped reads
HLA	Human leukocyte antigen
LD	Linkage disequilibrium
MHC	Major Histocompatibility Complex
NGS	Next-generation sequencing
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variants
YRI	HapMap Nigeria Yoruba population
ZFP57	Zinc Finger Protein 57 homolog

Summary

Extensive long-range linkage disequilibrium within the MHC region, commonly termed as conserved extended haplotypes (CEHs), are known to occur at relatively high frequency in the general populations. CEHs have been of long interest not just because of their unique genomic traits but also because they are reported to be associated with several diseases. However due to the inherent heterogeneity of the MHC region and the limitation of the technologies, thus far, the handful of studies on the genomic structure of MHC CEHs have been restricted to the European population. This lack of comparative MHC CEH sequence information in other ethnic groups hampers the efforts to elucidate the structure and genomic organization of CEHs.

In this study, the recombination patterns across the MHC region in three independent populations were examined. This analysis revealed population-specific recombination sites and these sites are seldom shared among populations, underlining the importance of recombination on haplotype diversity. Furthermore, from the SNPs analysis, I also uncovered two HLA haplotypes in the Singaporean Chinese population (A*33:03-B*58:01-DR*03:01 and A*02:07-B*46:01-DR*09:01) with no or minimal recombination across the MHC region, suggesting CEH properties in these haplotypes.

To have an in-depth genomic architecture on above two Singapore Chinese HLA haplotypes, multiple HLA homozygous B-LCLs were selected and subjected to whole genome sequencing. The analysis of this data revealed two significant findings. Firstly, extensive sequence conservation spanning a region of at least 3Mb of the MHC genomic region was found among the individuals carrying identical HLA haplotype. In addition, the intra-haplotypic variations within these CEHs were found to be exceptionally low comprising of approximate 0.008% of the MHC region. Novel single

nucleotide variation (SNV) not reported in other databases were found in 77/293 (26%) of A*33:03-B*58:01-DR*03:01 CEH and 50/238 (21%) of the A*02:07-B*46:01-DR*09:01 CEH. More importantly, SNVs found within the A*02:07-B*46:01-DR*09:01 CEH were associated with the expression of *ZFP57*, a transcription factor involved in DNA methylation maintenance; suggesting functional role in some of these polymorphic sites. The second major finding is that extreme sequence conservation extending up to 160kb at the HLA-DR region was found between the A*33:03-B*58:01-DR*03:01 haplotype and the European A1-B8-DR3 haplotype; implying individuals carrying these two haplotypes shared a common ancestor.

Next, the MHC transcription landscape of the Singaporean Chinese CEHs was also elucidated through RNA-sequencing. Interestingly, differences in gene expression between haplotypes affecting 26 genes were observed; this implies the influence of underlying MHC haplotypic structure on the transcription activity in the MHC region. Collectively, a comprehensive sequence and transcription description representative of the A*33:03-B*58:01-DR*03:01 and A*02:07-B*46:01-DR*09:01 haplotype was provided and I had showed that haplotype-specific sequence variations mediate the level of gene expression in the MHC. The availability of these alternate Asian MHC sequences would complement the eight European MHC haplotype sequenced by the MHC Haplotype Project and provides a framework to study the MHC diversity and disease association studies.

Chapter 1:

Introduction

1.1 Overview of the MHC

The human major histocompatibility complex (MHC) was first identified more than 50 years ago by Jean Dausset, Jon van Rood and Rose Payne [1] and because the MHC molecules were first detected on the surface of the human leukocytes, the human MHC was later also known as the human leukocyte antigen (HLA) complex. In the beginning, it was studied due to its role on the donor-patient compatibility following tissue graft and organ transplantation [2, 3], it was later realized that the MHC is critical in adaptive and innate immunity.

Decades of research has progressively established a remarkable genomic region on chromosome position 6p21.3 comprising of the HLA genes and the associative genes. This region is the most gene-dense region in the human genome holding more than 240 annotated genes, of which more than 60 have known or potential immune-related function [4, 5]. Furthermore, the extreme genetic variation within this region has a pivotal role in the disease susceptibility and indeed this four mega-base region has been reported to be associated with more than 100 diseases, including cancers, autoimmune diseases, infectious disease susceptibilities, neurodegenerative, cardiovascular, and metabolic disorders [6, 7]. In addition, the MHC region possesses unique genomic features offering an excellent model to study demographic events as well as to assess hypotheses pertaining to the dynamics of evolution [8, 9]. For these reasons, the MHC genomic architecture, diversity, gene expression and genetic interaction pathways have been studied intensively as frameworks for understanding the broader human genome.

1.2 MHC Gene Map and Organization

The human MHC region of approximate 4.6 Mb on the short arm of chromosome 6 (chr6:29.50 – 33.10Mb) is broadly categorized into three major gene clusters – classical HLA class I, class II and class III (Figure 1.1) [4]. The HLA class I gene cluster spanning about 1.9Mb with a total of 26 coding and pseudogene loci; comprises of the classical class I genes (*HLA-A*, *-B* and *-C*), the non-classical class I genes (*HLA-F*, *-G* and *-E*) and the class I resembling genes (*MICA* and *MICB*). The HLA class II gene cluster of approximate 0.9Mb in length with a total of 24 coding and 15 pseudogene loci; comprises of the classical class II genes (*HLA-DP*, *-DQ* and *-DR*) and the non-classical class II genes (*HLA-DM* and *–DO*). Each of these class II genes are expressed as hetero-dimers consisting of the α and β chain. The class III region with a relatively short physical length of 0.7Mb, harbors more genes than the other two clusters and is the most gene-dense region in the human genome. Although this region does not encode any of the HLA genes, it holds several sub-gene clusters that are crucial to the immune system such as those involved in inflammation and immune regulation (Figure 1.2). There are also other major gene clusters that border the MHC region and collectively are known as the extended MHC region covering 7.6Mb [4]. These gene clusters include the histone cluster (chr6:25.72 – 27.81Mb), the zinc finger cluster (chr6:28.04 – 28.56Mb), the olfactory receptor cluster (chr6:29.00 – 29.56Mb) and the extended class II sub-region containing the transporter genes (chr6: 33.13 – 33.38Mb).

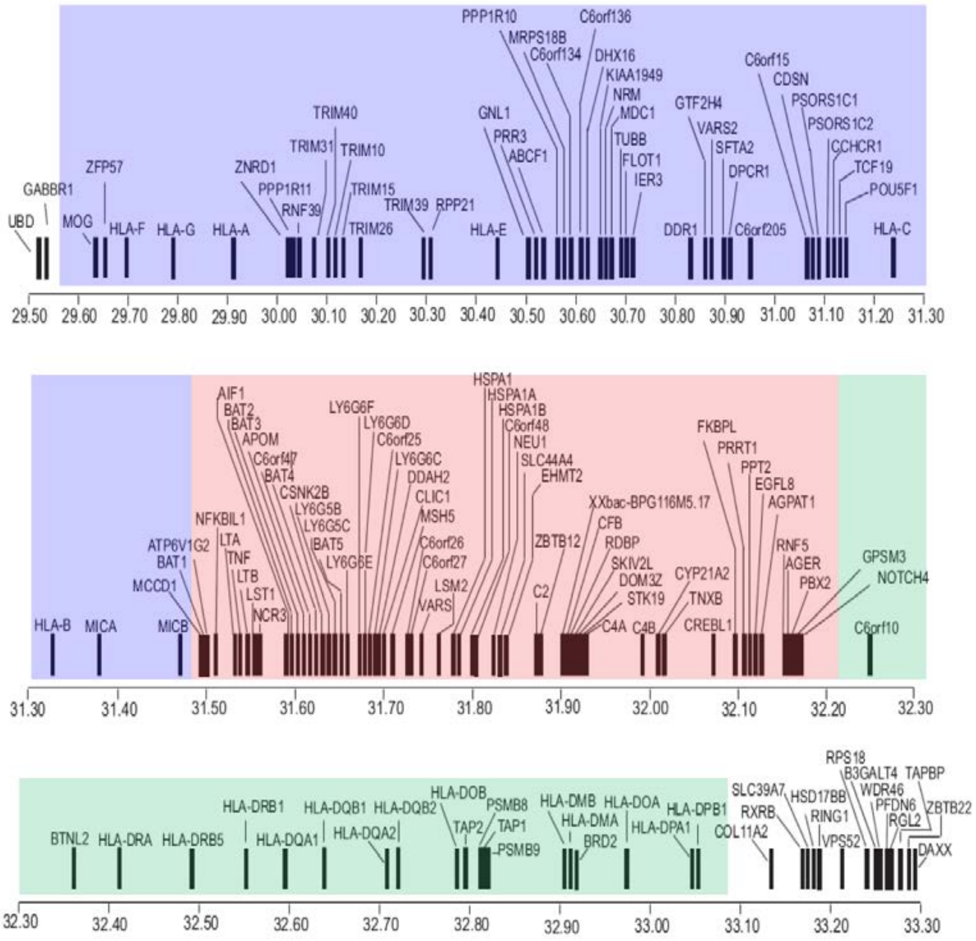


Figure 1.1 The human MHC region expressed genes. This region is divided into three major gene clusters: Classical HLA class I (blue), HLA class III (red) and HLA class II (green). This figure is adapted and modified from Trowsdale [7].

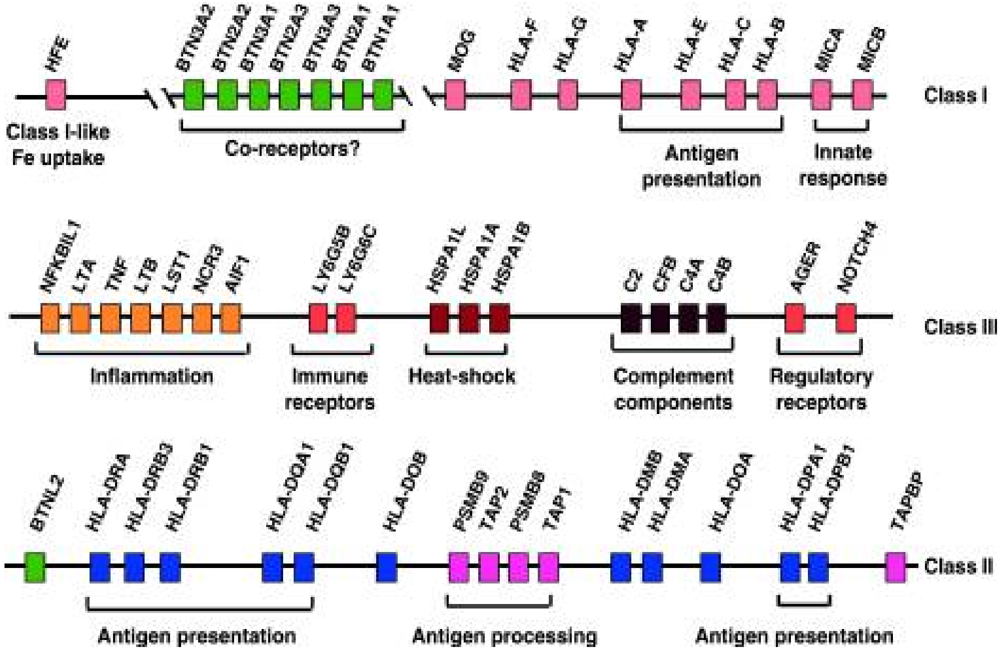


Figure 1.2 A simplified gene map of MHC showing the immune related gene based on their functions. This figure is adapted from Traherne [10].

1.3 MHC Diversity

1.3.1 HLA Molecules Diversity

Extreme sequence polymorphism is the hallmark of the HLA class I and class II genes. Since their initial discovery, hundreds of different variants of the genes or alleles has emerged and as of 2013, there are 9946 distinct HLA class I and class II alleles and amazingly new alleles are still being identified yearly, demonstrating the extreme variation found in these loci. With 3086 alleles, the *HLA-B* gene is the most polymorphic gene in the human genome followed by the *HLA-A* gene with 2432 alleles (Figure 1.3) [11]. For HLA class I genes, the majority of the nucleotide differences are located in the exon 2 and 3; and in exon 2 of the HLA class II genes with the exception of *HLA-DRA* gene. These exons are responsible for the coding of the antigenic peptides binding domain, hence non-synonymous nucleotide changes leading to amino acid modification in this domain alter the HLA molecule antigenic peptides binding capability [12]. Most of the HLA genes nucleotide variations are found in the exonic region as described above; this is in contradictory to other functional genes in the human genome where most variants are located in the introns [13]. In addition, these variations are effect of germline mutations rather than somatic mutations.

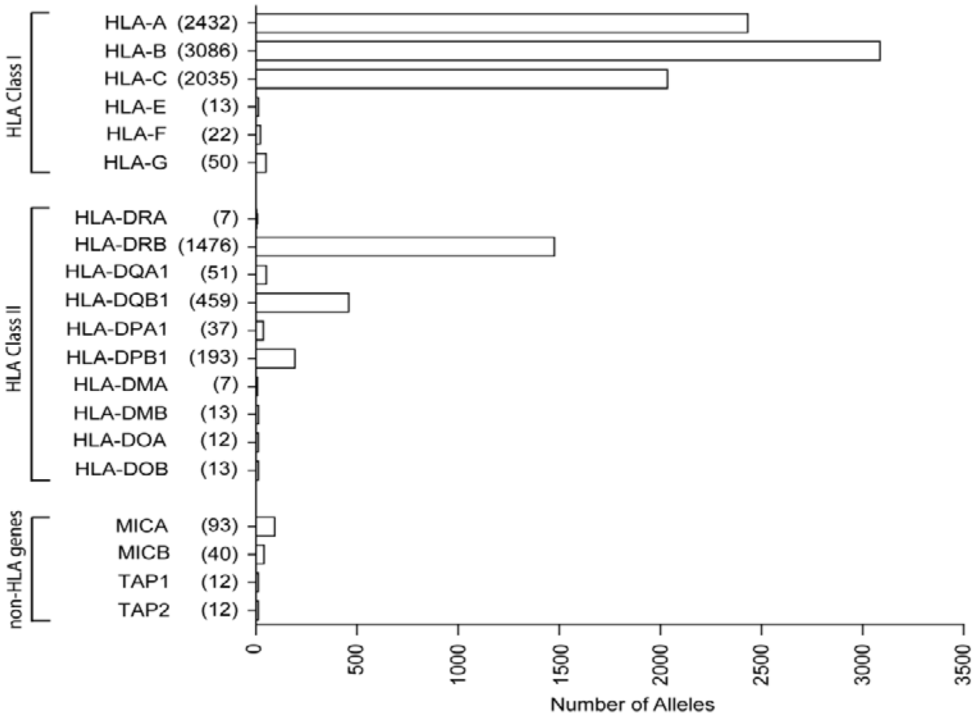


Figure 1.3 Number of alleles identified for class I, class II and non-HLA genes as of October 2013. Data was extracted from IMGT/HLA database (http://ww.ebi.ac.uk/ipd/imgt/nomenclature/).

The ever-increasing number of new HLA alleles being discovered necessitates a standardized nomenclature for the naming of the alleles. The WHO Nomenclature Committee for Factors of the HLA system established a standard HLA allele specification where each HLA allele is denoted by its gene name and an asterisk; this is followed by a distinct identification of four sets of numbers delimited by colons. The first set of number defines the allotype; alleles that differ by nucleotide changes that alter the amino acid in the encoded protein are discriminated by the second set of numbers while alleles differ by synonymous nucleotide changes in the coding region are differentiated by the third set of numbers. The last set of numbers defined the changes in the non-coding region.

1.3.2 Genetic Variation across the MHC Region

Besides the need to characterize of the HLA gene variation, information on the sequence variability in the MHC region as a whole is equally important; and in particular is relevant in the context of MHC disease association as well as the genealogical relationship between individuals or ethnic groups. This endeavor was first taken upon by the MHC Haplotype Project which aims to define the common sequence differences within the MHC region for MHC disease susceptibility studies [14-16]. Using bacterial artificial chromosome cloning (BACs) and shotgun sequencing, these studies sequenced eight common MHC haplotypes of European ancestry established from HLA-homozygous consanguineous B-lymphoblastoid cell lines and in the process characterized more than 37,000 single nucleotide variants and 7,000 short insertions/deletions. This extensive catalogue of variations enabled the establishment of linkage disequilibrium maps that facilitated the localization of susceptibility loci pertaining to multiple sclerosis and type I diabetes [17, 18].

Subsequently, independent groups employed a variety of approaches such as targeted amplicons-based methods and a combination of targeted sequence enrichment and next-generation sequencing to interrogate the distribution of variation across the MHC region [19-21]. Despite these impressive efforts, the MHC sequences described to date are mostly reflective of the European population; thus more efforts are still needed to acquire an adequately sized reference MHC sequences for other ethnic groups.

1.3.3 MHC Hyper-variable Regions

Several genomic regions within the MHC have undergone repeated segmental duplications resulting in structural copy number variation; most notably the HLA-DR region and in the class III region that holds the complement component C4 gene. The HLA-DR hyper-variable region contains a variable number of functional and pseudo *HLA-DRB* genes that result in different alternative arrangements of the *HLA-DRB* genes (haplogroups) [22]. To date, five such haplogroups have been identified; each possessing a specific arrangement of the *HLA-DRB* genes (Figure 1.4). The boundaries of all the haplogroups are characterized by the *HLA-DRA* gene at the telomeric end and the *HLA-DRB1* gene at the centromeric end and all the haplogroups possess the *HLA-DRB9* pesudogene located adjacent to *HLA-DRA*.

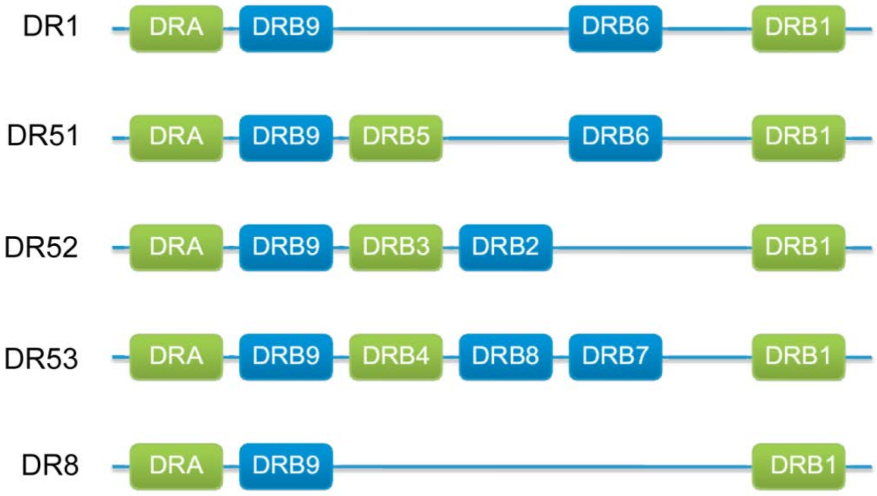


Figure 1.4 Five major HLA-DR haplogroups structure. Green boxes denote functional genes and blue boxes denote pseudogenes. This figure is adapted and modified from Marsh et al [23].

Additional HLA-DR genes are positioned between the *HLA-DRB9* and *HLA-DRB1*; the arrangement, the number and the type of HLA-DRB genes in this segment are dependent on the haplogroups. Interestingly, through in-vitro and in-vivo studies, differential transcripts expression was observed between haplogroups in resting peripheral B cells and within the same haplogroup the different DRB genes had equivalent transcripts abundance [24, 25]. However, the functional effect of having additional or lesser HLA-DRB gene(s) remains unclear.

The other region with sophisticated segmental duplication is in the class III region is called the RCCX module, named for its gene content, comprises of the serine/threonine Kinase 19 (*RP1*), the complement component (*C4*), the cytochrome P450 (*CYP21*) and the tenascin X (*TNX*) gene map to the chromosome in a chronological manner. Within the module, different variants of these genes can be presented; *RP1* or *RP2* (pseudogene), *C4A* or *C4B*, *CYP21A* or *CYP21A1P* (pseudogene) and *TXNA* (pseudogene) or *TNXB* [26]. The *C4A* and *C4B* can also exist in either long (*C4L*) or short (*C4S*) version which is differed by the presence or absence of the HERV-K endogenous retrovirus in the intron 9. Intriguingly, added to the complexity, a single chromosome can hold one to four copies of the RCCX module (Figure 1.5) [26]. Population studies had revealed high frequency of heterozgosity in the RCCX module configuration [27] and this can drive unequal crossovers during meiosis resulting in the acquisition of deleterious mutations [28-30]. In fact, there is a strong association of the RCCX copy number variation with human systemic lupus erythematosus (SLE) where low RCCX copy number was linked to increase in SLE risk [31].

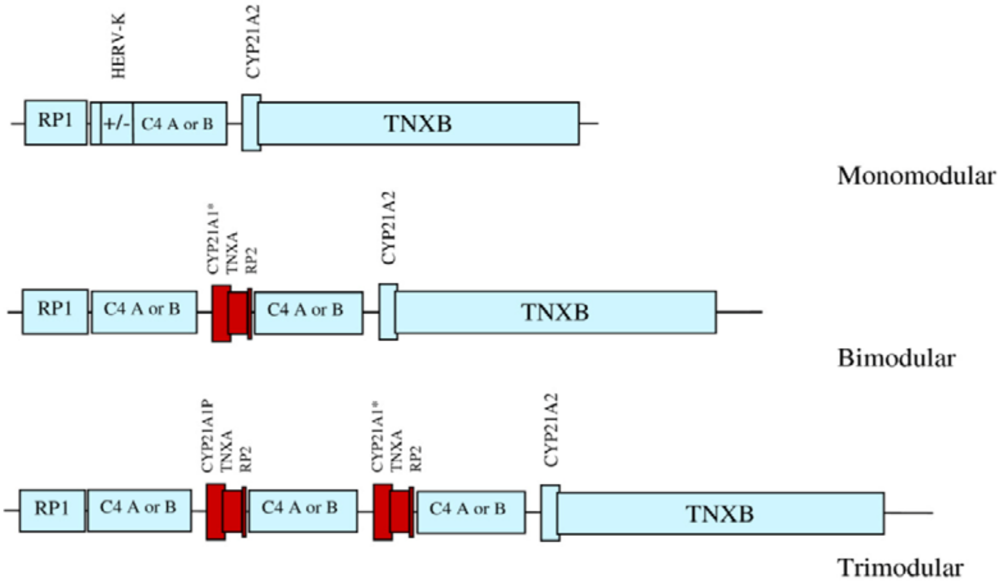


Figure 1.5 Organization of human RCCX modular variation. The *C4* gene can either exist as *C4A* or *C4B* with either the presence (*C4L*) or absence (*C4S*) of the endogenous retrovirus HERV-K. Pseudogenes are presented in red. This figure is adapted from Sweeten et al [32].

1.4 Haplotypes, Linkage Disequilibrium and Recombination

High linkage disequilibrium (LD) is a distinctive feature of the MHC region. Indeed, inheritance of non-random association of HLA alleles at numerous loci was observed within the MHC [33, 34] and this leads to the concept of "polymorphic frozen blocks" where combination of identical block sequences are shared among individuals in populations commonly termed as haplotypes [35]. Strong LD is noted between *HLA-A* and *HLA-B*; *HLA-B* and *HLA-C* as well as *HLA-DRB1* and *HLA-DQB1* evidenced by the extremely low recombination rates between these loci [36]. In addition, there exists long conserved sequences across multiple HLA alleles spanning over several mega-bases that are observed at relatively high frequency within populations and is referred as "Conserved Extended Haplotypes" (CEH) or "Ancestral Haplotypes" [37]. These CEHs appear to account for 33% of the total European MHC haplotypes and majority of the population possess at least one of these CEHs or their recombinants [37-39]. Despite the proposal of numerous theories, the mechanism behind the CEHs formation and age of their divergence from a common ancestor remain ambiguous [10, 13]. Interestingly, many of these CEHs are reported to be associated to multiple complex diseases. This is best exemplified by the A1-B8-DR3 CEH found in the Northern European population which is remarkably associated with several diseases such as for type 1 diabetes, systemic lupus erythematosus, rheumatoid arthritis and IgA deficiency and various other diseases [40-42].

Recombination within the MHC region plays an important role in maintaining the haplotype diversity in populations where the reshuffling of genomic segments leads to the generation of new haplotypes [43-45]. In the earlier years, the analysis of recombination activity in the human was restricted to pedigree studies [46]. Later, with the maturation of the sperm typing technique, enables screening of thousands of single sperms and this approach generated a high resolution MHC recombination map that definitively identified six recombination sites within the MHC [45, 47, 48]. However, due to the technical difficulties, such studies are limited in scale and are only male-specific. The advent of high density SNP genotyping assay allows the characterization of LD patterns which can be used to infer recombination events across the MHC region. Systematic analysis of the SNP genotype data from the European population revealed non uniform recombination and LD patterns across the MHC where regions of high LD are flanked by spikes in recombination activities [49-51]. In addition, these studies are able to confirm the recombination sites derived from the sperm typing experiments as well as detect novel recombination sites. A subsequent study went on to characterize the LD patterns in multiple populations and demonstrated that the MHC haplotypic structure is dependent on the underlying HLA allelic combinations [52]. On the whole, the current knowledge of MHC recombination pattern is derived from the LD structure analysis of admixed-population data and did not account for the background HLA allelic; hence recombination site in particular those specific to a single ethnic or population group could have been missed out. The availability of a comprehensive population-specific recombination and LD map can facilitate the mapping and localization of genetic segment associated with diseases in the MHC [53, 54].

1.5 Immune Function

The MHC molecules are essential components for the human immune surveillance. Both the MHC Class I and Class II molecules interact with antigenic peptides and present the peptides on the cell surface to CD8+ and CD4+ T cells respectively to initiate immune responses. The MHC class I molecules are expressed in all nucleated cells and present intracellular peptides of length 8-15 amino acids that are processed by proteasomes [55] (Figure 1.6). Furthermore, the presented MHC class I molecule-peptides complexes serve as ligands for the inhibitory killer cell immunologobulin-like receptors (KIRs) on natural killers (NK) cells and KIRs interactions with the MHC class I complexes inhibit the activation of NK cells. Reduction in the MHC class I expression in malfunction cells such as viral infected and tumor cells result in NK cells activation and elimination of the malfunction cells [56]. Unlike the MHC class I molecules which are expressed in most cells, the MHC class II molecules are predominantly expressed in professional antigen-presenting cells such as the dendritic cells and marcophages. MHC class II molecules bind to antigenic peptides generated from extracellular proteins processed by the endocytic pathway [55]. Cross-presentation is also possible where peptides derived from the endocytic pathway bind to the MHC class I molecules and are presented to CD8+ T cells [57]. It is well established that the MHC molecules can bind to a large repertoires of antigenic peptides within its peptide-binding groove [58].

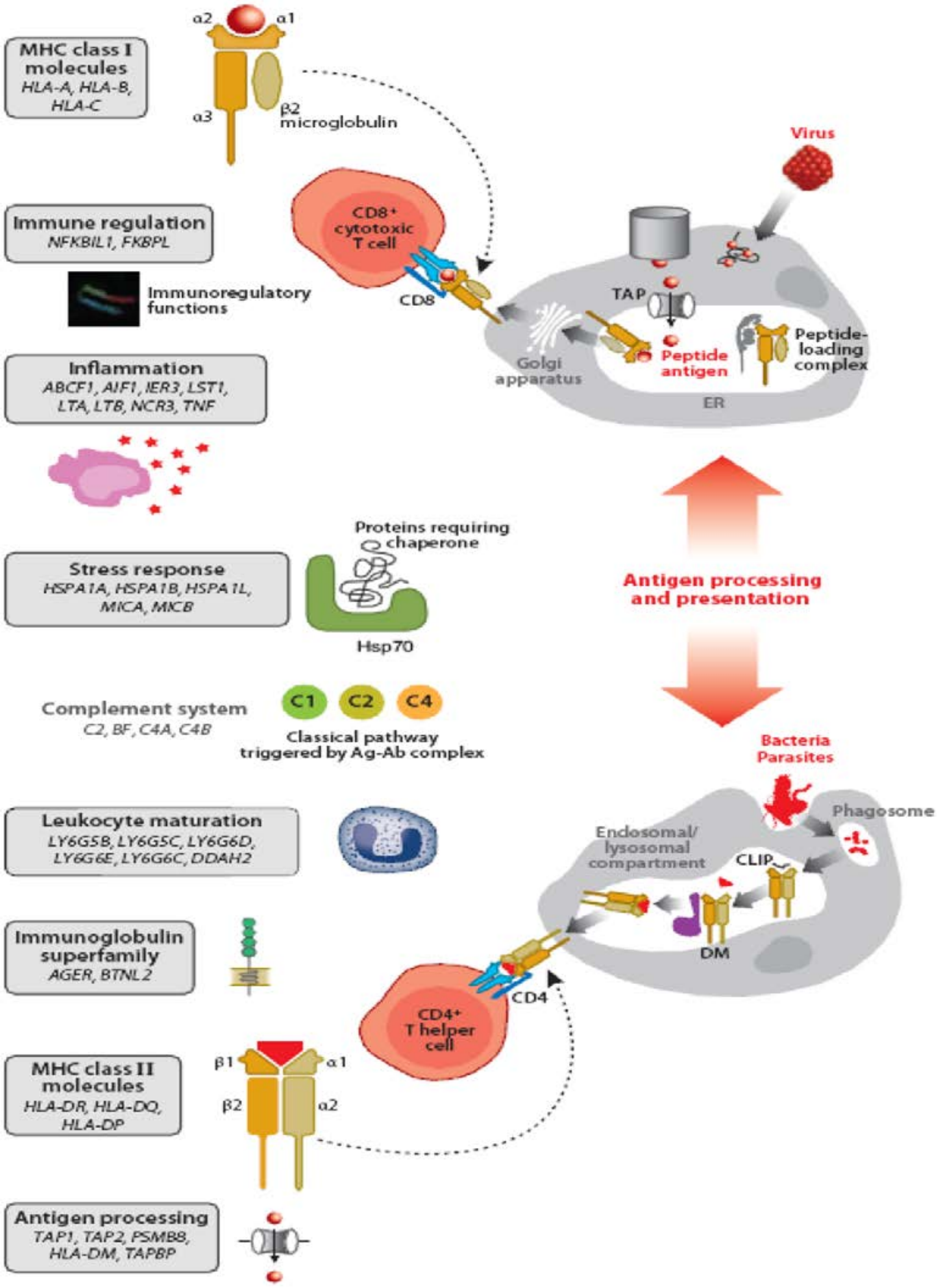


Figure 1.6 Immune functions of genes within the MHC region. MHC class I and II genes as well as non-HLA genes such as *TAP1*, *TAP2*, *PSMMB8/9*, tapasin play key roles in the antigen progressing and presentation pathway. The MHC region also contains genes such the *TNF* and *MICA/B* that have crucial roles in the human immunity system. This figure is adapted from Trowsdale and Knight [12].

Extreme polymorphism in these molecules especially in the peptide-binding groove ensure the broadest range of antigenic peptides to be recognized and hence protection against a variety of pathogens. It is interesting to note that the MHC alleles have overlapping peptide-binding specificities and this peptide-binding promiscuity is observed in alleles within each HLA gene as well as across alleles from different HLA genes [59-61].

Besides the MHC molecules, the MHC region also harbors many genes with important immune functions (Figure 1.6). Most notably is the set of genes that encode proteins involved in the antigen processing pathway. The *TAP1* and *TAP2* genes located in the class II region encode the transporter protein responsible for the delivery of the peptides into the endoplasmic reticurlum (ER) where the peptides are loaded on the MHC class I molecules and the *TAPBP* encodes the tapasin protein which is a dedicated MHC class I molecules chaperone involved in the peptide loading process in the ER [62]. In additional, the *PSMB9* and *PSMB8* encode the components of the proteasomes responsible for the fragmentation of cytosolic and nuclear proteins into short length peptides. It appears that the clustering of these genes related by their functions facilitated the gene expression and genetic exchange between the linkage genes [10]. Another important set of genes is the stress response genes *MICA* and *MICB*. These highly polymorphic genes encode molecules that serve as ligands for NKG2D which is an activating receptor expressed in NK cells [63]. In events of stress, infection or during tumorgenesis, *MICA/B* are over expressed on the surface of many cell types and the binding of these molecules to NKG2D activate the NK cells leading to the cytoxic response to the *MICA/B* expressing cells [64]. There is also a cluster of genes in the class III region that are involved in inflammation in particular the tumor necrosis factor (*TNF*) encoding gene. *TNF* is a proinflammatory cytokine involves in cellular and inflammatory reaction and it was reported that polymorphism in the TNF regulatory region influence the amount of TNF production which could affect inflammatory responses [65, 66].

1.6 Maintenance of Genetic Diversity in the MHC

The mechanisms that drive and maintain the extreme genetic polymorphism at the MHC loci have been an interesting field of study on its own especially in evolutionary biology. Due to the MHC's importance role in pathogen resistance, it has been long suggested that pathogen mediated balancing selection is the driving factor behind the maintenance of MHC diversity [67, 68]. Three main mechanisms of pathogen mediated balancing selection, supported by strong theoretical evidence, have been hypothesized; heterozygote advantage through over-dominance model [67], negative frequency-dependence [69] and fluctuating selection [70].

The heterozygote advantage through over-dominance states that individuals who are heterozygous at the MHC loci have a boarder range of recognized non-self antigenic peptides and as the result has a greater fitness against pathogens than individuals who are homozyguous at the MHC loci. Hence to improve pathogen resistance, MHC diversity is maintained in the population [71]. Evidence through empirical means has demonstrated that optimize rather than maximize amount of MHC diversity provides the greatest fitness advantage [72, 73]. It was put forward that having overly high amount of MHC diversity might result in the restriction of T-cell variation because of the removal of T-cells that response to MHC molecule self-peptide complex

[74]. The second likely MHC diversity driving mechanism is called the negative frequency-dependence whereby pathogens undergo selection to evade the recognition of the most common MHC alleles in the population and hence these common alleles are selected against and decrease in frequency in the population; while the frequency of novel alleles that provide improved resistance to the pathogen increases [75]. As the old alleles become rare, the pathogens resistance against them is diminished and causes their frequency to rise again. This recurring co-evolutionary competition results in the frequency fluctuation of the pathogens and the MHC alleles; in the process sustaining the MHC diversity in the populations [69]. The third proposed mechanism is the fluctuating selection whereby the existence of different pathogen strains in different populations results in the selection of different subsets of MHC alleles across different time and/or space [70]. As opposed to the negative frequency-dependence, selection imposed by fluctuating selection is directional rather than cyclical and pathogen evolution is independent of the MHC selection [9]. Determining which mechanism as the dominant factor is difficult; it is more likely that these mechanisms operate in conjunction with each another in the maintenance of the MHC diversity [76]. Besides pathogen mediated balancing selection, HLA intra-locus and inter-locus gene conversion via homologous recombination of short DNA fragment can contribute to the generation of the MHC diversity especially in the peptide-binding groove [77]. For instance, the HLA-B*46:01 allele is product of the inter-locus gene conversion of HLA-B*15:01 and HLA-C*01:02 allele [78]. Some of these newly generated alleles will establish in the population through genetic drift and selection.

The above models discussed are all established to understand the evolutionary processes that give rise to the observed polymorphisms in a single MHC locus and hence these models are unable to provide satisfactory explanation for HLA haplotype (multiple genes) diversity and the LD that creates this genetic fixation. To account for these, an alternate theory named associative balancing complex is proposed [79]. This theory states that deleterious nucleotide changes are accumulated in large genetic segments fixed by high LD (haploblocks) via the Muller's ratchet mechanism. Due to the high gene density in these segments, the deleterious nucleotide changes are often expressed as heterozygotes; as result purifying selection is ineffective in clearing these deleterious nucleotide changes and recombination is suppressed. Thus, the deleterious nucleotide changes are fixed into the haploblocks and propagate through generations.

1.7 Disease Associations in the MHC

The MHC region is linked to many diseases including infectious, inflammation-related and autoimmune conditions. In fact, no other region in the human genome is associated with more diseases than the MHC region [40]. Many of the early disease associations in the MHC were identified through hypothesis-based candidate gene approach by studying the HLA genes [80]. The tight association of genes in the MHC region allows the HLA genes to be used as the focal point to detect disease association; however isolating the causative genes are challenging. The availability of high-throughput single nucleotide polymorphisms (SNPs) enables the implementation of unbiased and hypothesis-free large-scale genome-wide association studies (GWAS). These GWAS studies detected the association of common genetic variants within the MHC region with a range of diseases and; to date, more than 100 MHC linked diseases have been identified and replicated in independent works (Figure 1.7). It is noted a number of these diseases such multiple sclerosis [81], nasopharyngeal carcinoma [82] and rheumatoid arthritis [83] are implicated by multiple genetic variants across the MHC region, suggesting multiple MHC genes may contribute to the disease condition. This multiple gene effects on diseases is not surprising given that genes within the MHC region are functionally involved in the similar pathway or system.

The prominent role of the MHC region in disease is without doubt; however establishing a direct genetic link between a MHC gene and a disease is problematic and is often confounded by at least three factors: the effects of multiple genes, high gene density and the extreme LD in the MHC region [12]. As the result, these prevent the unambiguous detection of disease causative variants in the MHC region The inability to identify disease-causing variants hampers the efforts to understand how these variants or genes within the MHC region influence the underlying mechanism that leads to the disease progression and development.

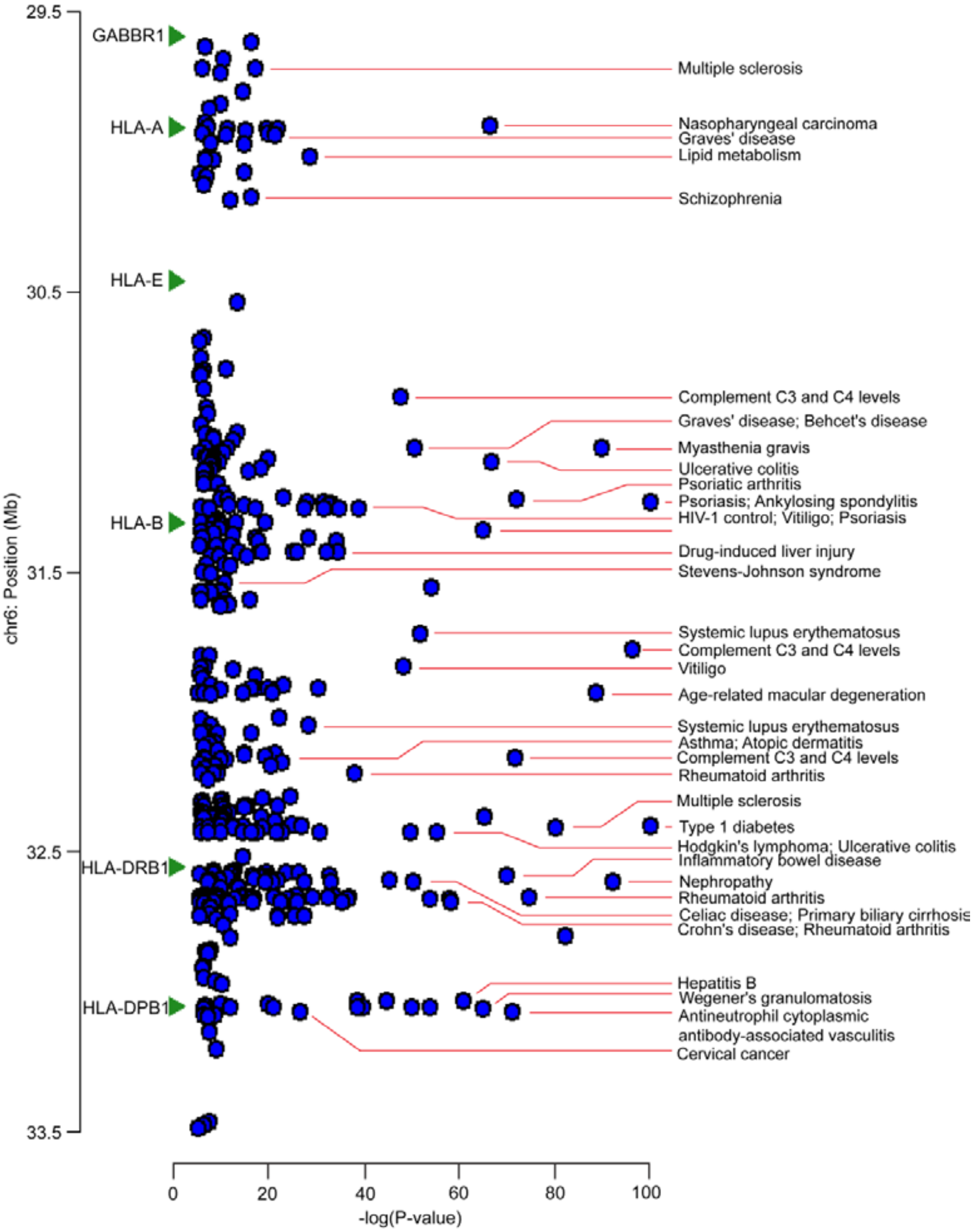


Figure 1.7 Disease-associated SNPs in the MHC region identified through genome-wide association studies. Each blue circle denotes the position of the SNP and its corresponding P-value. The SNP-disease associations listed are limited to those with P-values < 1x10^{-5}. The data was extracted from the NHGRI GWAS Catalog [84].

1.7.1 Infectious Diseases

Given that host response against viral infection depends heavily on HLA-restricted T-cell response, it is not surprising that most infectious diseases are associated with the HLA loci. Indeed, it was found that resistance to viral infection is linked to the HLA loci polymorphisms; most notably in HIV whereby escape variants of HIV-1 generates peptides that avoid T-cell recognition [85, 86]. For instance, the association of HLA-B*35 subtypes are linked to rapid HIV disease development and it appears that HLA-B*35-restricted HIV -1 variants evade CD8+ T cell recognition by affecting peptide binding as well as T-cell receptor interaction with the HLA complex [87]. Other than the influence the HLA loci polymorphism on viral infection, a variant at the 3' end of *HLA-C* was found to be associated to low HIV-1 viral load. It was revealed that the change of nucleotide at this position allows the binding of a miRNA and causes the down-regulation of *HLA-C* expression [88]. Heterogeneity at the HLA loci can also influence the clearing of viral infection and mortality. Individuals who are heterozygous at the HLA alleles have advantage over those who are HLA homozygous in the outcome of infection which in principal HLA heterozygosity increases the viral peptide repertoire pool and response to infection [89, 90]. In hepatitis B viral infection, heterogeneity across the HLA class II loci was demonstrated to have more favorable disease outcome [91].

1.7.2 Autoimmune Diseases

The MHC region is the major genetic risk contributor to most if not all autoimmune diseases [92-94]. GWAS studies have consistently revealed the associations of genetic variants within the MHC region in particular the HLA loci with autoimmune diseases and often these associations present the highest statistical signal and have considerably greater effect sizes than other part of the human genome. Despite the abundance of genetic studies linking the MHC region to autoimmune diseases, the underlying mechanism behind the MHC association still remains ambiguous. It has been long proposed that the interaction of T cell receptor with a self-peptide MHC complex could trigger an autoimmune response; although there is little or no evidence to prove the identity of these self-peptides [7].

Recently, large-scale studies with adequate statistical power revealed multiple independent associations within the MHC region in 7 prominent autoimmune diseases [95, 96]. These studies demonstrated that disease associations were not just restricted to the HLA class I and class II genes; but also possible independent contributions from non-HLA genes (Figure 1.8). For example, in systemic lupus erythematosus, in addition to the HLA genes, associations were found in *TNF*, *C4* and *Notch4* genes. Furthermore, these studies were able to show the primary HLA alleles driving the association; HLA-DRB1*15:01/HLA-DRB1*03:01 in systemic lupus erythematosus and HLA-B*44:02/HLA-DRB1*15:01 in multiple sclerosis. Despite these progresses, the dissection of individual genes contribution for psoriasis remains elusive. Psoriasis was associated with a genomic segment containing *HLA-C* and it was reported that both the HLA-C*06:02 and the nearby *C6orf10* were linked to the disease susceptibility [97, 98]. However due to the high gene density and LD in this region, the actual contribution of each loci remains unclear.

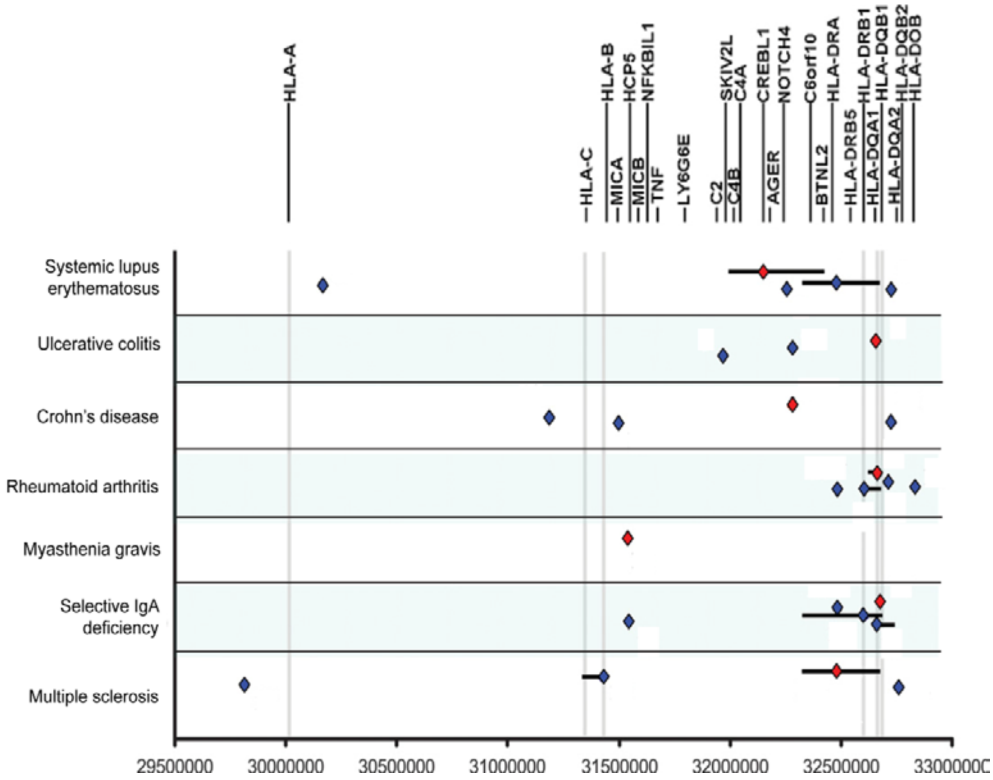


Figure 1.8 Location of loci linked to 7 autoimmune diseases across the MHC region. Primary signals associated with the disease are indicated in red while secondary signals are indicated as blue. Secondary signals are referred as independent association with P-value <0.001 that demonstrated pair-wise correlation <0.2 with the surrounding SNP loci after logistic regression analysis for the primary associations. This figure is adapted and modified from International MHC and Autoimmunity Genetics Network [96].

1.7.3 Other Conditions

Besides infectious and autoimmune diseases, other conditions such as cancers and drug-induced allergy conditions are implicated by genetic elements within the MHC region. Tumors development and progression are often linked to aberrations in genes involved antigen processing and presentation pathway [99]. It has been described that tumors in colorectal carcinoma, melamoma and cervical carcinoma alter the surface expression of HLA class I molecules to evade T-cells recognition [100-102]. In addition, defects in TAP either at the transcript or protein level resulting to the disruption of antigen processing pathway is found in several tumors [100, 103, 104].

Interestingly, several drug-induced acute reactions are associated with specific HLA alleles. Notably is the association of HLA-B*57:01 with sensitivity induced by abacavir as well as HLA-B*15:02 with Stevens-Johnson syndrome induced by carbamazepine [105, 106]. Recent developments have provided insights on the underlying mechanism that leads to the observed association. It was found that the abacavir drug molecule bound non-covalently into the peptide–binding groove of the HLA-B*57:01 molecule and changed its peptide binding specificity. This caused the binding of a novel set of self-peptides in the presence of abacavir to HLA-B*57:01 and the resulting to T cell reactivity [107]. Similarity, the binding of carbamazepine to HLA-B*15:02 altered its peptide repertoire and initiated T cell responses [108].

1.8 Epistasis

The interrelated functions of the genes within the MHC region have brought about extensive active epistasis and this genetic interaction between multiple genes plays a significant role in shaping the patterns of LD. The knowledge of epistatic interaction between multiple genes is an important component for successful disease association studies. The DRB region, involving specific alleles of two genes in complete LD – HLA-DRB1*15:01 and HLA-DRB5*01:01, is reported to be a major candidate for multiple sclerosis susceptibility [109]. A subsequent study, demonstrated the presence of functional epistasis that lead to the complete LD observed between these alleles [110]. More importantly, this study showed the HLA-DRB5*01:01 mediates the T-cell response initiated by the HLA-DRB1*15:01 and this epistatic interaction was linked to the less severe form of multiple sclerosis. Certain HLA allelic combinations, in particular the HLA class II region, are found at far higher frequency than the others in populations reflecting persistent selective pressure and such preferential allelic combination could be the effects of epistatic mechanism [111]. Interestingly, some of these haplotypes are associated to disease susceptibility such as the predisposition of the HLA-DRB1*04:01-DQA1*03:01-DQB1*03:02 haplotype to type 1 diabetes and the HLA-B*57:01-C*06 to host control of HIV [112, 113].

Besides the epistatic interaction between genes located within the MHC region; increasing there has been evidence for interaction of the HLA genes with genes outside of the region. Most notably is the association of HLA class I with the polymorphic KIR genes located on chromosome 19q13.4 whereby these genes encode activating and inhibitory receptors expressed on the NK cells. Similar to the HLA loci, the diversity seen in the KIR genes is being maintained through pathogen mediated selection and specific HLA-KIR combinations with NK activation properties are reported to confer resistance to infectious diseases [114]. For instance, the NK cells activations through the interactions of KIR3DL1 subtypes with HLA-B Bw4 alleles inhibit the HIV progression [115]. However, activating KIR-HLA pairings could also result in susceptibility to autoimmune diseases. Of prominence is the association of psoriasis with the HLA-C*06 alleles and KIR2DS1/KIR2DS2 [116, 117]. Overall, different KIR-HLA combinations give rise to variation in NK cells/T-cells activation or inhibition; resulting in resistance and vulnerability against infection and autoimmunity. This KIR-HLA class I specificity is a classical example for genetic epistasis whereby the presence of genes encoding the specific alleles are essential for functional responses.

1.9 Epigenetics

To date, most of the MHC disease association studies are central on the differences in the nucleotide sequences; but in most cases how these sequence differences are mechanistically correlated to the disease is unresolved. The study of epigenetic in the MHC region could provide meaningful explanation to these associations given its importance to the regulation of gene expression. Indeed, it is now well-established that HLA class I and class II genes are regulated by epigenetic events [118]. With the exception of HLA-G, the HLA class I genes are regulated by a number of conserved promoter elements – enhancer A, interferon-stimulated response element (ISRE) and the SXY-module (Figure 1.9). Transcription factor nuclear-factor (NF)-kB binds to the enhancer A and interferon regulatory factor (IRF) family members binds to the ISRE; thus HLA locus specific sequence variation in the promoter region would induce differential activation level among the HLA loci [119, 120]. The highly conserved SXY-module interacts with multiple transcription factors comprises of regulatory factor X (RFX), cyclic-AMP response element binding protein (CREB) and nuclear factor-Y (NF-Y) forming an enhanceosome that drive the genes transcription [121]. HLA-G, unlike other HLA class I loci, does not depend on the ISRE site for transcription regulation; instead it is regulated by CREB-1 and the transcription factor binding site is located further upstream (Figure 1.9) [122]. Under normal conditions, the expression of HLA-G is suppressed in most cell types other than trophoblast cells through chromatin remodeling via the transcription repressor Ras-responsive binding protein – 1(RREB-1) [123]. HLA-G expression is consistently found in tumor cells and thus it has been hypothesized that the expression HLA-G is a potential mechanism for the tumors to evade immunosurveillance [124, 125]. For the HLA class II genes, other than the presence of the SXY-module, their promoter region does not contain the enhancer A and ISRE (Figure 1.9). In fact, the transcription of HLA class II genes is controlled by the master activator class II transactivator (CIITA) [126]. In addition, the regulation of the class II genes is also associated with increase active histones modification and chromatin remodeling [127]. Other than the transcriptional regulation through cris-acting elements, it was found that the transcriptional insulator factor (CTCF) was involved in the class II expression through long-range chromatin interactions [128].

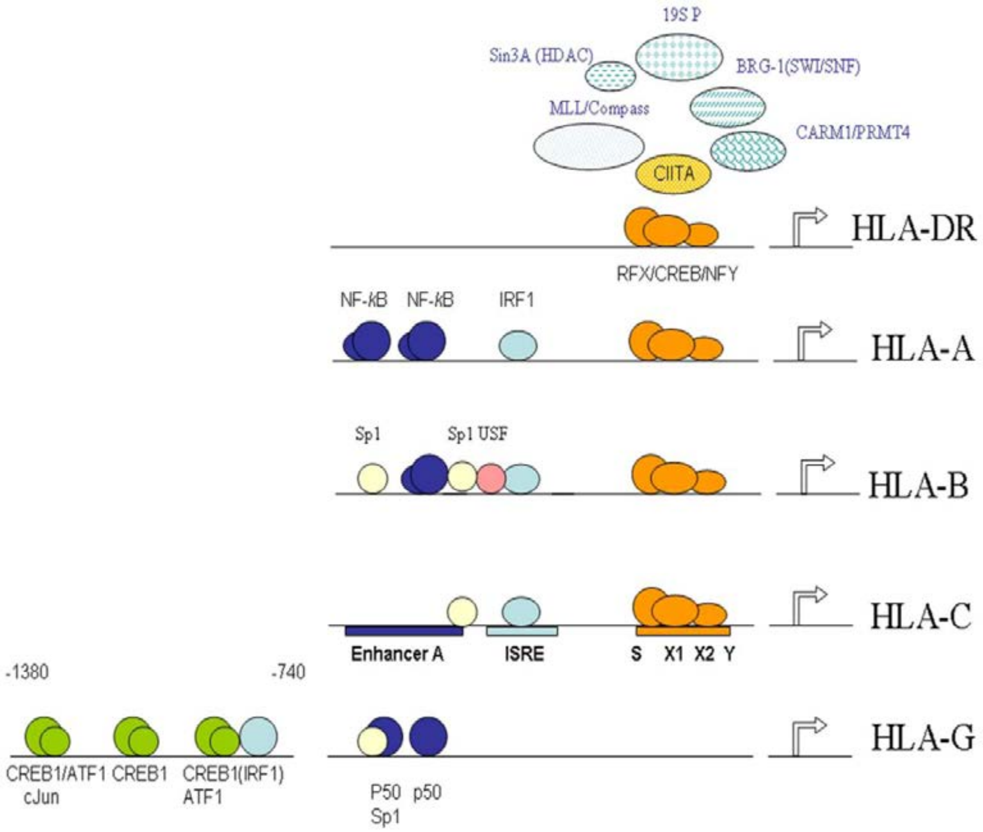


Figure 1.9 An outline of the regulatory elements, transcription factors and epigenetic activities mediating the expression of HLA class I and class II genes. With the exception of HLA-G, all genes shared the SXY-module in the gene promoter region and interact with the enchanceosome complex to regulate the transcription of the genes. This figure is adapted from van den Elsen [118].

1.10 Genomic Modulators of Genes within MHC Region

GWAS has successfully identified an abundance of new disease associated genomic loci in the MHC region that previously are undetectable using the traditional genomic approaches. However, in most cases, the exact location of the functional disease causative variants and how these polymorphisms affect gene expression leading to the observed disease phenotype remain unknown [129]. In addition, a number of these GWAS identified loci are found in the non-coding region and hence suggesting possible role in regulation of gene expression. Expression quantitative-trait mapping (eQTL) is one powerful tool to study the impact of sequence polymorphisms on transcription regulation. Indeed, many eQTL studies have showed strong associations in the MHC region [130-132]. More significantly, a recent study has demonstrated a HIV-1 control linked SNP located 35kb upstream of *HLA-C* is associated with differential *HLA-C* expression and it is later revealed that another variant linked to this SNP located at the HLA-C 3'UTR mediates the binding of a microRNA resulting to variation in *HLA-C* expression [88, 133]. However, caution needs to be taken when analyzing eQTL associations in the MHC region. This is because conventional expression microarray is unable to fully account for the extreme polymorphism in MHC region and hence this may result in spurious eQTL associations. Nevertheless, these imply the effect of individual variations in the MHC in the mediation of the immune-related processes and responses.

1.11 Objectives of Thesis

Several decades of research on the MHC region has outlined its genomic landscape providing extensive understanding to its gene organization, the linkage disequilibrium and haplotype structure. Indeed, important knowledge is gained in terms of its roles in the immune system, the nature and consequences of the extreme genetic polymorphisms in the region and its implication to numerous diseases. Despite the progress made, there is still a lack of understanding in (1) the biological and evolutionary mechanism driving the extreme genetic polymorphisms in the region, (2) the location of genetic variation with functional significance and (3) the underlying mechanism behind the numerous disease associations. The availability of MHC sequence information, in particular the structure and genomic organization of CEHs, is an essential resource to provide insights on above mention issues.

1.11.1 Key Objectives

1. Evaluate recombination profiles of HLA haplotypes from three distinct population groups and identify population-specific recombination sites.
2. Determine the presence of conserved extended haplotypes (CEHs) in the Singaporean Chinese population.
3. Establish an Asian reference MHC haplotype sequences.
4. Examine multiple MHC CEH sequences with identical haplotype profile to explore the scope of intra-haplotypic conservation and variation.
5. Determine the functional significance of intra-haplotypic variation.
6. Determine the effect of haplotype sequence variations on the transcription activity within the MHC region.

Chapter 2:

Materials and Methods

2.1 Subjects and HLA Sequence-based Typing

Peripheral blood mononuclear cells (PBMCs) were obtained with prior consent from 247 healthy Singaporean Chinese blood donors. Of these 211 are unrelated individuals while 36 are from comprising members of family trios. The genomic DNA of this collection was kindly prepared and extracted by a previous post graduate student, Chia Jer-Ming. Subsequently, B-lymphoblastoid cell lines (B-LCLs) were established from whole blood obtained from subjects who are HLA homozygous and the B-LCLs were prepared through *in-vitro* Epstein-Barr virus infection of the B-lymphocytes. The establishment of BLCLs were kindly performed and provided by WHO Immunology Center, National University of Singapore, Singapore. The two European B-LCLs - COX and QBL were purchased from the Research Cell Bank, Fred Hutchinson Research Centre, Seattle, WA.

The HLA allelic type at *HLA-A*, *-B*, *-C* and *–DRB1* loci of these Singaporean Chinese subjects were obtained from Yu *et al* [134]. This study used a sequence-based approach to interrogate the HLA allelic type. Briefly, the hyper-variable exons 2 and 3 of the *HLA-A*, *-B*, *-C* genes were examined by PCR amplification using specific primers, followed by direct DNA sequencing of the PCR products in the opposite directions. *HLA-DRB1* was sequenced and typed as previously described [135]. Purified PCR products were sequenced using the ABI BigDye Terminator v3.1 chemistry run on an ABI Prism 3100 Genetic Analyser (Applied Biosystems, USA). Excess dye terminators were removed by purification using an ethanol/EDTA/sodium acetate precipitation protocol.

2.2 HLA Homozygous Cell Lines Culture and Treatment

Under sterile conditions, frozen B58AL, B58SC, B58CF, B46BM, B46ZS, B46CM, COX and QBL B-LCLs were thawed in 37°C water bath and re-suspended in 11ml of Roswell Park Memorial Institute medium supplemented with 10% Fetal Bovine Serum (RPMI/10%-FBS). Subsequently, centrifugation was performed at 800rpm for 5min and the resultant cell pellets were re-suspended in 8ml RPMI/10%-FBS and moved into 25ml culture flasks. The cell cultures were then maintained at 37°C in a humidified incubator with 5% carbon dioxide in atmospheric air. When the cell confluency reached 85-100%, the cell medium was changed and 2/3 of the cells were transferred to 75ml culture flask and maintained at 37°C in humidified incubator with 5% carbon dioxide in atmospheric air. This process would be repeated till the number of cells required for experiments were attained.

At the 5th passage, for each B-LCL, the cells were collected and counted at approximate 1 X 10^6 cells per ml. The cells of each B-LCL were then seeded in 6-wells plate and stimulated with 200nm phorbol 12-myristate 13 acetate (PMA, Sigma) and 125nM ionomycin; equivalent amount of dimethyl sulfoxide (DMSO, Sigma) were added to unstimulated cell cultures to act as controls. Following, the cells were incubated for six hours in humidified incubator with 5% carbon dioxide in atmospheric air. After six hours, the culture were collected and pelleted down by centrifugation at 1000 rpm for 5min, 4°C. 1.5ml of supernatant was then transferred to 2ml eppendorf tubes and stored at -80°C for the ELISA experiment while the remaining supernatant was discarded. Cell pellets were then re-suspended in 500ul of phosphate buffered saline (PBS); transferred to 1.5ml eppendorf tubes and centrifuged at 1000 rpm for 5min, 4°C. The supernatant was removed and the resultant cell pellets were stored in -80°C for subsequent RNA extraction.

2.3 DNA and RNA Extraction

Genomic DNA was isolated from the cell pellets of B58AL, B58SC, B58CF, B46BM, B46ZS and B46CM BLCLs using the QIAGEN® DNeasy Blood and Tissue Kit; following the manufacturer's protocol. Extracted DNA was elute in 200ul of 10mM Tris-Cl/0.5mM EDTA (AE buffer) and stored at -20°C.

For the purpose of whole genome sequencing, high DNA quality and accurate quantification are important to achieve favorable sequencing results. To ensure these, gel electrophoresis run on a 0.8% agarose gel was performed to detect DNA degradation. The gel electrophoresis showed relatively tight band for the DNA of every cell line indicating that there was no or little degradation (Figure 2.1). The DNA quantification was performed using the Picogreen assay. Briefly, the quantification assay first requires the establishment of a 6-points DNA standard curve with concentration range from 0 to 10ng/ul of λ-DNA (Invitrogen, USA). DNA standards were generated from the 10X serial dilution of the 10ng/ul of λ-DNA working solution with the 10mM Tris-Cl/1mM EDTA (TE buffer) and 10ul of the DNA standards was pipetted into a 96-well full skirted black plate.

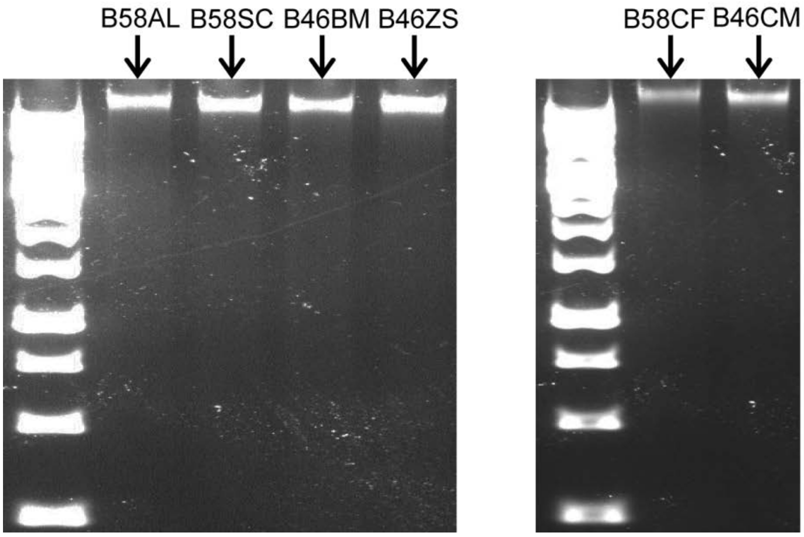


Figure 2.1 Gel electrophoresis with 20ng of DNA per lane on 0.8% agarose gel. The result indicates no or little DNA degradation.

Next, 100ul of 1X Picogreen (Invitrogen, USA) was then pipetted into each well and incubated at room temperature for 5 mins followed by centrifugation at 1000rpm for 30sec. Flurorescene output emitted from the interaction between the double-stranded DNA and Picogreen was measured using the Tecan Genios fluorescence reader (Tecan, Switzerland) and the fluorescence readings for the DNA standards were used to construct a standard curve. Lastly, using the same protocol, the concentration of the DNA samples were inferred by comparing their fluorescence readings against the standard curve. Overall, we were able to obtained at least 15ug of DNA for all the samples.

Total RNA was extracted and purified using the RNeasy Mini Kit (Qiagen, Germany); following the manufacturer's protocol. During the purification, additional steps for DNA digestion using the On-column DNase digestion kit (Qiagen, Germany) were included to ensure high quality RNA. The quality and quantity of the purified RNA were determined using the ND-1000 Nanodrop spectrophotometer (Thermo Fisher Scientific, USA) and we were able to obtained at least 15ug of good quality total RNA for each sample to be used for RNA-sequencing.

2.4 Enzyme-linked Immunosorbent Assay (ELISA) Experiment

The amount of TNF-alpha and IL6 in the supernatant collected after six hours of PMA and ionomycin stimulated were determined by using the human TNF-alpha and IL6 Quantikine ELISA kits (R&D Systems). Briefly, samples were (200ul for TNF-alpha; 100ul for IL6) loaded in triplicates into the plate coated with mouse monoclonal antibody against the proteins together with the assay diluent (buffered protein base with preservatives) provided, incubated for 2 hours and 1 hours at room temperature respectively and followed by the addition of the respective protein conjugate for additional 2 hour incubation at room temperature. The ELISA reaction was then detected by the additional of 200ul of substrate solution (lyophilized NADPH with stabilizers) per well and incubated for 30min at room temperature in dark followed by the addition of 50ul stop solution (2N sulfuric acid). The absorbance level was measured at 450nm using the Infinite® 200 PRO plate reader (Tecan, Switzerland).

The pates were washed five times with the washing buffer (1X buffered surfactant) after each step. The standard curve was established by an 8-points serial dilution of 1X calibrator diluent to be used as a reference for quantification.

2.5 RT-qPCR for ZFP57 Expression Quantification

cDNA was generated using Maxima® First Strand cDNA Synthesis Kit (Thermo Fisher Scientific) as per manufacturer's instruction using 1ug of total RNA as template. The reaction mixture comprised of 4ul of 5X reaction mix (reaction buffer, dNTPs, oligo dTs and random hexamer primers), 2ul of maxima enzyme mix (reverse transcriptase and RNase inhibitor), 1ug of total RNA and top up with nuclease-free water to 20ul. This was then incubated for 10min at 25°C and followed by 15 mins at 50°C and at 85°C for 5 mins to terminate the reaction. The resultant cDNA was diluted 10 times with RNase/Dnase free water and stored at -20°C for subsequent use.

qPCR by KAPA SYBR® FAST Roche LightCycler® 480 2X qPCR Master Mix (Kapa Biosystems, Woburn, MA) was performed with triplicates for each of the two biological replicates on the Roche LightCycler® 480 System

(Roche Applied Science). The reaction mixture consisted of 5ul Sybr green, 0.5ul of combined forward and reverse primer (2ng), 2ul of cDNA and 2.5ul of nuclease-free water. The qPCR cycling conditions were set at 95°C for 10 mins, 45 cycles of [95°C for 10 secs; 60°C for 10 seconds; 72°C for 10 seconds] and followed by 95°C for 60 seconds. Ct values calculation using the second derivative maximum method and melting curve analysis were carried out with gene-specific primer pairs. The *ZFP57* expression was normalized to Hypoxanthine Phosphoribosyltransferase 1 (*HPRT1*) and determined using the ΔCt method. Primer pair sequences for the respective genes are showed in Table 2.1.

Table 2.1 Primer sequences for *ZFP57* and *HPRT1* gene.

Gene	Forward Primer	Reverse Primer
ZFP57	TGAGGATGTGGCAGTGAATTT	GTGTTTGGGAGATGGACAAAC
HPRT1	GTAATGACCAGTCAACAGGGGAC	CCAGCAAGCTTGCGACCTTGACCA

2.6 Elucidation of RCCX Copy Number Variations (CNV) in the Cell Lines

The assay used to determine the RCCX modular duplication in each cell line was based on and developed from a modified version of the real-time PCR assay as previously described in Wu *et al* [136]. This modified version used SYBR Green chemistry instead of Taqman chemistry to measure the level of mRNA. Primers specific for *C4A*, *C4B*, long *C4* (*C4L*), short *C4* (*C4S*) were designed to resolve the RCCX modular duplication (Table 2.2). The difference between *C4L* and *C4S* is the insertion of the endogenous retrovirus HERV-K segment between exon 9 and exon 10. As such the amplicons for *C4L* and *C4S* shared a common forward primer, and their reverse primers were designed to differentiate between the long and short *C4* gene (Figure 2.2). In addition, the copy number of the *TNXA* gene, which equals to the number of RCCX modules minus 2, was also interrogated. The assignment for the number of copies of each targeted gene involved two calibration steps. The first calibration step was a quantitative real-time PCR endogenous control using the *RP1* gene, which is positioned upstream of the RCCX module and always present as 1 copy per chromosome. Quantified levels of target genes were compared to levels of *RP1* in order to obtain relative copy numbers of target genes. However, in this approach, there is an intrinsic underestimation of the targeted gene copy number. To correct for this underestimation, a second calibration step was performed. A calibrated plot created from 13 reference human cell lines (COX, QBL, MOU, PGF, SSTO, DBB, WT51, MADURA, CB6B, WT8, DAUDI, MANIKA and HOM) with known RCCX modular number was used to unambiguously assign the targeted gene copy number (Table 2.3).

The genomic DNAs of these reference cell lines were purchased from the Research Cell Bank, Fred Hutchinson Research Centre, Seattle, WA. For each sample, the number of copies of (*C4A* + *C4B)/2*, (*C4L* + *C4S)/2*, and (*TNXA* + 2)/2 were the same, acting as an internal validation.

Table 2.2 Primer sequences to determine the copy numbers of the *C4A*, *C4B*, C4 long, C4 short, *TNXA* and *RP1* genes.

Gene	Forward Primer	Reverse Primer
C4A	CCTTTGTGTTGAAGGTCCTGAGTT	TCCTGTCTAACACTGGACAGGGGT
C4B	TGCAGGAGACATCTAACTGGCTTCT	CATGCTCCTATGTATCACTGGAGAGA
C4L	TTGCTCGTTCTGCTCATTCCTT	GTTGAGGCTGGTCCCCAACA
C4S	TTGCTCGTTCTGCTCATTCCTT	GGCGCAGGCTGCTGTATT
TNXA	TCCTGCAGTCATCTTTGTCTTCAG	GAGCTGCAGATGGGATACCTTTAA
RP1	GACCAAATGACACAGACCTTTGG	GACTTTGGTTGGTTCCACAAGTC

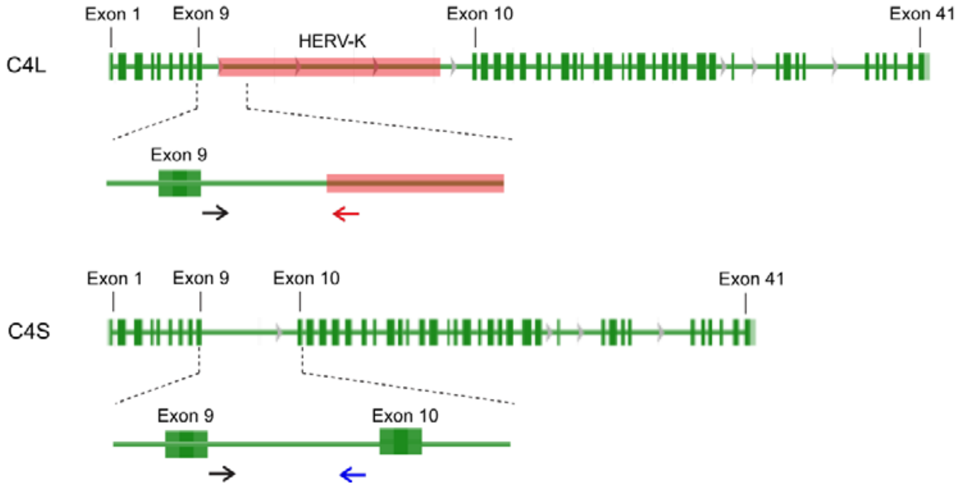


Figure 2.2 Verification of the copy numbers of *C4L* and *C4S* genes. An insertion of the endogenous retrovirus HERV-K segment is found in *C4L* but not *C4S*. To differentiate the two genes, the reverse primer for the *C4L* amplicon targets the 5’ sequence specific to HERV-K (red) while the reverse primer for the *C4S* amplicon targets the upstream sequence of exon 10 specific to *C4S* gene (blue).

Table 2.3 RCCX modular structure in 13 reference human cell lines to establish the calibration plot.

Cell line	RCCX Structure	C4A	C4B	C4L	C4S	TNXA
COX	Monomodular	0	2	0	2	0
QBL		2	0	2	0	0
MOU		0	2	2	0	0
PGF	Bimodular	2	2	4	0	2
SSTO		2	2	4	0	2
DBB		2	1	2	2	2
WT51		4	0	4	0	2
MADURA		2	2	0	4	2
CB6B	Quadrimodular	4	4	6	2	6
WT8	Heterozygous	1	2	3	0	1
DAUDI		2	1	2	1	1
MANIKA		2	3	2	3	3
HOM		3	2	4	1	3

2.7 SNP Genotyping and Selection

SNPs interrogation of MHC region in the Singapore Chinese (CHSG) population was performed on the Illumina GoldenGate MHC Panel platform (Illumina, USA). This platform was designed examine the genotype of 2360 SNPs residing in the MHC genomic region from chr6:28.97 – 33.88Mb. The SNP coordinates were mapped to the Human Reference Sequence Assembly 36.1 (NCBI 36.1). Genotyping results were filtered using the following criteria: SNP loci deviating from Hardy Weinberg equilibrium using a Fisher's exact test at a significance level of 0.001; SNPs loci with a call rate of less than 95% and SNP loci with a minor allele frequency of less than 5% were discarded. In addition, for familial data, SNP genotypes that were discordant with the parental structure in more than one family were discarded. After the quality control checks, 1877 SNP loci were left for further analysis.

Genomic DNA of the six samples (B58AL, B58SC, B58CF, B46ZS, B46BM and B46CM) was subjected to genome-wide SNP genotyping using the Illumina Human 1M-Duo BeadChip Kit (Illumina, USA). The genotype profiles of 1,169,675 SNPs across the entire human genome were interrogated. The SNPs coordinates were mapped to the Human Reference Sequence Assembly 37.2 (NCBI build 37.2), and all samples had overall call rates of more than 95%. SNP loci that were not called in any of the samples or that deviated from Hardy Weinberg equilibrium at a significance level of 0.001 were not included in the downstream analysis.

Genotype data consisting of 30 European (CEU) trios and 30 Yoruban (YRI) trios were obtained from the HapMap phase II collection [137]. SNPs genotype data within the MHC region interrogated by the Illumina GoldenGate assay were selected and subjected to the screening procedure as described above. A total of 1360 common SNPs loci found across the CEU, YRI and CHSG populations were chosen for further downstream analysis.

2.8 Haplotype Inference

Haplotype inference was carried out using a Bayesian-based approach implemented in PHASE 2.1 [138, 139]. This approach computes the posterior distribution of unknown haplotypes based on the observed SNPs genotype to estimates the expected haplotype structures found in the sample population (prior information). This prior information is estimated using the coalescent-based model where recent haplotypes are derived from the ancestor haplotypes through recombination and mutation; as these events are relatively rare over short genetic distances, new haplotypes will resemblance to one of the observed haplotypes in the population [140]. Haplotypes for each individual is then inferred by selecting the most probable haplotype from the posterior distribution. Among the population-based haplotype inference algorithm, PHASE 2.1 is considered as the benchmark for haplotype phasing [141] and hence was selected. To ensure maximum accuracy, several factors that influence phasing accuracy [142] were taken into consideration. Rare SNP loci which are computationally difficult to resolve, would not be considered and in addition, the phasing procedure for each population group was performed independently to preserve the ethnic relatedness. For the Singapore Chinese population, the haplotype inference was performed in 2 stages. First stage involved the phasing of 12 family trios to attain 48 distinct phase unambiguous haplotypes. Second, this set would serve as “known haplotypes” to aid the haplotype inference for the 211 unrelated individuals. Each of the HLA alleles was represented by unique digit and the haplotype information was resolved together with the SNPs. The HapMap populations *HLA-A, -B* and *–C* typing were attained from [143] and phased together with the selected SNPs.

2.9 Extended Haplotype Homozygosity (EHH) and Recombination Analysis

EHH analysis was performed to assess the level of LD decay across a sample of haplotypes tagged with their respective HLA alleles. Essentially, EHH calculated at a position *x* is defined as the probability that 2 chromosomes, carrying an allele (or haplotype) of interest at an anchor locus, have identical SNPs sequence from defined core locus to the position *x* [144]. In the context of this project, this core locus is described by the HLA allele of the haplotype.

The EHH of a selected core locus *t* is computed as the following:

$$EHH_t = \sum_{i=1}^{S}\binom{e_{ti}}{2} \Big/ \binom{c_t}{2}$$

c is the number of haplotypes of a particular HLA allele
e is the number of total number of haplotypes
s is the number of haplotypes with a unique SNPs sequence

A constant EHH value indicates the transmission of a haplotype without recombination. Hence, probable recombination sites were scored when either EHH decay is observed in 2 or more HLA haplotypes across the SNPs interval or when at least 10% of a single HLA haplotype diverged from the core haplotype.

The historical recombination rates for HLA chromosomal SNPs haplotypes were estimated using the program LDHat [145]. LDHat implements a coalescent-based approach that accounts for the patterns of genetic variation and observed linkage disequilibrium to infer recombination. The estimate was performed with a block penalty of 5 and 10,000,000 iterations. LDHat presents the population-scaled recombination rates *ρ*, hence to obtain the per-generation recombination rates *r* (cM/Mb), the following equation was applied.

$$r = \rho / 4N_e$$

N_e is the effective population size and was set at 10,000 [146-148]

2.10 Runs Of Homozygosity (ROH) and Identity-By-Descent (IDB) Analysis

The genomic region in chromosome 6 from 25MB to 35MB covering the extended MHC region was selected and screened for Runs Of Homozygosity (ROH) segment analysis. The ROH analysis was implemented by an algorithm in the PLINK package [149] with the following parameters - sliding window size : 50kb, minimum length for ROH segment : 1000kb, number of heterozygotes genotype call allowed in a window : 3 and maximum distance between adjacent SNPs in order to be considered in a segment : 50kb. The genotype data of B58AL, B58SC, B58CF, B46ZS, B46BM and B46CM assayed from Illumina Human 1M-Duo BeadChip were used for ROH analysis. In total, 10215 SNPs markers were found within the genomic region of interest and were used for the ROH analysis.

Similarly genotype information of the six cell lines from the Illumina Human 1M-Duo BeadChip was used for the Identity-By-Descent (IBD) analysis. 96,387 independent SNPs not in linkage disequilibrium with one another were randomly selected from the genome-wide SNPs data and were used to test for IBD among the samples. The PLINK package was used to estimate the following IBD parameters in each sample pair: probability of genetic markers sharing 0 allele (IBD=0), probability of genetic markers sharing 1 allele (IBD=1) and probability of genetic markers sharing 2 alleles (IBD=2). Therefore, high values of IBD=1 and IBD=2 would denotes high degree of relatedness between 2 individuals while high value of IBD=0 denotes otherwise.

2.11 Genome Sequencing

The genomic DNA extracted from B58AL, B58SC, B58CF, B46BM, B46ZS and B46CM BLCLs was subjected to whole genome sequencing using the next generation technology implemented by Complete Genomics (Mountain View, CA, USA). The sequencing was performed using Complete Genomics (CG) proprietary sequencing instruments and technology [150]. Briefly, the CG sequencing technology called DNA nanoball sequencing involves the searing of the isolated DNA into approximately 500bp genomic fragments; the insertion of four synthetic adaptor sequences into each fragment followed by the circulation of the fragments to generate 70-base reads (35-base pair-end reads). Each of these circular reads is then amplified into a head to tail concatemer resulting in a long single-strand DNA. This single-stranded DNA are folded in a nanoball and captured by a microarray sequencing flow cell. The sequencing assay called combinatorial probe-anchor ligation (cPAL) chemistry is carried out where independent non-iterative sequencing reactions interrogate the nucleotide profiles of the DNA through the ligation of fluorescent probes to the DNA. There are several advantages of the CG technologies over approaches that employed sequencing by hybridization method. Firstly is the ability to interrogate simple repeats. Secondly, the independent non-iterative sequencing reaction minimizes the accumulation of sequencing errors and provides superior fault tolerance. Lastly, the cPAL approach enable massive parallel base reading in a single cycle and therefore reduces the consumption of reagents and sequencing time.

The CG proprietary bioinformatics pipeline was used to map and assemble the genome [151] and the reads were aligned to the Human Reference Sequence Assembly 37.2 (NCBI build 37.2). Using a combination of Bayesian and de Brujin graphed-based approaches, the CG assembler is able to annotate SNPs, short insertion/deletions and block substitutions. The variants called were annotated against a variety of public databases - NCBI gene annotation Build 37.2, dbSNP build 137, catalogue of somatic mutations in cancer v61 (COSMIC) and miRBase version 19 and Data of Genomic Variants version 9 (DGV).

2.12 MHC Haplotype Variation Classification and Comparison

For intra-haplotype comparisons, only nucleotides at positions with high quality score metric generated from the CG assembly protocols were considered for comparison; nucleotide at positions with low confidence score were considered as an ambiguous call and would not be used for comparison. Consensus sequence for each of the two Asian MHC haplotypes was established by evaluating the samples sequences. For each haplotype, the 3 genomes available were each divided into two haploid chromosomes, and the 6 resulting haploid chromosomes were then compared. If 2 or more chromosomes had ambiguous or low confidence calls at a position, it was no-called (N) in the consensus sequence. A variant was called if two or more haploid chromosomes showed an alternate nucleotide call.

To compare the two Asian haplotypes against the eight European MHC haplotypes sequenced by the MHC Haplotype Project, BED files of the eight European MHC haplotypes [14-16] were downloaded from http://www.ucl.ac.uk/cancer/medical-genomics/mhc/#HaplotypeData and the coordinates of these BED files were aligned to the Human Reference Sequence Assembly 37.2 (NCBI build 37.2). Genetic variants found between the MHC haplotypes were annotated using the annovar software [152]. Construction of the consensus sequences as well as the intra and inter-haplotype comparisons were performed using in-house generated R-scripts.

2.13 Assessment of Sequencing Accuracy

To assess the discordant in the nucleotide calls generated from the CG platform and the SNPs genotyping platform, a total of 48 nucleotide positions were selected for PCR re-sequencing. Primer pairs to assess the variant of interest were designed using NCBI Primer-BLAST and checked against Human Reference Sequence Assembly 37.2 (NCBI build 37.2). Only primer pair's sequences matching to the sequence of the samples were selected. This cross-checking of primer sequences with the sample genomes was performed using in-house written R-scripts. The sequences of these primer pairs and the location of the variants can be found in Table 2.4. The PCR reactions were carried out in volume of 20ul reaction mixture comprising of 2ul of 10X buffer (100mM Tris-HCl and 500mM KCl), 2.4ul of 25mM $MgCl_2$, 1ul of 2.5mM dNTPs, 0.2ul of Hi Fidelity Taq DNA polymerase (Roche Applied Science, Germany), 3ul of 2ug forward primer, 3ul of 2ug reverse primer, 3.4ul of nuclease-free water and 5ul of 20ng DNA template.

Table 2.4 Re-sequencing experiments to assess mismatches between SNP genotyping and CG sequencing platform in cell lines carrying A33-B58-DR3 and A2-B46-DR9. Ta: annealing temperature.

A33-B58-DR3

Forward Primer	Reverse Primer	Position Examined	Length	Ta
ACTGACAGAATGAACCTGCAGAC	AATCACTCTCTGGTACAGGATCTGG	29,796,376	722	60
TGAGAACTGGCGGGGAGATA	TCTCTTGCTGGCTCAGCTTT	29,819,909	490	60
CTGACTCATATCAAGGGCCAGAAA	AGAGAGGAAAGTCAGGACACAATAC	30,383,046	719	60
ATAAAAACAGGCTGCATGTGGTAAA	AGTTGAGGTTTTTCTGTTATGCCTG	30,418,354	776	60
CAGAACCAGGGAGATGAGACATAC	TGTTCCTGCTTCTCTTTTCACTTTC	31,170,514	621	60
GAACATATGCTACAAAAGGCCAGAG	GGTGTGGAGAAGGCTGTGGG	31,321,327	774	60
CTCTTGAAGGACTCTGGGTTAGAAG	GCACCAGAGTTCAAGAGAGAAAATTA	31,639,979	707	60
CAATGCTTATAGGGTATCCCCAGTC	GCAGTGTACACACACAGATACTGAT	31,655,438	597	60
TAGGGTCTCTAATCTCCAAAACACC	CTAAAAGCCAGAGCTCCCAGTCC	31,697,558	739	60
CCTTTATGAGACCTGCATTGAACC	GGTACTCCAACACTGATCATAGGG	32,130,937	660	59
TCAGATTGAATTTTTCCTCCCTTCC	GATTACAGCTTCCACAAGTTCCATT	33,036,549	560	59

A2-B46-DR9

Forward Primer	Reverse Primer	Position Examined	Length	Ta
AGTACATGTAGACAGCTCACAGT	GCACAGGGAATGTGTTCTCG	29,801,958	420	58
GGGGTTTCTTTGCATTGGATGTATT	TTGTCTCTTGATACCACAAGGAGAT	29,913,509	733	60
ATAGAATTAGAAAGAGGCTGGGGTC	GTGCTAATGAAAGTTGGGCCTTAG	29,942,191	708	59
CTTTCAGTTCTCTTCTGTGTCTCCA	AGTATATTAGGTTAGCGGGTGGTAG	30,704,985	697	60
CCGTGGGGATGGCTAGAAAA	CCCTGAGGGAATCTGGGGTA	31,079,236	538	59
GATTCCAGACTTGGAGTTTCAACAG	GAGTAAAGGACTGAGAGGATGGGA	31,082,304	563	59
CCTTTATGAGACCTGCATTGAACC	GGTACTCCAACACTGATCATAGGG	32,130,937	660	59
TCAGATTGAATTTTTCCTCCCTTCC	GATTACAGCTTCCACAAGTTCCATT	33,036,549	560	59

The PCR reactions were performed on Applied Biosystems® GeneAmp® PCR System 9700 machine (Applied Biosystem, USA). The following cycling conditions were used: 94°C for 2 mins, [94°C for 30 seconds, annealing temperature (T_a) for 30 seconds, 72°C for *x* seconds (dependent on the length of template)] for 25 cycles, and 72°C for 20 mins. The quality and integrity of the PCR product was verified by 1% agarose gel electrophoresis. The PCR products were then purified using the QIAquick gel extraction kit (Qiagen, Germany) according to the manufacture's protocol and the purified PCR templates were sequenced using their respective primers.

Primer sequences AGCAGTCACAAGTCACAGGG and CAGCCCATCGCATGCTCAAT were used to interrogate the missense mutation (chr6:29,913,037, HLA-A exon 7) found in the cell lines carrying A33-B58-DR3 was selected. TA-cloning was then performed to verify the missense mutation using pGEM®-T Easy Vector Systems (Promega, Fitchburg, WI) following manufacturer's instructions. Briefly, 3ul of purified PCR product were ligated with 1ul of pGEM®-T Easy Vector in a reaction mixture comprising of 5ul of 2X ligation buffer and 1ul of T4 DNA ligase and this reaction mixture was incubated for 1 hr at room temperature. Next, transformation was performed using the JM109 High Efficiency Competent Cells where 1ul of the ligated products were added into 50ul of JM109 and the mixture was incubated on ice for 20 mins; followed by heat shock for 45 sec at 42°C and immediate incubation for 2 mins on ice. The recovery of the cells was performed by the addition of 950 ul of S.O.C medium (Invitrogen) and the reaction mixture was incubated in thermo-mixer shaking at 1400rom for 1 hr at 37°C. 100ul from the reaction mixture was plate on LB/ampicillin/IPTG/ X-Gal plate and the plate was incubated overnight at 37°C. Thereafter white colonies were selected, inoculated in 3ml of LB with

100ug/ml of ampicillin and incubated with shaking overnight at 37°C. Plasmid DNA was extracted using QIAprep® Miniprep Kit (Qiagen, Germany) and isolated plasmids were sequenced with the T7 promoter and SP6 promoter primers.

2.14 Phylogenetic Analysis and Estimation of Haplotype Divergence

The SNP sequences of the Asian and European haplotypes spanning across the extended MHC region (chr6: 29.65-33.0Mb) were comprise of 18,781 common SNPs annotated in dbSNP build 137. SNP positions with heterozygous call in the Asian haplotypes were denoted as missing data. Phylogenetic trees were constructed based on the maximum likelihood statistical method and the Kimura 2-parameter substitution model was used to calculate the likelihood on a given tree. To evaluate the reliability of branching points, bootstrap test of phylogeny was performed (n=500). The tree building process was implemented in MEGA5 [153].

Nucleotide diversity metric (π) was used to measure the level genetic variation between the MHC sequences computed by the following equation.

$$\pi = 2\sum_{i=1}^{n}\sum_{j=1}^{i-1} x_i x_j \pi_{ij}$$

n is the number of sequences

x is the frequency of the *i*th and *j*th sequences

π_{ij} is the number of nucleotide differences per nucleotide site between *i*th and *j*th sequences

The age of the Asian haplotypes was determined by the following equation.

$$t = {}^{Ny}\!/_{nLu}$$

t is the length of time since the haplotype sequences shared a common ancestor

N is the number of nucleotide differences between the haplotype sequences

y is the number of years per generation and was set at 20 years

n is the number of haplotype sequences

L is the length of the sequences

u is the mutation rate per nucleotide per year and was set at $1.1X10^{-8}$ [154]

2.15 RNA-seq Preparation and Analysis

Total RNA of at least 15ug isolated from B58AL, B58SC, B46BM, B46ZS, COX and QBL cell lines were subjected to RNA-seq. Prior to the library construction step, cytoplasmic rRNA were removed from the RNA samples using the human Ribo-Zero^{TM} rRNA Removal Kits (Epicentre, USA). For each RNA sample, total RNA libraries of 75bp pair-end reads with DNA fragments size range 120-225bp were prepared using the TruSeq RNA kit (Illumina, USA). A total of 12 libraries were prepared including libraries for the sample biological replicates. Sequencing was carried out on the Illumina HiSeq 2000 machine with two libraries pooled together per sequencing lane, resulting in approximately 79 million reads per library. The library preparation and sequencing were performed by the Duke-NUS Genome Biology Facility.

Quality control (QC) was performed on the raw sequence data to remove low-quality reads. Here, only reads with reads with 70% of the base positions meet the Phred score cuff-off of 20 were retained for the further downstream processes. This filtering was conducted using the NGS QC Toolkit [155] and the sequence biasness was assessed using the FastQC software (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc).

The QC reads from each library were mapped independently using TopHat2 v2.0.7 [156] against the human reference transcriptome (NCBI gene annotation Build 37.2) and the human reference genome (NCBI Build 37.2 reference sequence) using the default settings. The transcripts annotation GFF file and sequence indexes information were downloaded from Illumina's iGenomes project (ftp://ussd-ftp.illumina.com). The TopHat2 alignment algorithm involves a three-step process. First the reads are mapped against the known transcriptome. The remaining unmapped reads are then mapped against the genome; reads that are spanned within a single exon are treated as mapped reads while multi-exon spanning reads are treated as unmapped reads. Together with previously aligned reads with low scores, the third step will split the unmapped reads into non-overlapping 25bp segments and mapped against the genome to identify the most probable splicing sites.

To estimate the relative abundance of the transcripts, the approach implemented in the Cufflinks suite (version 2.1.1) software was used [157]. RNA-seq enables quantitative measure on the abundances of RNA transcripts in the form of the number of reads mapped to the targeted transcripts. However, because of the inherent technical biases introduced in the library preparation step and the variation in the number of reads generated between different sequencing runs as well as due to the difference in the length of the RNA transcripts, normalization procedure is essential to attain meaningful interpretation from the analysis. To account for these issues, Cufflinks uses fragments per kilo base of transcript per million mapped reads (FPKM) that normalized the read counts by the length of the transcripts and the total number of mapped reads in the sample. Cufflinks were performed with the NCBI gene annotation Build 37.2 to output FPKM values for known annotated genes. For differential expression analysis, we used Cuffdiff in the Cufflinks suite to perform the estimation. In Cuffdiff, genes variance across replicates is modeled as a non-linear function of mean counts using a combination of normal and negative binomial distributions and T-test is used to derive the P-values for differential expression. To correct for multiple hypothesis testing, Benjamini-Hochberg adjustment was implemented.

2.16 Databases and Tools

Allele Frequency Net Database (AFND)

The AFND [158] was used to search and extract the HLA allele and HLA haplotype frequencies in worldwide populations. (http://www.allelefrequencies.net/)

The International Immunogenetics Information System® HLA (IMGT/HLA)

The HLA allele sequences used in this study was extracted from the IMGT/HLA [11] (http://www.ebi.ac.uk/ipd/imgt/hla/, Release 3.13.1).

UCSC Genome Browser

The ENCODE histone modification data was assessed through the UCSC Genome Browser (http://genome.ucsc.edu/). In addition, the Table Browser tool in the genome browser was used to retrieve the Human genome reference sequence and annotation data in the MHC region as well as to extract the SNPs information residing in the eight alternate MHC reference sequences.

International HapMap Project

The genotype data of the European population (CEU) and Nigeria Yoruba population (YRI) were obtained from the HapMap database (release 24). The data was QC and all the SNP alleles were all referred to the positive strand before analysis.

Gene Expression Omnibus (GEO)

The expression profiles of the BLCLs from the Centre d'Etude du Polymorphisme Humain (CEPH) collection under the accession number of GSE29158 were extracted from the GEO database.

R Project for Statistical Computing

The R programming environment was used to perform data processing and graphics plotting. R packages were also used for basic statistical analysis.

NGS tools

The Integrative Genomics Viewer (IGV) [159] was used to view and check on the RNA-seq reads mapping results. The SAMtools - version 0.1.18 (http://samtools.sourceforge.net/) and Picard (http://picard.sourceforge.net/index.shtml) were used to manipulate and process the NGS data files.

Chapter 3:

Recombination in the Major Histocompatibility Complex Haplotypes

3.1 Introduction

Recombination has a significant role in the generation of high haplotype diversity found within the human MHC region. The occurrence of these recombination events enables the assortment of DNA blocks within the MHC region [10] and in the process creating diverse combinations of MHC haplotype across populations. Typically, recombination sites are localized to within a 1 – 2kb genomic segment that is flanked by regions with low recombination [160] and are found at multiple points along the chromosomes in a non-random manner. These would lead to the breakdown of linkage disequilibrium (LD) defining the discrete haplotype blocks [161].

As compared to other parts of the genome, the identification of recombination sites within the MHC region is proven to be particularly difficult. Previously, the laboratories of Mary Carrington and Alec Jeffreys employed single-sperm genotyping approach to detect the frequency and distribution of recombination within the MHC region and these studies were able to definitely identify six MHC-residing recombination sites [47, 48]. However, this approach is experimentally laborious and the resolution of the identified recombination sites (average 73.2kb segment per site) is poor. The availability of high-throughput single nucleotide polymorphism (SNP) genotyping assay has provided an alternate method to infer local recombination rates through *in silico* modeling using population genotype SNPs information. This approach is best exemplified by the undertaking of the International HapMap Project where 3.1 million SNPs information extracted from individuals of European, African and Asian ancestries was used to characterize the patterns of LD in the human genome and provide means to estimate recombination rates [160]. However, this approach of using admixed population data to infer recombination sites is not comprehensive enough when applied to the MHC region. This is because the intense genomic rearrangement in the MHC region, not found in other parts of the human genome, is driven by the underlying HLA allelic gene combinations which are uniquely found in distinct population groups complicates the mapping of recombination sites in the human MHC region [162, 163]. Population-distinct HLA allelic gene combinations maintained by natural selection and other evolutionary forces may result in recombination breakpoints specific to each population. Hence, the exclusion of HLA allelic typing information in the recombination sites inference process could possibly conceal the presence of population-specific recombination sites.

In this chapter, we devised a method to detect the location and the frequency of recombination sites within the MHC region. This approach relied on the LD genetic map generated using both the HLA allelic information and the SNPs haplotype phase information from an Asian cohort comprising of Singapore Chinese (CHSG) to infer MHC-residing recombination sites. Using this approach, we also examined the recombination profiles in individuals of European and African ancestries, and hence offered a comprehensive MHC recombination map across populations.

3.2 Results

3.2.1 Linkage Disequilibrium Structure within MHC Region

The inference of recombination sites would require the generation of population LD structure. To perform this, individuals HLA allelic typing and the MHC region SNPs genotype information are essential. The HLA information of 247 CHSG individuals (211 unrelated and 36 comprising members of family trios) were obtained from Yu *et al* [134] and they were typed for four HLA loci at *HLA-A*, *-B*, *-C* and *–DRB1* using sequence-based typing method. DNA of the same cohort was then subjected to Illumina GoldenGate MHC Panel to interrogate the SNPs genotype status covering 4.9Mb (28,970,148 - 33,882,048) of the extended MHC region. By incorporating both the HLA allelic and SNPs genotyped data, SNPs haplotypes phase with their corresponding HLA allelic gene combination were derived using PHASE 2.1 [138]. The resulting 470 chromosomes were categorized according to their HLA alleles and haplotypes. The most frequent HLA haplotypes found in the CHSG population were A*02:07-C*01:02-B*46:01 (12.8%), C*01:02-B*46:01-DRB1*09:01 (9.6%), A*33:03-C*03:02-B*58:01 (9.0%), C*03:02-B*58:01-DRB1*03:01 (7.7%), A*02:03-C*07:02-B*38:02 (4.3%), A*11:01-C*07:02-B*40:01 (4.0%), C*08:01-B*15:02-DRB1*12:02 (3.6%) (Table 3.1). SNPs haplotypes pooled according to their HLA alleles with frequency > 5% were then subjected to extended haplotype homozygosity (EHH) at three separate 1 mega-base genomic segments (chr6: 29.5 – 30.5 Mb, chr6: 31.0 – 32.0 Mb and chr6: 32.2 – 33.2Mb) covering the HLA genes loci. EHH is the measure of LD decay at varying locations from a defined locus [144] and hence provides an overview on the LD structure of the region of interest.

Table 3.1 Common HLA haplotype frequency in CHSG **(A)** 3-locus HLA-A-C-B **(B)** 3-locus HLA-C-B-DRB1 **(C)** 4-locus HLA-A-C-B-DRB1

A

HLA-A-C-B		
Haplotype	**Counts**	**Observed Freq. (%)**
A*02:01-C*07:02-B*40:01	7	1.49
A*02:01-C*15:02-B*40:01	8	1.70
A*02:03-C*07:02-B*38:02	20	4.26
A*02:07-C*01:02-B*46:01	60	12.77
A*11:01-C*03:04-B*13:01	15	3.19
A*11:01-C*08:01-B*15:02	16	3.40
A*11:01-C*07:02-B*40:01	19	4.04
A*11:01-C*01:02-B*46:01	10	2.13
A*11:02-C*12:02-B*27:04	7	1.49
A*24:02-C*03:04-B*40:01	7	1.49
A*24:02-C*07:02-B*40:01	8	1.70
A*33:03-C*03:02-B*58:01	42	8.94

B

HLA-C-B-DRB1		
Haplotype	**Counts**	**Observed Freq. (%)**
C*03:04-B*13:01-DRB1*15:01	12	2.55
C*03:04-B*13:01-DRB1*16:02	7	1.49
C*08:01-B*15:02-DRB1*12:02	17	3.62
C*07:02-B*40:01-DRB1*09:01	15	3.19
C*01:02-B*46:01-DRB1*08:03	13	2.77
C*01:02-B*46:01-DRB1*09:01	45	9.57
C*03:02-B*58:01-DRB1*03:01	36	7.66

C

HLA-A-C-B-DRB1		
Haplotype	**Counts**	**Observed Freq. (%)**
A*02:03-C*07:02-B*38:02-DRB1*16:02	7	1.49
A*02:07-C*01:02-B*46:01-DRB1*08:03	8	1.70
A*02:07-C*01:02-B*46:01-DRB1*09:01	36	7.87
A*11:01-C*03:04-B*13:01-DRB1*15:01	7	1.49
A*11:01-C*08:01-B*15:02-DRB1*12:02	12	2.55
A*11:01-C*07:02-B*40:01-DRB1*09:01	7	1.49
A*33:03-C*03:02-B*58:01-DRB1*03:01	31	7.02

In this study, the core locus for the three respective genomic segments was defined at the *HLA-A*, *HLA-B* and *HLA-DRB1*. Generally, the EHH value of a HLA haplotype recorded a drop, at incremental distance away from the defined core HLA locus, whenever there is decay in LD. Unchanged EHH values over long genomic distance imply regions of high LD and sequence conservation.

Region of high LD with EHH >= 0.9 were observed in extended regions proximate to the HLA-A, *HLA-B* and *HLA-DRB1* genes (Figure 3.1). At the chr6: 29.5 – 30.5 Mb chromosomal segment, a 190kb (position 29,838,709 – 30,027,753) region of extensive LD encompassing *HLA-A, HLA-H and HLA-G* was observed (Figure 3.1A). Haplotypes telomeric of HLA-A were generally well conserved while at the centromeric end, two distinct patterns emerged. Although HLA-A*02:07 and A*33:03 haplotypes stretched a further 200kb or more before breaking up, haplotypes carrying A*02:01, A*24:02, A*11:01 alleles broke right after HLA-A. For the chr6: 31.0 – 32.0 Mb chromosomal segment (Figure 3.1B), high LD was observed for at least 213kb (position 31,325,794 - 31,538,700) in all the common HLA-C-B haplotypes. This region of strong LD extended past *HLA-C* towards the telomeric end, and stretched over the *MICA* gene at the centromeric end. Of note, the centromeric boundary whereby all HLA-C-B haplotypes break corresponded to a HapMap inferred recombination region at position 31.54 Mb. The region of strong homozygosity at the chr6: 32.2 – 33.2Mb segment stretched for at least 181kb and was flanked by 2 recombination sites derived from the sperm typing assay (Figure 3.1C) [47]. The sperm typing recombination site at the telomeric end spans over a 105.15kb region enveloping several EHH drops. The most significant EHH drop was found at the 32,447,054 to 32,448,850 interval in all of the seven major HLA-DRB1 haplotypes.

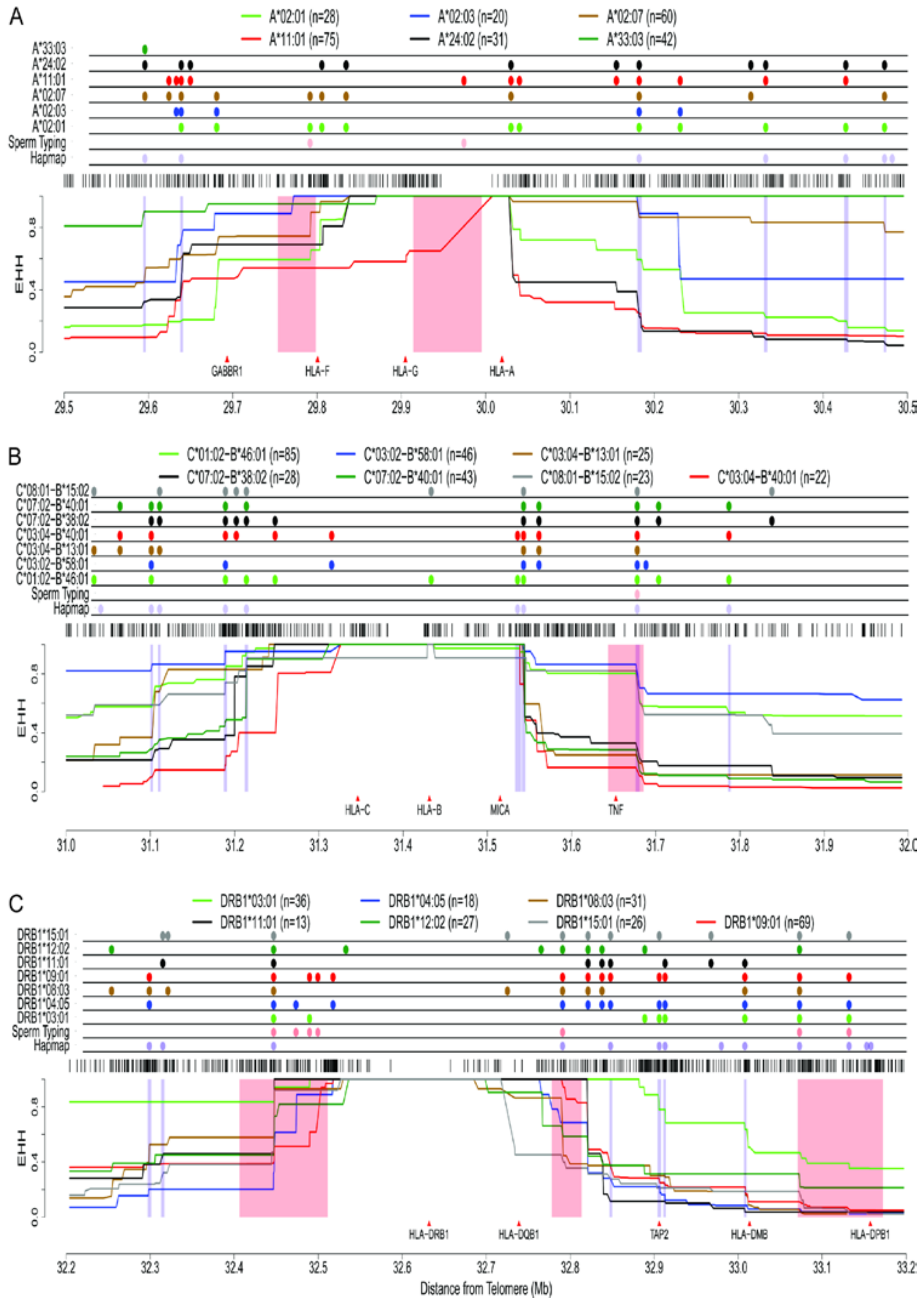


Figure 3.1 Extended haplotype homozygosity (EHH) plots of SNP haplotypes for common HLA alleles. Plots covering 1Mb region with the classical HLA loci/haplotypes used as anchor positions: **(A)** *HLA-A*, **(B)** *HLA-C-B* and **(C)** *HLA-DRB1* respectively. Positions of recombination sites and their relative sizes are mapped onto the plots as follows: six recombination segments identified by sperm recombinants (highlighted in pink columns) and HapMap-inferred hotspots that coincide with EHH drops (highlighted in purple columns).

At the centromeric end, 5/7 HLA-DRB1 haplotypes registered EHH drops at the same location that coincided with the sperm typing derived recombination interval (position 32,447,054 - 32,448,850).

3.2.2 MHC-residing Recombination Sites within the CHSG Population

LD breakages denoted by decay of EHH values along the chromosome is indicative of probable occurrence of recombination events. The EHH plots of the three genomic segments were marked by distinct step-wise drops occurring at non-random discreet interval (Figure 3.1), suggesting the presence of recombination sites. Putative recombination sites are recorded only either when two or more HLA haplotypes independently register drop in EHH value across the same SNPs interval or when at least 10% of the chromosomes carrying a unique HLA allelic haplotype diverge from the core pool. In addition, *in silico* recombination rates estimation using LDhat [145] was performed based on the pooled 470 chromosomes.

A total of 69 recombination sites were characterized across the three genomic segments based on the above criteria (Figure 3.2 and Table 3.2). In contrast, the approach using the sperm typing assay and the HapMap study were only able to detect six and 29 recombination sites across the extended MHC region respectively. All the six sperm typing sites overlapped with EHH drops and as the sperm typing approach had poor resolution ranging in size from 35.4kb to 116.3kb; it was not uncommon to see several EHH drops within each sperm typing derived recombination region.

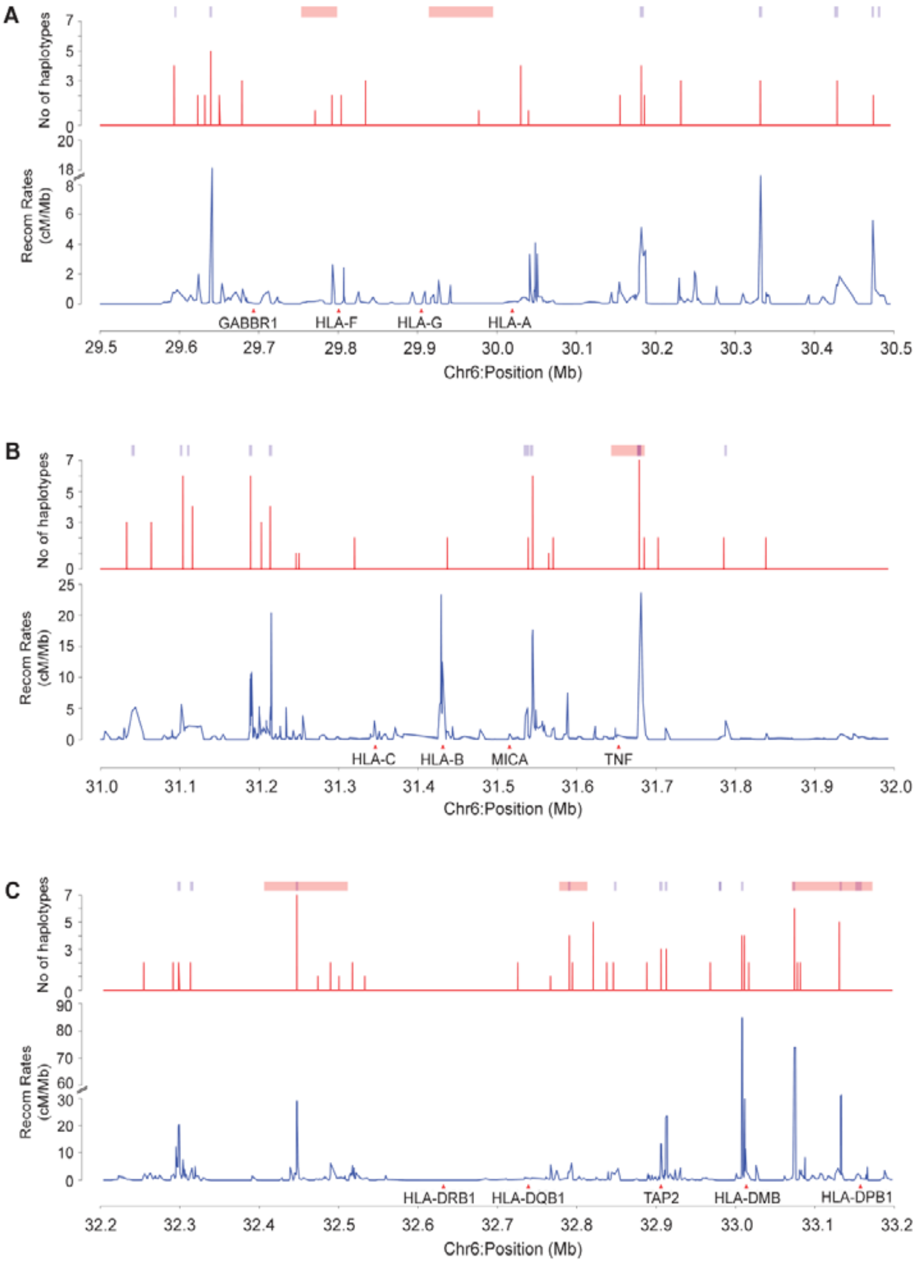


Figure 3.2 Recombination regions identified by SNPs interval displaying EHH decay. Plots illustrate the number of unique HLA allelic haplotypes independently registered drop in EHH (red) and the recombination rates (blue) across **(A)** *HLA-A*, **(B)** *HLA-C-B* and **(C)** *HLA-DRB1*. Recombination segments identified by sperm recombinants are indicated by the pink colored bars and the HapMap-inferred recombination sites are indicated by the blue colored bars.

Table 3.2 List of the recombination sites identified through the EHH approach. 69 recombination sites were identified and among them, 37 were not found in any of the previous studies.

	From	rsID	To	rsID	Size	Marker
*	29591947	rs1592410	29594887	rs9257890	2.94	LOC100507362
	29622209	rs398616	29624221	rs3094576	2.01	UBD 3’
	29630691	rs2534791	29634919	rs362536	4.23	UBD Intronic
*	29637733	rs1233405	29641274	rs388234	3.54	UBD 5’
	29648840	rs362509	29651625	rs362525	2.79	GABBR1 3’
	29677934	rs3025643	29679588	rs10946999	1.65	GABBR1 Exon
†	29769435	rs7772169	29772431	rs3131886	3	MOG 3’
†	29791787	rs1632962	29792613	rs2517911	0.83	HLA-F
	29803481	rs1628578	29804097	rs3817826	0.62	HLA-F-AS1
	29830494	rs9391630	29838709	rs1737069	8.22	IFITM4P
†	29946621	rs2844821	30007656	rs2524005	61.04	HLA-H-HCG4B
	30027753	rs7747114	30032404	rs3893538	4.65	HLA-A
	30038598	rs7739434	30040979	rs3873283	2.38	HCG9 5'
	30154225	rs2394734	30155944	rs7382061	1.72	GABRA3
*	30179089	rs2240070	30184609	rs2240068	5.52	TRIM31 Exon
	30184734	rs2074483	30186994	rs2284163	2.26	TRIM31 Intronic
	30228099	rs2285797	30235302	rs9261535	7.2	TRIM10
*	30329421	rs2516723	30334283	rs4526237	4.86	HLA-L
*	30427909	rs6905389	30429340	rs984802	1.43	RPP21
*	30471924	rs3130113	30476614	rs3130118	4.69	HLA-E
	31032003	rs12212418	31033964	rs11753326	1.96	DPCR1
	31063660	rs1634731	31063908	rs2517416	0.25	MUC21
*	31101937	rs2523897	31105671	rs9262549	3.73	MUC22
*	31110595	rs4248154	31120975	rs3131927	10.38	MUC22 3’
*	31188411	rs2233969	31189722	rs3823402	1.31	C6orf15
	31199971	rs3094204	31205162	rs3130558	5.19	PSORS1C1 Intron
*	31213289	rs1265100	31214247	rs3130573	0.96	PSORS1C1 Exon
	31245144	rs3130503	31247469	rs879882	2.33	POU5F1
	31248720	rs1265158	31251561	rs3131018	2.84	PSORS1C3
	31314185	rs3130685	31325794	rs2894189	11.61	HLA-C
	31435639	rs1811197	31437994	rs2523567	2.36	HLA-B
*	31538700	rs3099840	31538778	rs2596473	0.08	HCP5 5’
*	31543970	rs2523676	31544768	rs2523674	0.8	HCP5 3’
	31551302	rs12660382	31556133	rs2523651	4.83	HCG26
	31557757	rs2523647	31559192	rs2516507	1.44	HCG26 3'
	31561619	rs2516500	31567721	rs2516415	6.1	MICB 5’
	31569068	rs3130922	31571470	rs2516408	2.4	MICB
*†	31676448	rs2857595	31680935	rs2844479	4.49	NCR3
*†	31683255	rs9348876	31686751	rs2844477	3.5	AIF1 5'
	31701455	rs2260000	31703466	rs2736171	2.01	PRRC2A
*	31783744	rs2242653	31786709	rs805287	2.97	LY6G6F
	31837338	rs707938	31838993	rs707937	1.66	MSH5
	32253685	rs2269423	32255674	rs3130349	1.99	RNF5
	32290737	rs206015	32292323	rs404860	1.59	NOTCH4 Intron
*	32297819	rs715299	32299317	rs3830041	1.5	NOTCH4 Intron
*	32311515	rs3130299	32315371	rs416352	3.86	NOTCH4 5’
*†	32447054	rs2050190	32447818	rs6913309	0.76	C6orf10 Intron
†	32473558	rs3129954	32474399	rs4248166	0.84	BTNL2 Intron
†	32489714	rs7759742	32489917	rs743862	0.2	BTNL2 5’
†	32497626	rs3135363	32503546	rs2187818	5.92	HLA-DRA 5’
	32516713	rs3129878	32518115	rs3129883	1.4	HLA-DRA Intron
	32529776	rs10947279	32536263	rs6903608	6.49	HLA-DRA 3’
	32717405	rs9272723	32734064	rs7744001	16.66	HLA-DQA1
	32766693	rs2858330	32767136	rs5002702	0.44	HLA-DQB1
*†	32789623	rs3916766	32791669	rs6935940	2.05	HLA-DQA2 5’
†	32793528	rs3916765	32795336	rs3104401	1.81	HLA-DQA2 3’
	32820225	rs9276431	32821245	rs2239800	1.02	HLA-DQA2 Exon
	32835883	rs1023449	32839688	rs2071550	3.81	HLA-DQB2 Exon
*	32844122	rs9296044	32847866	rs1383265	3.74	HLA-DQB2 5’
	32887974	rs5009557	32888702	rs11244	0.73	HLA-DOB
*	32905515	rs241439	32906773	rs241433	1.26	TAP2 Intron
*	32912195	rs3819714	32913448	rs2071465	1.25	TAP2 Exon
	32965779	rs241414	32970718	rs241407	4.94	LOC100294145
*	33007463	rs3132131	33008629	rs154972	1.17	HLA-DMB 3'
	33010561	rs10751	33011878	rs151719	1.32	HLA-DMB
	33013724	rs194675	33019792	rs2395296	6.07	HLA-DMB 5'
*†	33072674	rs206765	33075719	rs12216336	3.05	BRD2 3'
	33077435	rs172274	33078428	rs206762	0.99	HLA-DOA 3'
	33080668	rs592625	33082379	rs2581	1.71	HLA-DOA
*†	33129170	rs7743563	33132251	rs435549	3.08	HLA-DPA1 3’

† indicates intervals that reside within sperm typing segments and ***** indicates intervals that overlapped with the HapMap recombination sites. Each recombination site was mapped to the Human Reference Sequence Assembly 36.1 (NCBI 36.1) and assigned to a marker that is in the closest proximity to the site.

For example, the sperm typing derived recombination segment chr6: 32, 063,170 – 32,511,466 harbored five independent recombination sites. Of the 29 HapMap inferred recombination sites, 24 of them corresponded to the EHH drops and 37 of the EHH-derived recombination sites were not found in any of the previous studies. In addition, the examination of unambiguous HLA allelic SNPs haplotypes derived from trios data showed that the positions of LD breakage along these chromosomes coincide with the EHH-derived recombination sites (Figure 3.3). This high correlation of the EHH-derived recombination sites with the sperm typing, HapMap data and the positions LD breakage along haplotypes derived from the trios data illustrates the validity of the EHH mapping approach.

Peaks of recombination rates were located at SNPs intervals where multiple haplotypes independently displayed EHH drop. For instance, a recombination rate peak of 4cM/Mb was located at chr6: 30,027,753 – 30,032,404 and A*02:01, A*02:07, A*11:01 and A*24:02 haplotypes all exhibited EHH drop at this genomic interval. EHH-derived recombination sites were also found in regions with low recombination rates. These recombination sites were specific to a unique HLA haplotype and as such the number of chromosomes was not significant enough to result in elevated recombination rate which was estimated from population-pooled chromosomes. This highlights the importance of accounting for the HLA genes allelic information to enhance sensitivity and specificity in the inference of recombination sites across the MHC region. Furthermore, the HLA gene allelic EHH approach was able to produce excellent recombination map resolution, with 28/69 (41%) mapped to <2kb in size and another 29/69 (42%) falling between 2 to 5kb.

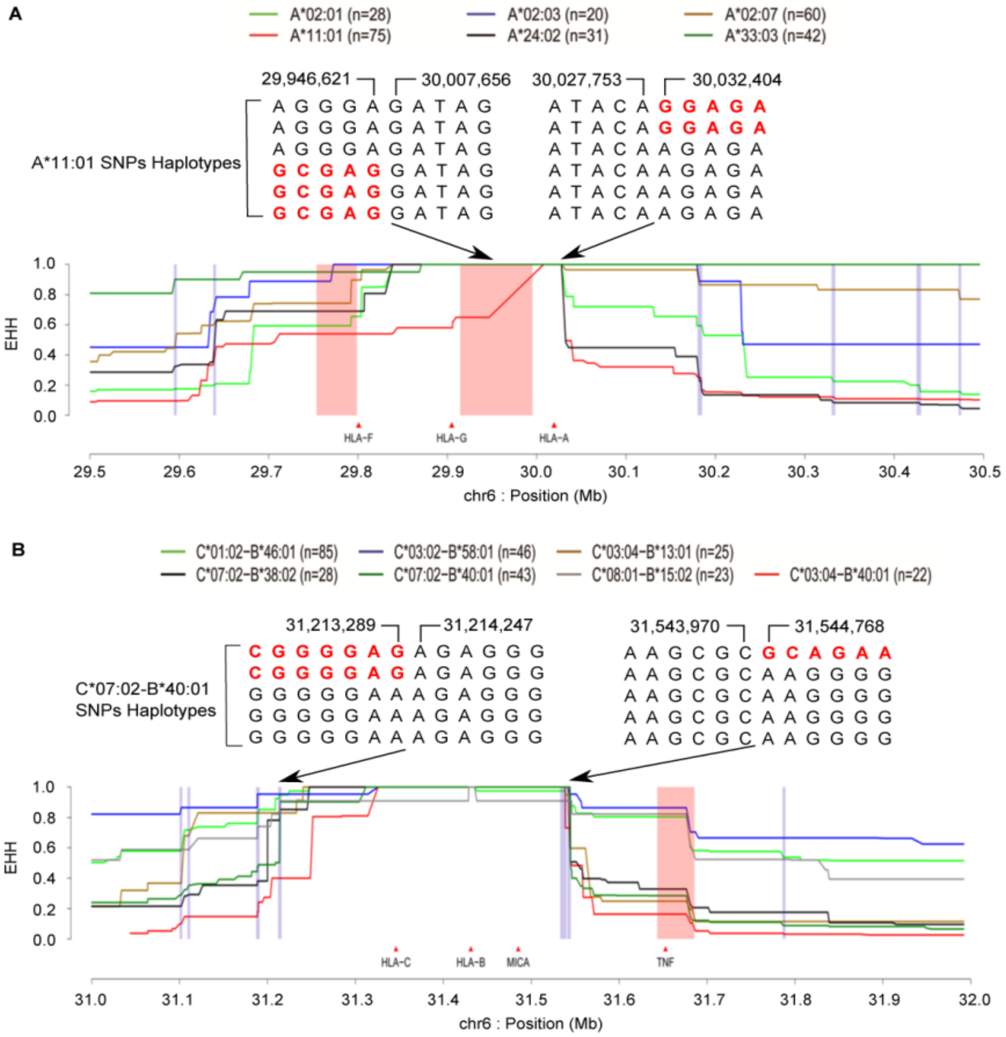


Figure 3.3 Recombination sites derived from trios data. **(A)** Recombination sites supported by trios derived A*11:01 SNPs haplotypes. **(B)** Recombination sites supported by trios derived C*07:02-B*40:01 SNPs haplotypes. The arrows point to the location of EHH decay where the LD of the SNPs haplotypes is disrupted.

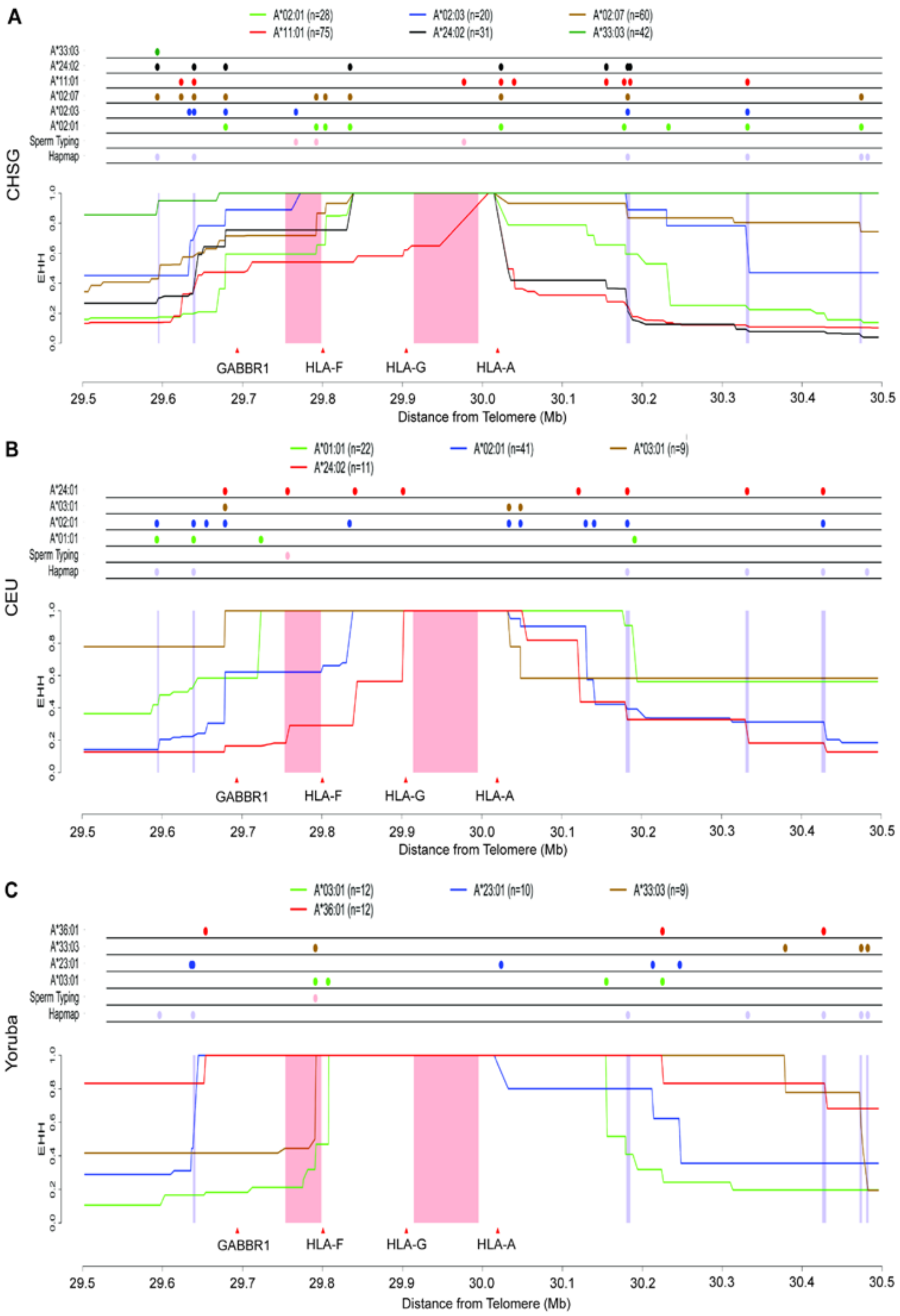


Figure 3.4 EHH plots of SNP haplotypes for common HLA-A haplotypes. **(A)** CHSG, **(B)** CEU and **(C)** YRI. Dots in the panels above each plot indicate the SNPs interval where haplotypes break, with each color denoting a specific HLA-A haplotype. Positions of recombination sites and their relative sizes are mapped onto the plots as follows: recombination segments identified by sperm recombinants (highlighted in pink columns) and HapMap-inferred sites that coincide with EHH drops (highlighted in purple columns).

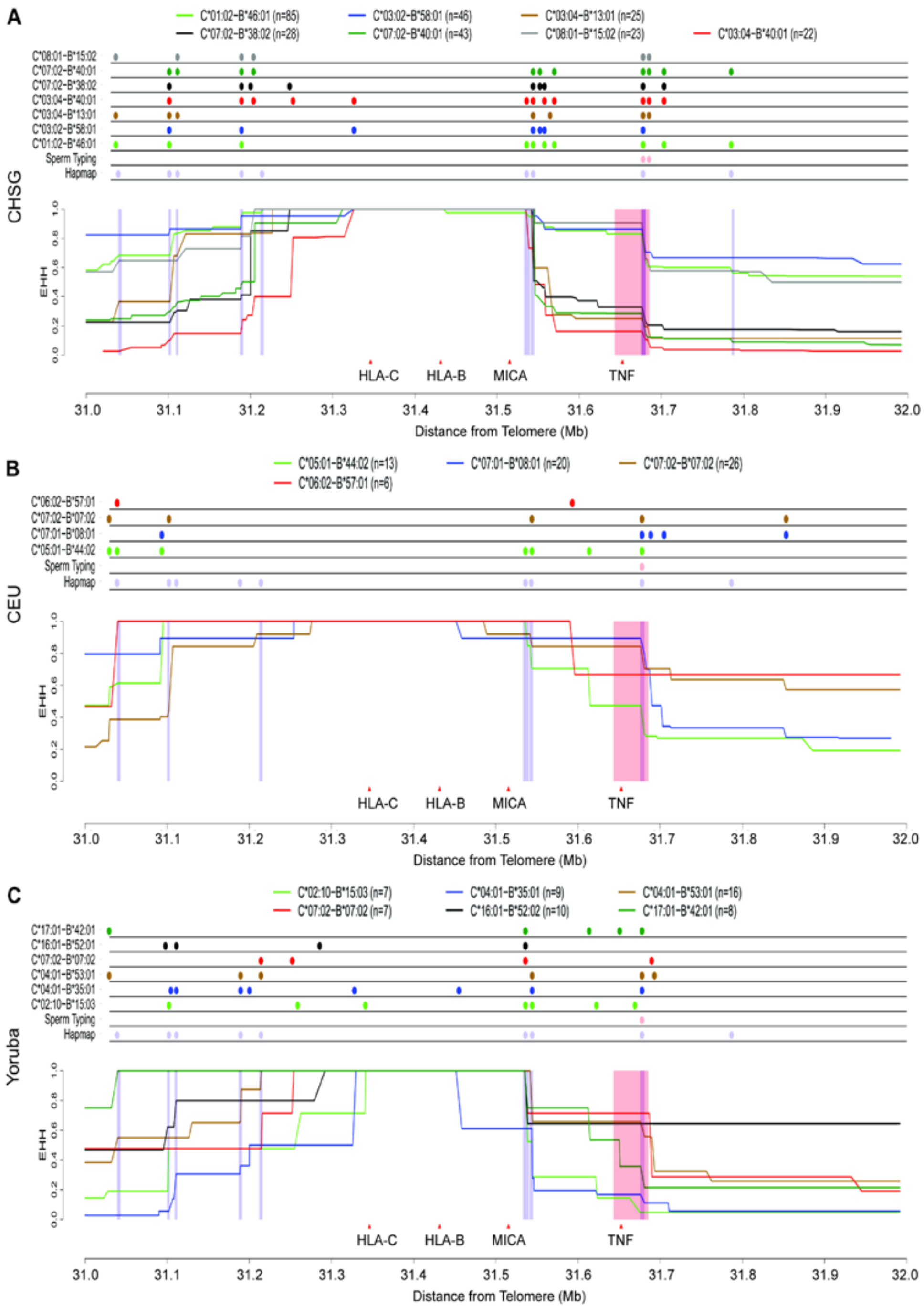


Figure 3.5 EHH plots of SNP haplotypes for common HLA-C-B haplotypes. **(A)** CHSG, **(B)** CEU and **(C)** YRI. Dots in the panels above each plot indicate the SNPs interval where haplotypes break, with each color denoting a specific HLA-C-B haplotype. Positions of recombination sites and their relative sizes are mapped onto the plots as follows: recombination segments identified by sperm recombinants (highlighted in pink columns) and HapMap-inferred sites that coincide with EHH drops (highlighted in purple columns).

3.2.3 Population-specific Recombination Sites within MHC Region

To date, the knowledge of MHC recombination map variation across populations is limited and not well studied. The EHH approach integrated with the HLA gene allelic information offers an opportunity to perform comparative study of the recombination profiles between different population groups. Here, in addition to the CHSG population, we applied this approach to two HapMap populations, the Europeans (CEU) and the Nigeria Yorubans (YRI). The SNP genotype data of these populations were extracted from the HapMap depository while the *HLA-A*, *-B* and *-C* allelic typings were obtained from Erlich *et al* [143]. To enable comparison across the three populations, 1360 SNPs loci whose genotypes were known in the three populations were selected and subjected to the EHH analysis. As *HLA-DRB1* gene allelic information is not available for the HapMap populations, EHH analysis was only performed at the two genomic segments covering the HLA class I genes. Similar to the observation in the CHSG population, maintenance of LD proximate to the HLA genes were observed in the CEU as well as the YRI population and the range of high LD region at the HLA genes varied according to the underlying HLA allelic background (Figure 3.4 and 3.5).

Collectively, there were 37 probable recombination sites detected in the CHSG, 30 in the CEU and 38 in the YRI (Figure 3.6). We were also able to recover >90% of the CHSG recombination sites identified in the previous section, albeit at larger segment interval. Interestingly, >50% of the identified sites in each population (CHSG – 56.8%, CEU – 50.0% and YRI – 63.2%) were uniquely population-specific. Only <16% of the sites were shared among the three populations and all of these sites fell within the segments determined by the sperm typing experiment or the *in silico* modeling approach (Table 3.3).

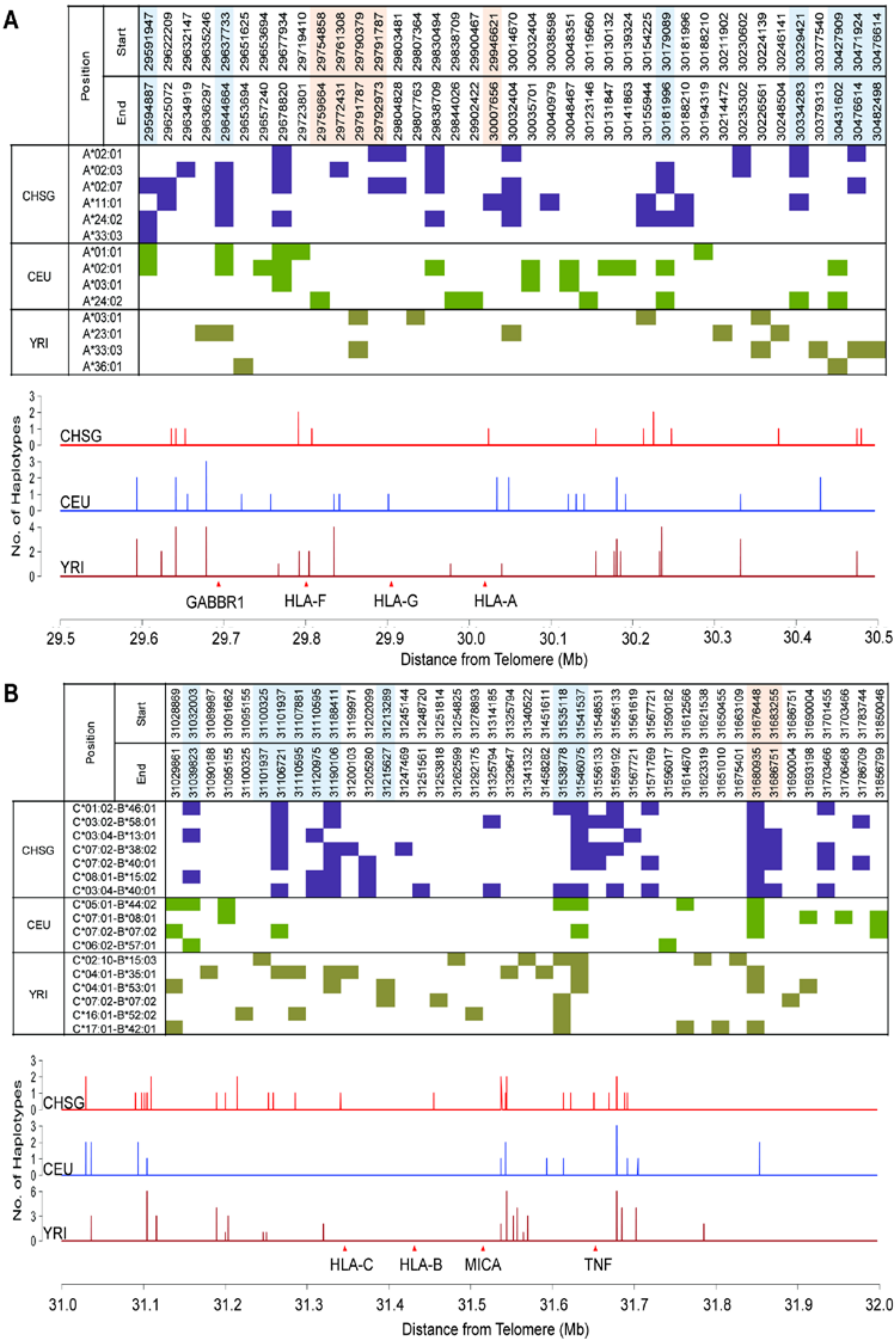


Figure 3.6 Recombination sites across the CHSG, CEU and YRI population. **(A)** *HLA-A* **(B)** *HLA-C-B*. The upper section of each panel displays the HLA haplotypes recombination sites interval across the 3 populations. The lower section of each panel displays plots that indicate the number of unique HLA allelic haplotypes independently registered EHH drop in the CHSG, CEU and YRI.

Table 3.3 Number of common and population-specific recombination sites identified in the CHSG, CEU and YRI across *HLA-A* and *–B* region. Only a small proportion of recombination sites are shared among populations.

Region	CHSG-specific	CEU-specific	YRI-specific	CHSG & CEU	CHSG & YRI	CEU& YRI	All
HLA-A	9	10	9	5	3	1	1
HLA-B	12	5	15	1	2	3	4

In addition, recombination sites distinct to a specific population were noted to be in close proximity to other recombination sites unique to another population resulting in a boarder genomic segment where recombination activities are likely to occur across populations. This was especially evident in regions downstream of *HLA-B*. For instance, population-specific recombination sites detected in the three populations were within a 31kb segment (chr6: 31,089,987 – 31,120,975) downstream of *MUC22* gene. This observation is in agreement with the previous studies where elevated recombination activities are likely to occur within a cluster flanked by regions of low recombination [48].

The differences in recombination sites among populations reflect the variation of haplotype pool in each population. Of note, for example, A*02:01 haplotype can be found in both CHSG and CEU populations; however there is a difference in the distribution of recombination sites in A*02:01 haplotype carry by these 2 populations. This suggests that the differences observed in the distribution of recombination sites might not entirely due to the difference in haplotype data from each population. Hence, the consideration of HLA haplotypic variation within and across populations is important to improve the resolution of the MHC recombination map.

3.2.4 Absence of Recombination in Common Asian HLA Haplotypes

Conserved extended haplotype (CEH) is defined as a genomic segment with distinctive long-range sequence conservation coupled with suppression of recombination events [37, 38]. The presence of CEHs have been reported among a number of common HLA haplotypes in European (A1-C7-B8-DR3) [164] and Japanese populations (A24-C12-B52-DR15 and A33-C14-B44-DR13) [165]. In our studied population, a number of four locus HLA haplotypes are found at relatively high frequency (>1%); notably the A*33:03-C*03:02-B*58:01-DRB1*03:01 (A33-B58-DR3) and A*02:07-C*01:02-B*46:01-DRB1*09:01 (A2-B46-DR9) haplotypes (Table 3.1). Next, we aim to investigate whether the common HLA haplotypes in the CHSG population would display the characteristic of CEHs. Using the phase chromosomes comprising of 1877 SNP markers, the major allele frequency (MAF) of the SNP markers was computed for each of the common CHSG HLA haplotypes. Contiguous SNP loci with MAF value of 1 indicates extensive conservation and therefore implies the presence of CEH.

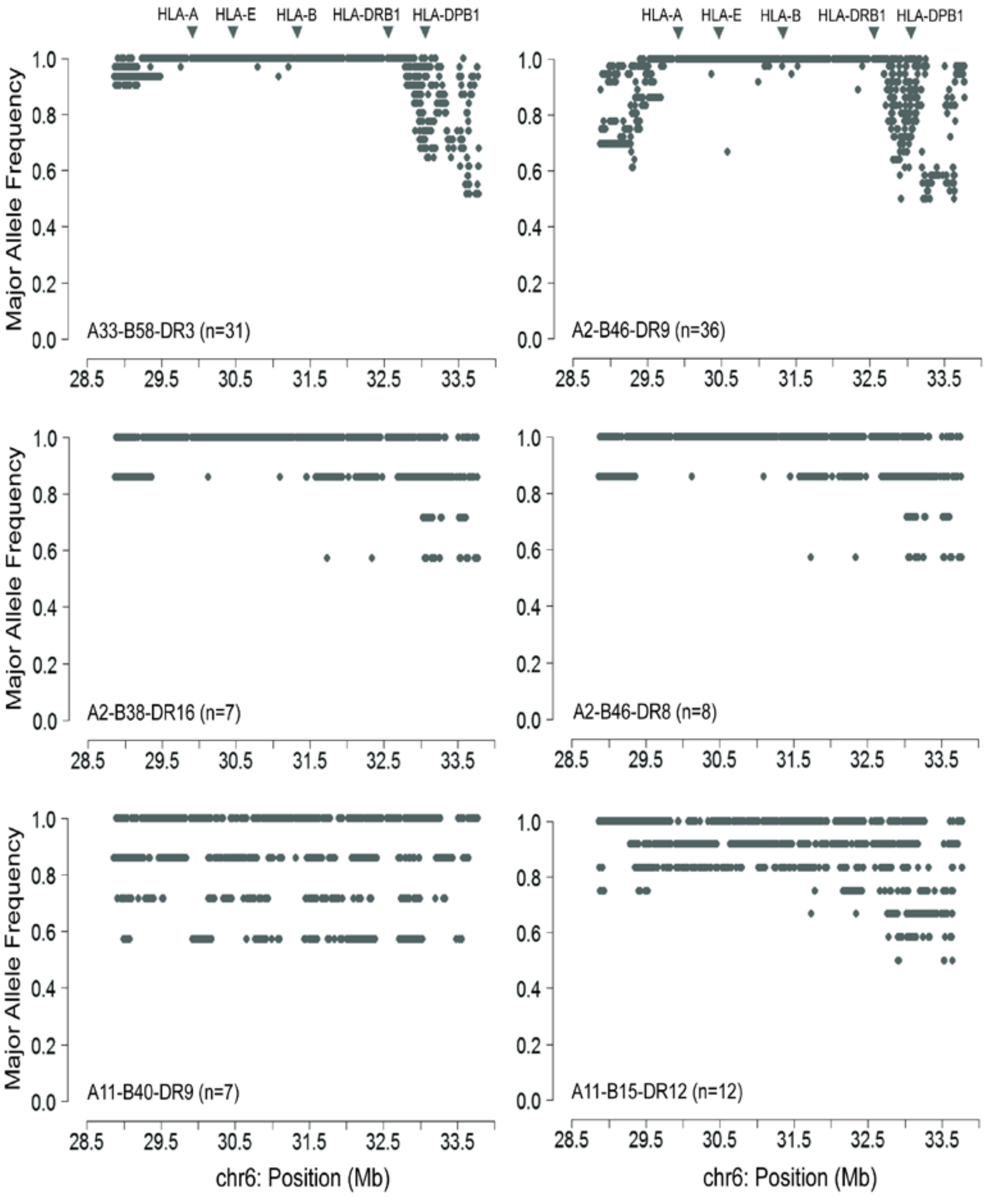


Figure 3.7 Major allele frequencies of SNP markers across the MHC region for common four-locus CHSG HLA haplotype. The plots were derived from 1877 SNP loci. Of the common Singapore Chinese HLA haplotypes, A33-B58-DR3 and A2-B46-DR9 displayed extensive SNPs sequence conservation across the MHC region.

The MAF analysis revealed an extended region of SNPs invariant between *HLA-F* and *HLA-DQB1* covering at least 3Mb for both A33-B58-DR3 and A2-B46-DR9 HLA haplotypes; suggesting the absence of recombination events across the extended MHC region (Figure 3.7). In contrast, for the other four common HLA haplotypes, SNPs invariant was restricted only at regions adjacent to HLA loci and the intervening regions were characterized by the decay of LD. This data suggests that high frequency HLA haplotypes observed in a population does not necessarily imply the extended conservation of high LD and recombination suppression at the genomic level across the MHC region.

To generate further evidence for presence of CEH in A33-B58-DR3 and A2-B46-DR9, DNA purified from six lymphoblastoid cell lines (B58AL, B58SC and B58CF HLA homozygous for A33-B58-DR3; B46BM, B46ZS and B46CM HLA homozygous for A2-B46-DR9) was analyzed by Illumina Human 1M-Duo BeadChip SNP array. After SNPs quality filtering, 10215 SNP markers were found in the genomic segment of chr6:25.0 – 35.0Mb; of which 7509 SNPs fell within the extended MHC region (28.5Mb to 33.5Mb). These SNP markers were then subjected to Runs Of Homozygosity (ROH) analysis, implemented to screen for the SNPs' genotype homozygosity profile in each cell line. Regions of conservation were identified for each cell line based on levels of homozygosity within the genome. Homozygosity and intra-haplotype conservation were found not only at the HLA loci, but also across the extended MHC region (Figure 3.8A). For A33-B58-DR3, at least 99.5% of SNPs were found to be homozygous across 4.66 Mb region (Table 3.3) and the genotype calls of these homozygous SNPs were consistent in all the three A33-B58-DR3 cell lines.

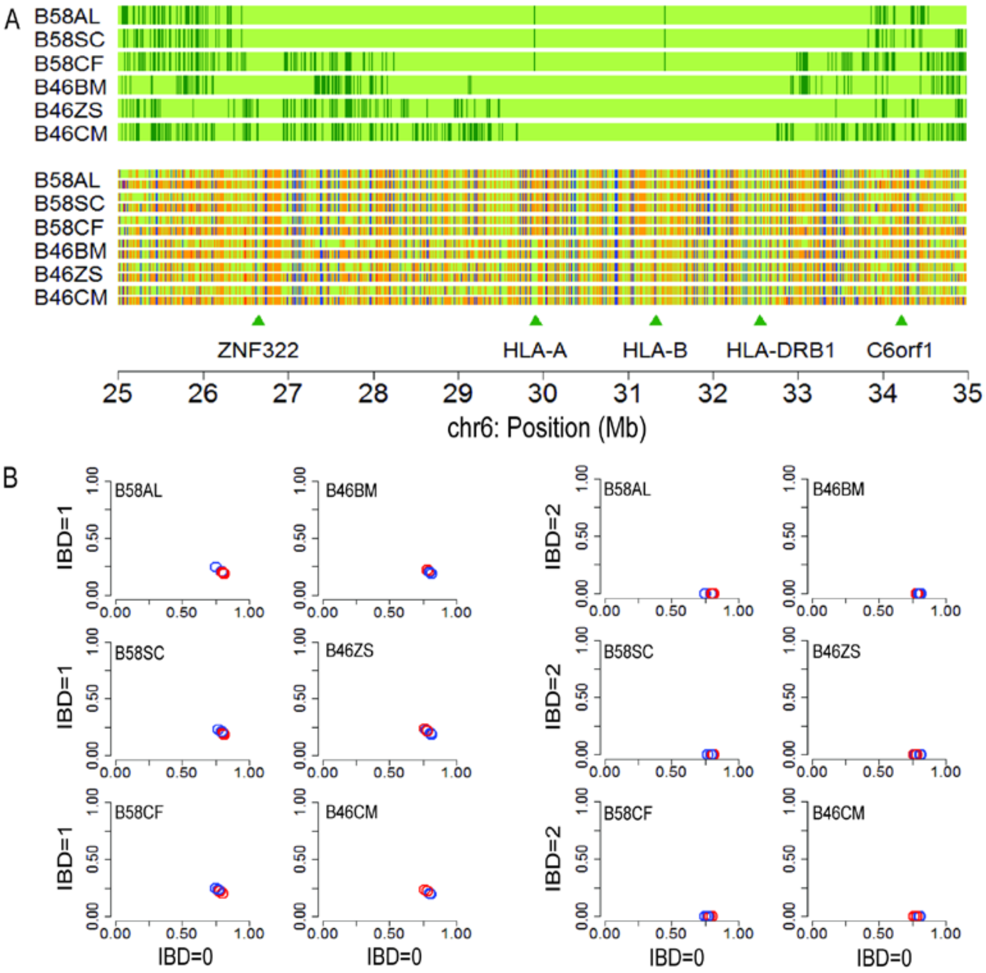


Figure 3.8 Conserved extended haplotype in A33-B58-DR3 and A2-B46-DR9 across MHC region. **(A)** SNPs alignment of six HLA homozygous individuals carrying specific haplotype. This homozygous individuals SNPs alignment includes the genotype status of each SNP (green vertical bars indicate a homozygous SNP call and dark green vertical bars indicate a heterozygous SNP call) and the SNP allelic call with reference to its position (green = adenine, red = cytosine, orange = guanine, and blue = thymine). 10125 SNP markers are involved in the alignment **(B)** Identity-by-descent analysis. Pairwise IBD plots (IBD=1 vs IBD=0 and IBD=2 vs IBD=0) of a reference individual with the other respective individuals. Blue circle indicates pairwise analysis of individuals carrying A33-B58-DR3 while the red circle indicates pairwise analysis of individuals carrying A2-B46-DR9.

Table 3.4 Range of conserved extended region for the six cell lines derived from ROH analysis of the SNPs information. Htz: heterozygous; Hmz: homozygous.

Cell line	Start	End	Length (Mb)	No. of SNPs	No. of htz SNPs	% of hmz SNPs
B58AL	26,922,906	33,853,071	6.93	8660	26	99.7
B58SC	26,922,906	33,820,059	6.90	8637	26	99.7
B58CF	28,310,997	32,973,794	4.66	6997	35	99.5
B46BM	29,350,854	32,903,900	3.57	6238	25	99.6
B46ZS	29,577,617	33,910,884	4.33	6975	21	99.7
B46CM	29,728,209	32,739,888	3.01	5283	16	99.7

Likewise for A2-B46-DR9, the conservation region is slightly shorter at 3.01Mb with 99.6% of the encompassing SNPs genotyped as homozygous (Table 3.4) and again the genotype calls of the homozygous SNPs were consistent in all the three A2-B46-DR9 cell lines. To verify that the extended segment of conservation observed within the MHC region is not merely due to undetected familial relatedness among these individuals, an identity-by-descent (IBD) analysis was performed using the whole-genome SNPs information of the six individuals. All the pairwise IBD analysis between every possible individual clustered at the bottom right quadrant (Figure 3.8B) indicating no strong evidence for relatedness among the individuals. Therefore, the SNP sequence conservation observed within the MHC region among individuals carrying similar HLA haplotypes is not due to familial relationship. These analyses reveal that for both the A33-B58-DR3 and A2-B46-DR9 HLA haplotypes, the linkage of the HLA alleles is not restricted only to the HLA loci, but rather that mega-bases of genomic segment are inherited together in linkage disequilibrium with minimal recombination.

3.3 Conclusion

In this chapter, a single population cohort Singapore Chinese and their corresponding HLA type were used as a mean to partition the haplotypes. This is a great improvement over the sperm typing method which also recognizes the advantage of a haploid genome as a “cleaner” read-out. In addition, this method also counts females (which sperm typing lacks), thus providing a more balanced assessment of the study population. The single haplotype information was able to reveal unambiguously, positions along the MHC genome where recombination events had occurred, leading to breakage of SNP linkage disequilibrium and the results can thus be visualized as EHH plots. From the 470 chromosomes studied, we were able to identify 69 recombination sites of which 37 recombination sites were novel. By applying the above approach to 2 other populations, European and African, we were able to show that each population has its own unique signature of recombination sites within the MHC. This has not been empirically defined till now and even more interesting is that the population-specific recombination sites are seldom shared (or seen) among the 3 populations studied; highlighting the role of recombination in generating haplotype diversity. Through this study, for the first time, we revealed two HLA haplotypes in the Singaporean Chinese population (A2-B46-DR9 and A33-B58-DR3) with little or no recombination activity for at least 3Mb across the MHC region.

Chapter 4:

Intra-haplotypic Variation in MHC Conserved Extended Haplotype

4.1 Introduction

Typically, the linkage disequilibrium (LD) breakdown becomes evident over certain regions along the MHC with the extent and arrangements markedly defined by the HLA alleles. However, in the previous chapter, we revealed extensive LD lacking any recombination events covering a region of at least 4.66 Mb in individuals carrying A33-B58-DR3 and 3.01Mb in A2-B46-DR9. CEHs draw great interests not just because of their unique genomic traits but also because common conserved extended haplotypes (CEHs) are known to be associated with numerous diseases [35, 166]. For instance, the A1-B8-DR3-DQ2 haplotype alone is a risk factor for type 1 diabetes, systemic lupus erythematosus, rheumatoid arthritis and IgA deficiency and various other diseases [40-42]. Of note, studies have consistently demonstrated the association of A33-C3-B58-DR3 and A2-C1-B46-DR9 common HLA haplotypes with nasopharyngeal carcinoma [167, 168], myasthenia gravis [169] and type 1 diabetes [170]. Despite the strength of the risk associations, genetic dissections of the exact disease-causing variants and genes have been difficult. Firstly, due to the long and extensive LD on these CEHs, it is often difficult to distinguish between disease-causing variants and other benign variants in linkage within the same haplotype, causing difficulty in identifying the genes or variations responsible for causing disease. Secondly, the extent of intra-haplotypic variation within the conserved region of the CEHs, which might separate disease-affected haplotype carriers from unaffected haplotype carriers, is not well-established. Previous attempts to decipher the difference between type-1A diabetic A1-B8-DR3-DQ2 haplotypes and non-type-1A diabetic A1-B8-DR3-DQ2 yielded uninformative results, despite the high resolution of the common SNPs (MAF>=5%) genotyping platform [164]. Such attempts may be futile because they did not include intra-haplotypic variants that might be rare at the population level. Indeed, studies have shown that functional variants whose frequency were too low to be detected by genome-wide association exhibit significant disease susceptibility effects [171, 172]. Thus, further efforts are needed to characterize both the extent of LD and the extent of intra-haplotypic variation within CEHs.

Attempts to quantify the level of polymorphism within MHC haplotypes haven been carried out on eight European MHC haplotypes using bacterial artificial chromosome cloning (BACs) and shotgun sequencing [14, 15]. Among these cells, only the MHC haplotype of PGF (A3-B7-DR15-DQ6), and COX (A1-B8-DR3-DQ2), were sequenced completely and the MHC reference sequence of PGF was incorporated into the mosaic NCBI Build 37.2 reference sequence [16]. Each of these haplotypes was assembled into a haploid sequence from a single consanguineous cell line using BAC derived sequences; however, as the parental origin of each BAC sequence was uncertain and with only one representative from a particular HLA haplotype, there is no information on intra-haplotypic variation, hence the characteristic features of these haplotypes cannot be determined. Smith and colleagues went a step further, using PCR primer pairs, covering the MHC region, to perform partial re-sequencing for 19 independent A1-B8-DR3-DQ2 chromosomes and in the process identified only 11 single-nucleotide variants (SNVs) between HLA-A and HLA-DQ gene [21]. Unfortunately, in this study only 15% of the conserved region was sequenced, and it was thus unable to definitively explore the scope of variation in the CEHs. Subsequent studies employed target region and next generation re-sequencing approaches to interrogate variations residing within the MHC region [19, 20] but cell lines were either HLA heterozygous or did not exhibit CEH characteristics in the

MHC genomic region and as a result, haplotype sequence assignment was problematic. The sequences of the eight common European MHC haplotypes and the variations characterized from the analysis of these sequences were often used as a framework and resource for MHC disease susceptibility studies. Though useful, given the immense diversity of the MHC region, these are not sufficient to provide a complete description of the region in particularly for individuals from other ethnic backgrounds.

In this study, we report the characterization of two Asian CEHs, A*33:03-C*03:02-B*58:01-DRB1*03:01 (A33-B58-DR3) and A*02:07-C*01:02-B*46:01-DRB1*09:01 (A2-B46-DR9) which are present in relatively high frequencies of about 7% and 6% respectively in the Singapore Chinese population who are predominantly descended from Southern China [173, 174] . In contrast to earlier studies which examined only one representative of a particular MHC haplotype, we compared 3 unrelated individuals for each CEH and subjected them to whole genome sequencing. The data will provide an in-depth, nucleotide-resolution view of these prominent Asian CEHs, and assess intra-haplotypic conservation and variation in the extended MHC region.

4.2 Results

4.2.1 Fine-scale Mapping of A33-B58-DR3 and A2-B46-DR9 CEHs using Deep-sequencing

The use of high density SNP typing at 1 SNP per 665 bp as described in the previous chapter demonstrated that conserved extended haplotypes of HLA identical independent individuals appeared to be indistinguishable. It is as yet not known if this conservation is maintained at the nucleotide level. To do so, six HLA homozygous cell lines B58AL, B58SC, B58CF, B46BM, B46ZS and B46CM were subjected to whole-genome sequencing (WGS) using the Complete Genomics (CG) platform. Raw reads were processed by the Complete Genomics Standard Sequencing Pipeline 2.0, and assembled according to the Genome Reference Consortium Human genome build 37 (GRCh37) [150]. For each sequenced genome, the mean coverage per base pairs was at least 37.13 times covering no less than 94.18% of the extended MHC region (28.5Mb to 33.5Mb) (Table 4.1). Nearly half of the uncovered base pairs were within the 32,435,000 – 32,660,000 segment encompassing the HLA-DRB, HLA-DQA1, HLA-DQB1, pseudogenes and 43.66% - 60.42% of this 225kb region was not covered (Figure 4.1). This low rate of sequence calling is likely due to the highly polymorphic nature of the HLA-DRB region where multiple insertions and deletions of large genomic sequence result in haplotype-specific rearrangements of the HLA-DRB genes and its pesudogenes, making the alignment of this segment particularly difficult. Hence, the HLA-DRB region, together with another region between 31,210,000-31,235,000 where low coverage was observed due the repetitive sequence, were excluded from comparative analysis.

Table 4.1 CG sequencing coverage and performance. The mean coverage is at least 37.13 per cell line, resulting in high confidence calls for at least 94% of the MHC region for each of the cell line.

Cell lines	Mean Coverage per bp	% bp Coverage >5X	% bp Coverage >20X	% bp Coverage >40X	% of MHC region covered
B58AL	44.67	98.14	91.09	61.40	95.31
B58SC	43.69	98.10	90.57	59.44	95.15
B58CF	39.06	97.96	88.14	49.10	94.66
B46BM	43.54	98.13	90.31	58.93	95.03
B46ZS	44.30	98.11	90.88	60.94	95.06
B46CM	37.13	97.86	86.64	44.20	94.18

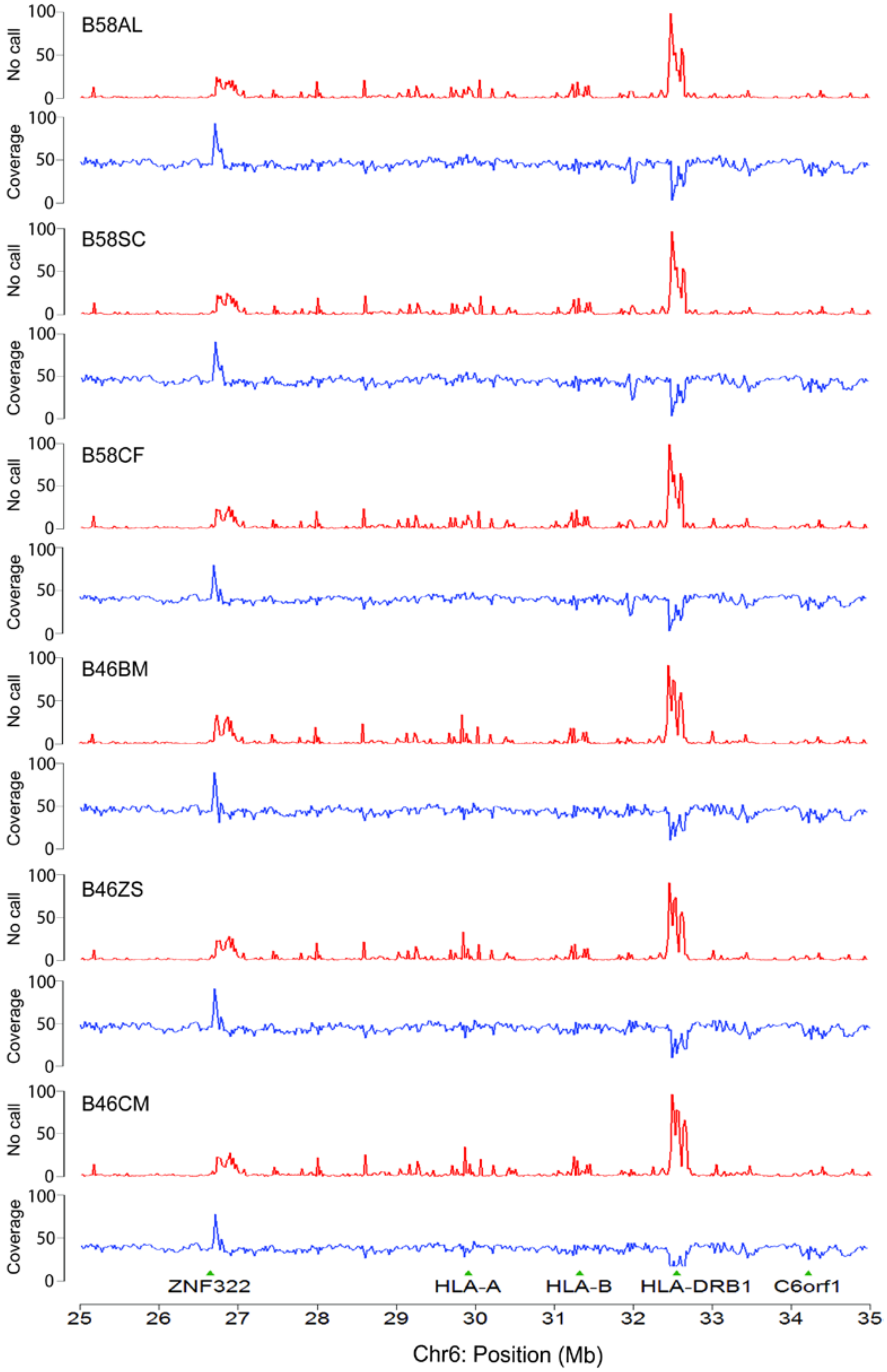


Figure 4.1 No-call and coverage profile for each cell line across the MHC region. The average GC corrected coverage and the no-call counts were binned into non-overlapping 20kb windows. Low coverage and high no-call rate were predominantly located at the HLA-DR region.

To assess the quality of data generated from the CG platform, genotypes calls between 25Mb to 35Mb from the Illumina Human 1M-Duo BeadChip SNP array were compared with the deep sequencing data. The results showed a high concordance rate of 99.5% for A33-B58-DR3 and 99.6% for A2-B46-DR9 between SNPs genotyping and the CG data (Table 4.2). 48 call sites among the discordant data were randomly selected for validation by PCR re-sequencing. Of these, 45/48 were found to be consistent with the CG data, while only 3/48 agreed with the SNP genotyping array data (Table 4.3), indicating that the CG sequencing platform in general delivers higher call accuracy than the SNP genotyping platform.

Next, the range of genomic conservation in each sample was assessed using homozygosity levels in the deep sequencing data. Nucleotides within the 25Mb -35Mb region of chromosome 6 were binned into windows of 5kb, and the numbers of homozygous and heterozygous reference single nucleotide variants (SNV) calls within each bin were examined. Stretches of homozygosity were defined to be regions with no more than four consecutive windows having the zygosity SNVs ratio (number of homozygous SNVs against total number of SNV in a given window) of less than 0.95. The resulting conservation region in each genome coincided with the region determined using the SNPs genotyping platform (Table 4.4) and regions outside the conserved segments have comparably much higher number of heterozygous reference SNVs (Figure 4.2). Within the conserved segment boundaries,

Table 4.2 SNP genotype call differences between Illumina Human 1M-Duo BeadChip and CG sequencing. Hmz indicates homozygous genotype call and htz indicates heterozygous genotype call.

Cell line	Positions Compared	Differences hmz/hmz	Differences hmz/htz	Differences htz/hmz	Differences Total	% Match
B58AL	9916	15	1	23	39	99.6
B58SC	9876	14	4	26	44	99.6
B58CF	9880	12	8	27	47	99.5
B46BM	9908	11	8	14	33	99.7
B46ZS	9903	13	3	14	30	99.7
B46CM	9904	12	10	13	35	99.6

Table 4.3 Re-sequencing experiments to assess mismatches between SNP genotyping and CG sequencing platform in samples carrying A33-B58-DR3 and A2-B46-DR9 haplotype. 45/48 of the positions assessed were in agreement with the CG data.

A33-B58-DR3											
			B58AL			B58SC			B58CF		
Position	rsID	Ref	SNP Array	CG	PCR	SNP Array	CG	PCR	SNP Array	CG	PCR
29,796,376	rs12722477	C	C/C	A/A	?/?	C/C	A/A	A/A	C/C	A/A	A/A
29,819,909	rs2508053	C	C/T	C/C	C/C	C/T	C/C	C/C	C/T	C/C	C/C
30,383,046	GA005234	C	C/G	C/C	C/C	C/G	C/C	C/C	C/G	C/C	C/C
30,418,354	rs34111681	G	T/T	G/G	G/G	T/T	G/G	G/G	T/T	G/G	G/G
31,170,514	rs9263870	A	A/G	G/G	G/G	A/G	G/G	G/G	A/G	G/G	G/G
31,321,327	rs9266095	A	A/G	A/A	A/A	A/G	A/A	A/A	T/T	G/G	?/?
31,639,979	rs9267532	C	C/T	T/T	T/T	C/T	T/T	T/T	C/T	T/T	T/T
31,655,438	rs10573	G	A/G	A/A	A/A	A/G	A/A	A/A	A/G	A/A	A/A
31,697,558	rs707916	G	A/G	A/A	?/?	A/G	A/A	?/?	A/G	A/A	A/A
32,130,937	rs10680	T	C/C	T/T	T/T	C/C	T/T	T/T	C/C	T/T	?/?
33,036,549	rs17509489	T	G/G	T/T	?/?	G/G	T/T	T/T	G/G	T/T	?/?

A2-B46-DR9											
			B46BM			B46ZS			B46CM		
Position	rsID	Ref	SNP Array	CG	PCR	SNP Array	CG	PCR	SNP Array	CG	PCR
29,801,958	rs2743944	T	T/C	T/T	T/C	T/C	T/T	T/C	T/C	T/T	T/C
29,913,509	rs1062405	T	C/C	T/T	T/T	C/C	T/T	T/T	C/C	T/T	T/T
29,942,191	rs2232236	T	G/G	del/del	del/del	G/G	del/del	?/?	G/G	del/del	del/del
30,704,985	rs28380598	T	C/C	T/T	T/T	C/C	T/T	T/T	C/C	T/T	T/T
31,079,236	rs1265055	G	A/G	A/A	A/A	A/G	A/A	A/A	A/G	A/A	A/A
31,082,304	rs3130554	T	G/T	T/T	T/T	G/T	T/T	T/T	G/T	T/T	T/T
32,130,937	rs10680	T	C/C	T/T	T/T	C/C	T/T	T/T	C/C	T/T	?/?
33,036,549	rs17509489	T	G/G	C/T	C/T	G/G	T/T	T/T	G/G	T/T	T/T

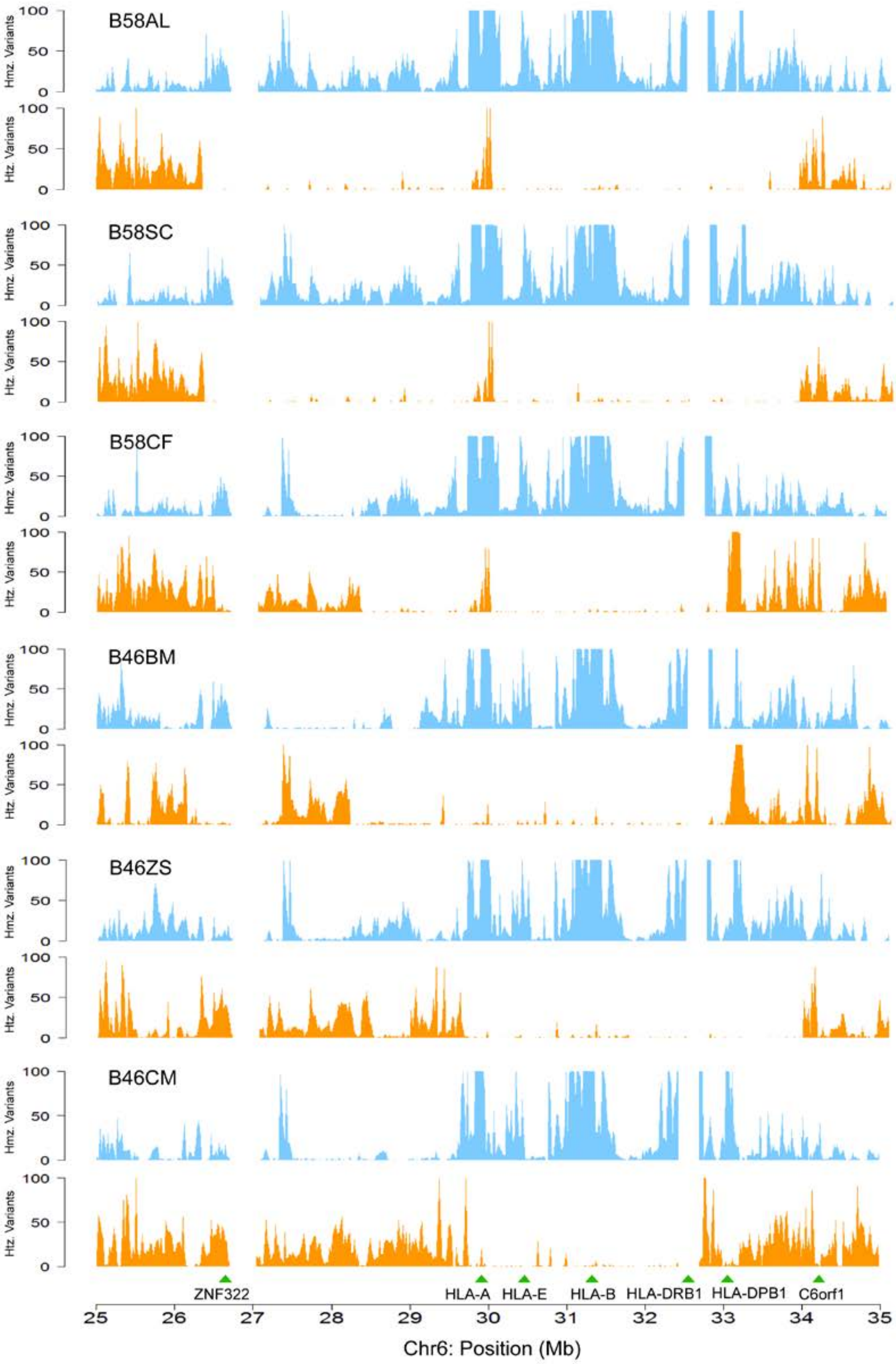


Figure 4.2 Zygosity profile of variants derived from deep sequencing. The region chr6: 25Mb – 35Mb was binned into non-overlapping windows of 20kb. The number of homozygous (hmz) and heterozygous (htz) variants with respect to the NCBI Build 37.2 reference sequence was calculated in each bin. For each of the six individuals, the upper panel plot represents the homozygous variants counts while the lower panel plot represents the heterozygous variants counts across the region of interest.

Table 4.4 Range of conserved extended region for the six cell lines derived from CG data.

Cell line	Start	End	Length (Mb)	No. of htz variants
B58AL	26,630,000	33,830,000	7.48	595
B58SC	26,350,000	33,810,000	7.46	578
B58CF	28,350,000	32,950,000	4.60	455
B46BM	29,367,500	32,917,500	3.55	221
B46ZS	29,630,000	33,890,000	4.26	146
B46CM	29,550,000	32,650,000	3.10	167

CG data showed at least 99.99% homozygosity in all samples, with the A33-B58-DR3 haplotypes samples having on average 8.33 heterozygous calls per 100kb, while the A2-B46-DR9 haplotypes samples having on average 4.89 heterozygous calls per 100kb. Large numbers of homozygous reference variant calls were observed within or proximate to the HLA genes, highlighting the difference in the HLA allelic combination between the reference genome and our samples. Despite the extensive number homozygous reference variants, small pockets of heterozygous variants randomly spread across the conserved MHC region were observed. Of interest, spike of heterozygosity centromeric of the HLA-A gene was detected in the A33-B58-DR3 samples even though high density SNP profiling have indicated high level of homozygosity in this genomic region. This characteristic has not been observed in the previous studies of CEH as the high density SNPs geneotyping platform is not able to provide the necessary resolution.

4.2.2 Intra-haplotypic Conservation and Variation

The use of three HLA homozygous diploid samples of each Asian CEH offered the opportunity to characterize the extent of intra-haplotypic variation and conservation. Using the CG platform, the three diploid samples of each haplotype yielded 6 haploid sequences, which were compared to each other at each nucleotide position across the length of the sample with the shortest range of homozygosity. The range compared for A33-B58-DR3 was chr6:28,350,000-32,950,000, and the range compared for A2-B46-DR9 was chr6:29,630,000-32,650,000. Nucleotide positions having ambiguous or low confidence score in two or more haploid would be regarded as no-called (“N”) in the consensus sequence. Accounting for the gaps within the region, we were able to establish 4,135,945 bp of phase-discrete sequence representative of the A33-B58-DR3 haplotype and 2,720,646 bp of phase-discrete sequence representative of the A2-B46-DR9 haplotype. Next, the number of positions with SNVs and insertions/deletions (indels) between the six sequences were computed, and plotted against chromosome position (Figure 4.3). Within the conserved region, the degree of intra-haplotype variation was found be exceptionally low; 293 SNVs and 52 indels were identified in the A33-B58-DR3 haplotype, while 238 SNVs and 51 indels were observed in A2-B46-DR9 haplotype. A closer inspection revealed that majority of the intra-haplotypic variations in each MHC haplotype was localized to a single region. For example, spikes of variation localized to a 120kb region covering the *ZFP57* and *HLA-F* gene (chr6:29,600,360 – 29,721,396) were found in the A2-B46-DR9 haplotype but not in A33-B58-DR3 haplotype (Figure 4.4A) and these variations make up >70% (171/238) of the total A2-B46-DR9 intra-haplotyic SNVs. Similarly, elevated number of A33-B58-DR3 intra-haplotypic variations accounting for 90% (262/293) of SNVs were observed at a 240kb region covering the *HLA-A* gene (chr6: 29,733,502 – 29,971,973) while the number of intra-haplotypic variation in the A2-B46-BR9 in this region was distinctly lesser (Figure 4.4B). The estimated nucleotide diversity value (π) between the A33-B58-DR3 haploid sequences and A2-B46-DR9 haploid sequences was 7.08 X 10^{-5} and 8.75 X 10^{-5} respectively. In comparison, these values are 38- to 48-fold lower than the nucleotide diversity found between PGF and COX (3.4 X 10^{-3}) [15], the two common MHC haplotype found in the European population and at least 5-fold lower than the nucleotide diversity between any two haplotypes across the human genome [175, 176]; indicative of extreme low nucleotide diversity in the A33-B58-DR3 and A2-B46-DR9 MHC haplotypes.

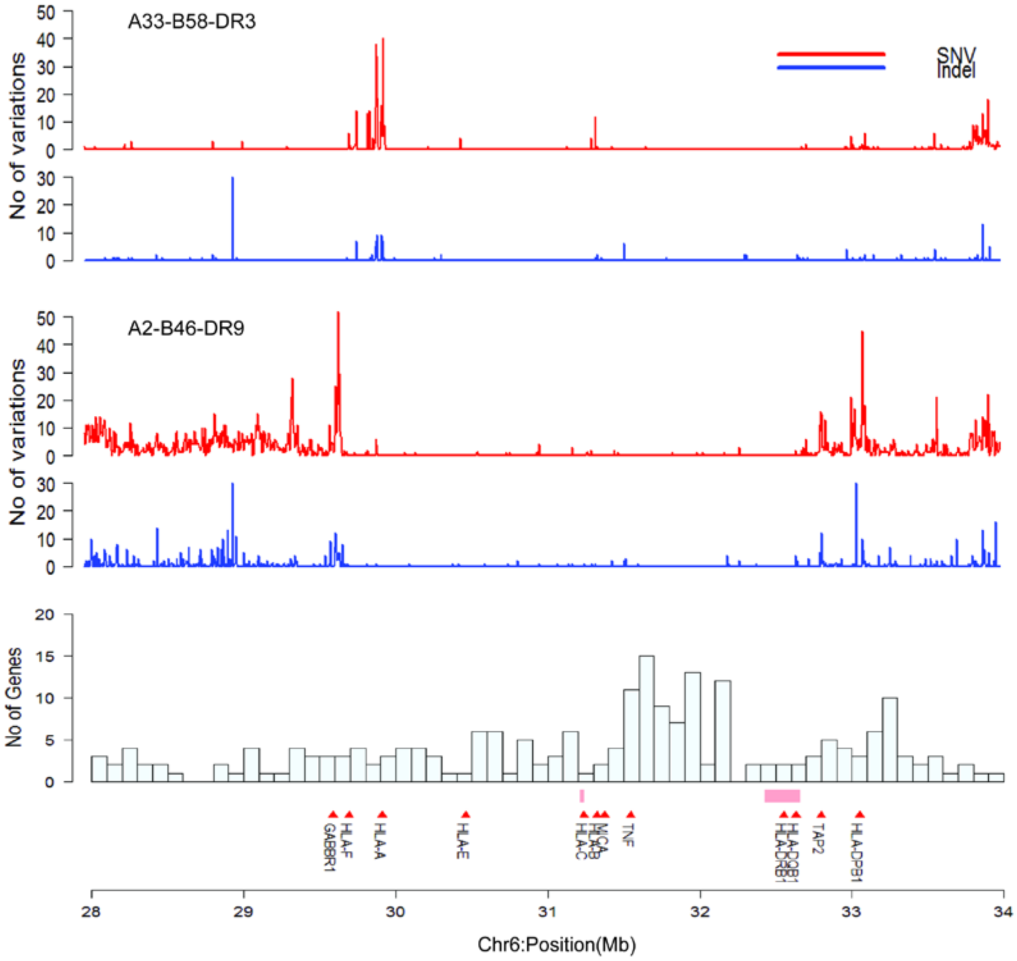


Figure 4.3 Distribution of intra-haplotypic variations across the MHC region. Each data point on the plot represents the number of SNV counts (red) and the number of indel counts (blue) in a non-overlapping 5kb window. The number of variations for each haplotype was derived from the comparisons of six haploid chromosomes at every possible nucleotide position across the MHC region. The pink bars indicated regions where the sequences are ambiguous and nucleotide positions within these regions are not compared.

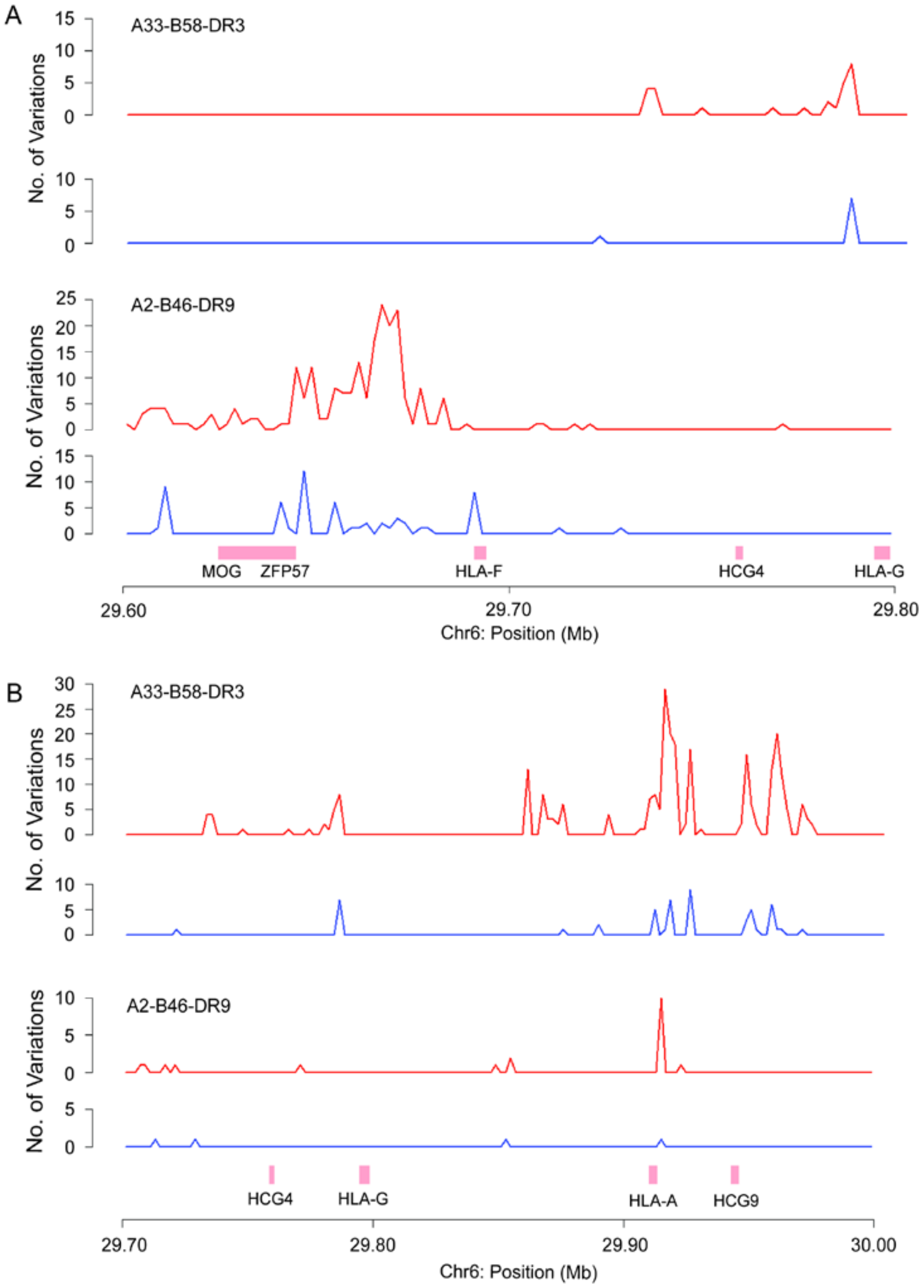


Figure 4.4 Spikes in Intra-haplotypic variations found in **(A)** A2-B46-DR9 haplotype and **(B)** A33-B58-DR3 haplotype. The variation counts were binned into non-overlapping 2kb windows. Heighten intra-haplotypic variation genomic segments extended up to 120kb and 240kb were identified in the A2-B46-DR9 and A33-B58-DR3 haplotype respectively.

Outside the boundaries of conserved region, the number of variations between the haploid sequences was significantly increased.

Using NCBI RefSeq Build 37.2 gene annotation, intra-haplotypic variations were grouped into functional categories (Table 4.5) with the majority of the variants (~80%) resided in the non-coding region, while less than 1.5% of all variants were located in the exonic region (Figure 4.5). Coding region variants showed equal proportion of missense and synonymous mutations (Table 4.6). Surprisingly, missense mutations were found in A33-B58-DR3 haplotype samples, on exon 7 of *HLA-A* and exon 1 of *HLA–B*. To validate these variants, the *HLA-A* exon 7 (position 29,913,037) missense variant was cloned and re-sequenced. The results confirmed its presence in all three samples of A33-B58-DR3 (Figure 4.6, Figure 4.7 and Figure 4.8) and within this 479 bp cloned fragment, 13 heterozygous and 15 homozygous variants were in agreement with the CG data (Table 4.7). Using *in-silico* tools to predict the possible functional effect of these missense mutations [177, 178], it was found that the induced amino acid substitutions would have minimal effect on the protein function (Table 4.6). In total, 77 novel SNV variants were identified within the A33-B58-DR3 samples, and 50 novel SNV variants within the A2-B46-DR9 samples that were not annotated in dbSNP build 132, the 1000 Genome Project and the International HapMap Project (Table 4.8 and Table 4.9). The discovery of these novel SNPs provides a potentially powerful set of markers in disease-association studies.

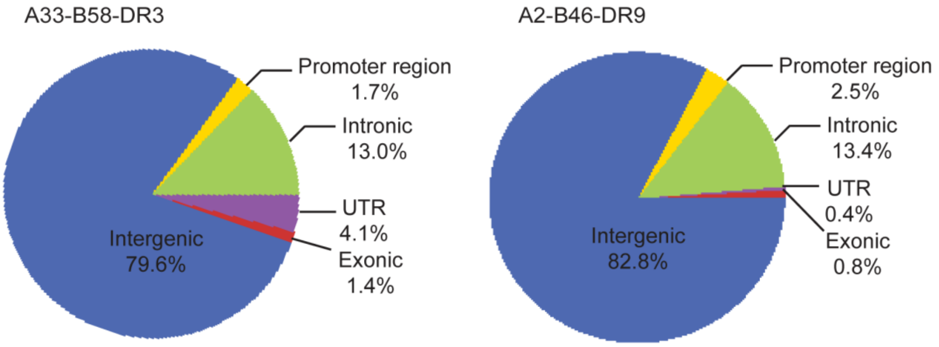


Figure 4.5 Functional annotation of variants for each haplotype based on NCBI gene annotation Build 37.2. It was noted majority of the variants were found in the non-coding region while only less than 1.5% were found in the exonic region for both haplotypes.

Table 4.5 Intra- and inter-haplotypic comparisons: single nucleotide variants (SNVs) and insertion/deletion (indels).

	Intra-haplotype		Inter-haplotype					
			A2-B46-DR9	A2-B46-DR9	A33-B58-DR3	A2-B46-DR9	A33-B58-DR3	PGF
			vs	vs	vs	vs	vs	vs
Variant Type	A2-B46-DR9	A33-B58-DR3	A33-B58-DR3	PGF	PGF	COX	COX	COX
SNVs								
Coding	2	4	161	126	204	160	206	190
Missense	1	2	84	68	111	91	113	99
Nonsense	0	0	1	2	1	0	1	2
Synonymous	1	2	76	56	92	69	92	89
ncRNA exonic	0	0	138	106	119	148	152	115
UTR 5'	1	1	24	28	34	24	32	37
UTR 3'	0	11	75	54	87	64	79	78
Intronic	32	38	1745	1320	1704	1577	1859	1529
Promoter region	6	5	217	228	242	267	222	195
Intergenic	197	234	5141	4250	6795	4779	6549	7911
Total	**238**	**293**	**7501**	**6112**	**9185**	**7019**	**9099**	**10055**
SNVs/100kb	**8.58**	**6.84**	**270**	**220**	**211**	**253**	**217**	**237**
Indels								
Coding	0	0	3	2	2	5	4	3
Frameshift	0	0	3	2	2	3	3	1
Non-frameshift	0	0	0	0	0	2	1	2
ncRNA exonic	1	0	17	9	18	18	24	19
UTR	1	1	12	6	18	9	16	15
Intronic	9	12	208	135	180	298	363	381
Promoter region	3	2	26	19	19	37	44	36
Intergenic	37	37	418	298	516	688	993	1298
Total	**51**	**52**	**684**	**469**	**753**	**1055**	**1444**	**1752**
Indels/100kb	**1.8**	**1.22**	**24.6**	**16.9**	**17.3**	**37.9**	**34.4**	**41.2**

Table 4.6 Intra-haplotype SNVs within coding sequence region.

A2-B46-DR9

Start	End	Alternate Allele	Reference Allele	Variant Type	Mutation	Gene	Exon	Amino acid change	Effect	
									PolyPhen2	SIFT
29,634,002	29,634,003	C	G	SNV	missense	MOG	3	V171L	Benign	Tolerated
32,190,483	32,190,484	A	G	SNV	synon	NOTCH4	3	-	-	-

A33-B58-DR3

Start	End	Alternate Allele	Reference Allele	Variant Type	Effect	Gene	Exon	Amino acid change	Effect	
									PolyPhen2	SIFT
29,913,036	29,913,037	A	G	SNV	missense	HLA-A	7	V358M	Benign	Tolerated
31,324,886	31,324,887	C	G	SNV	missense	HLA-B	1	L17V	Benign	Tolerated
31,324,887	31,324,888	T	G	SNV	synon	HLA-B	1	-	-	-
31,324,890	31,324,891	T	C	SNV	synon	HLA-B	1	-	-	-

SNV: Single nucleotide variant; synon: synonymous substitution

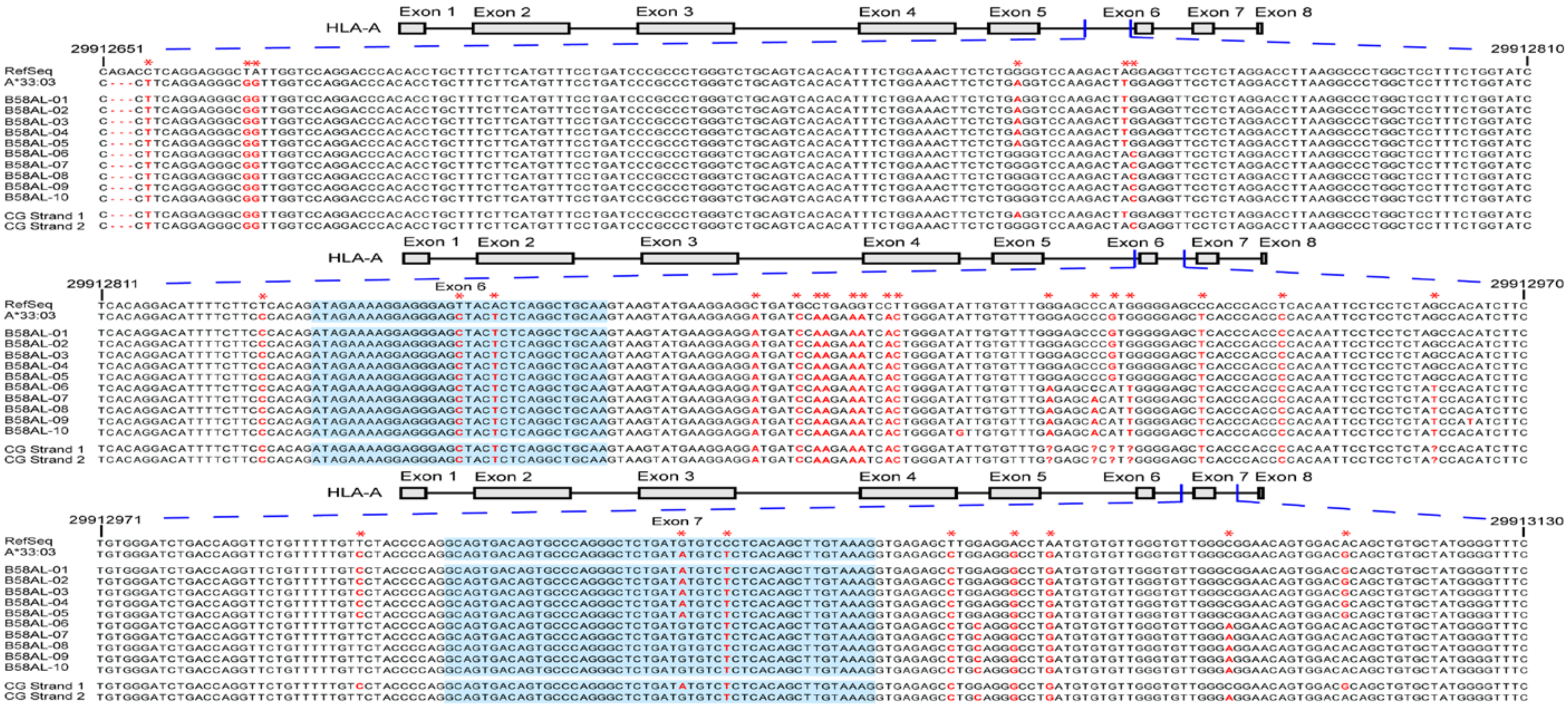


Figure 4.6 Alignment of B58AL DNA templates derived via TA cloning. The DNA template was mapped to NCBI Build 37.2 reference sequence chr6:29,912,732 – 29,913,130 located within the HLA-A gene. The derived DNA sequences were also aligned with the sequence of A*33:03 allele downloaded from IMGT/HLA database (http://www.ebi.ac.uk/ipd/imgt/hla/, Release 3.13.1). Strong correlation in the sequences derived from TA cloning with the CG data.

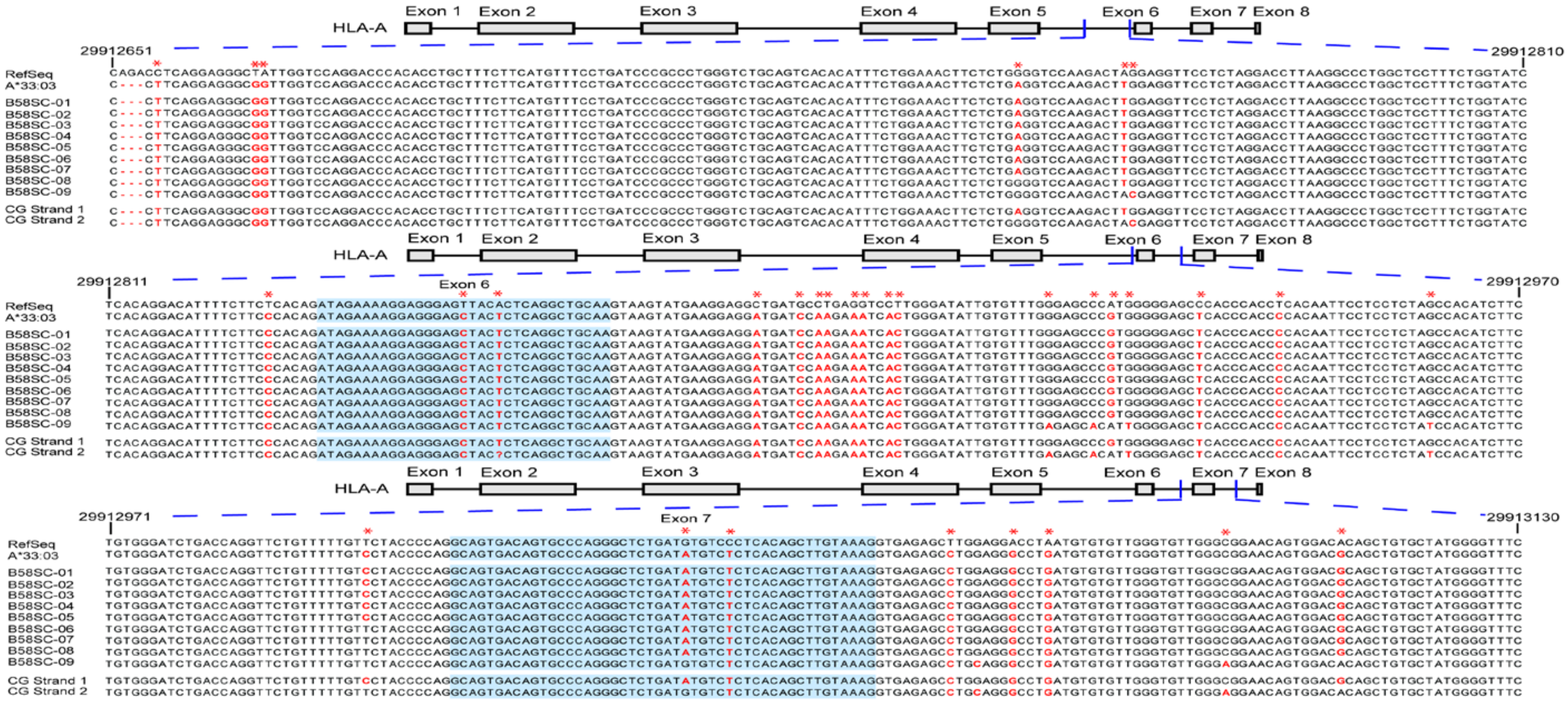


Figure 4.7 Alignment of B58SC DNA templates derived via TA cloning. The DNA template was mapped to NCBI Build 37.2 reference sequence chr6:29,912,732 – 29,913,130 located within the HLA-A gene. The derived DNA sequences were also aligned with the sequence of A*33:03 allele downloaded from IMGT/HLA database (http://www.ebi.ac.uk/ipd/imgt/hla/, Release 3.13.1). Strong correlation in the sequences derived from TA cloning with the CG data.

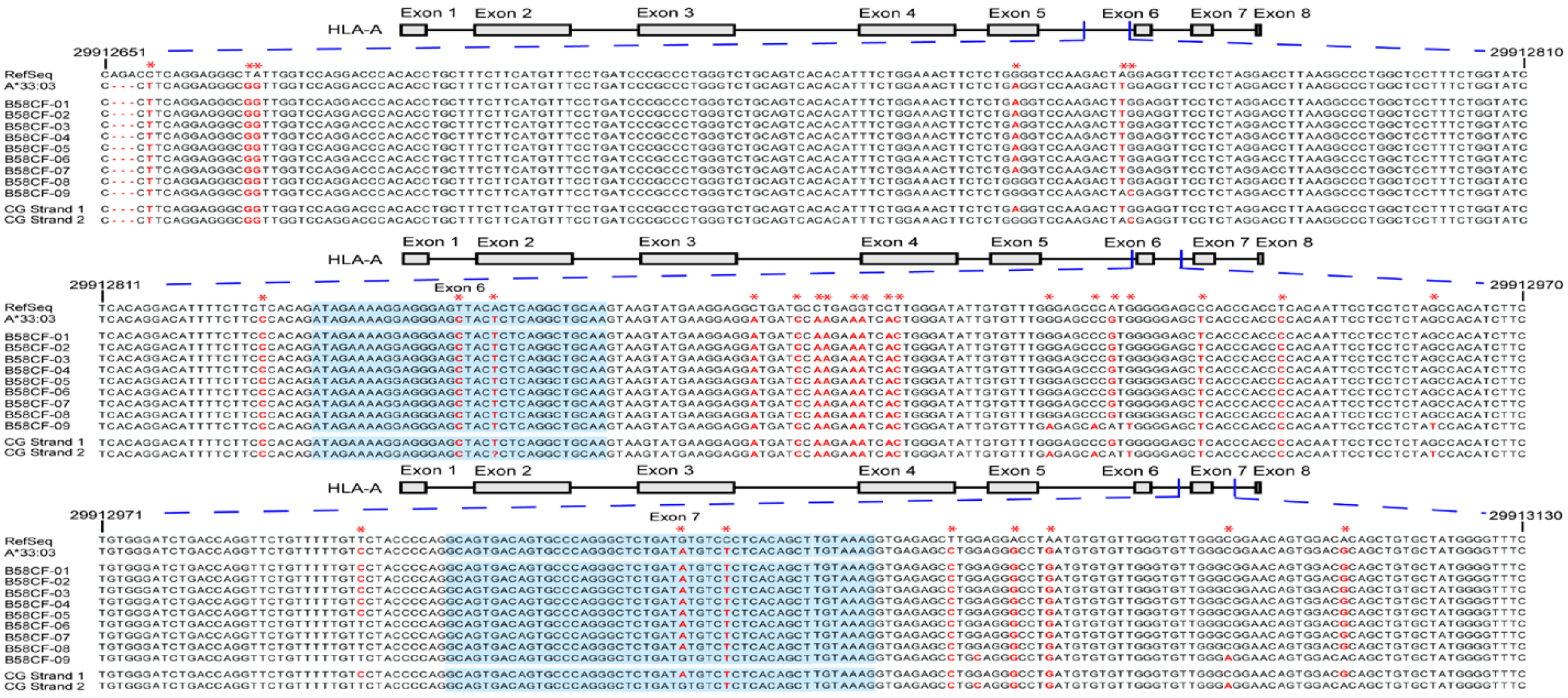


Figure 4.8 Alignment of B58CF DNA templates derived via TA cloning. The DNA template was mapped to NCBI Build 37.2 reference sequence chr6:29,912,732 – 29,913,130 located within the HLA-A gene. The derived DNA sequences were also aligned with the sequence of A*33:03 allele downloaded from IMGT/HLA database (http://www.ebi.ac.uk/ipd/imgt/hla/, Release 3.13.1). Strong correlation in the sequences derived from TA cloning with the CG data.

Table 4.7 Variants within DNA template chr6:29,912,732 – 29,913,130.

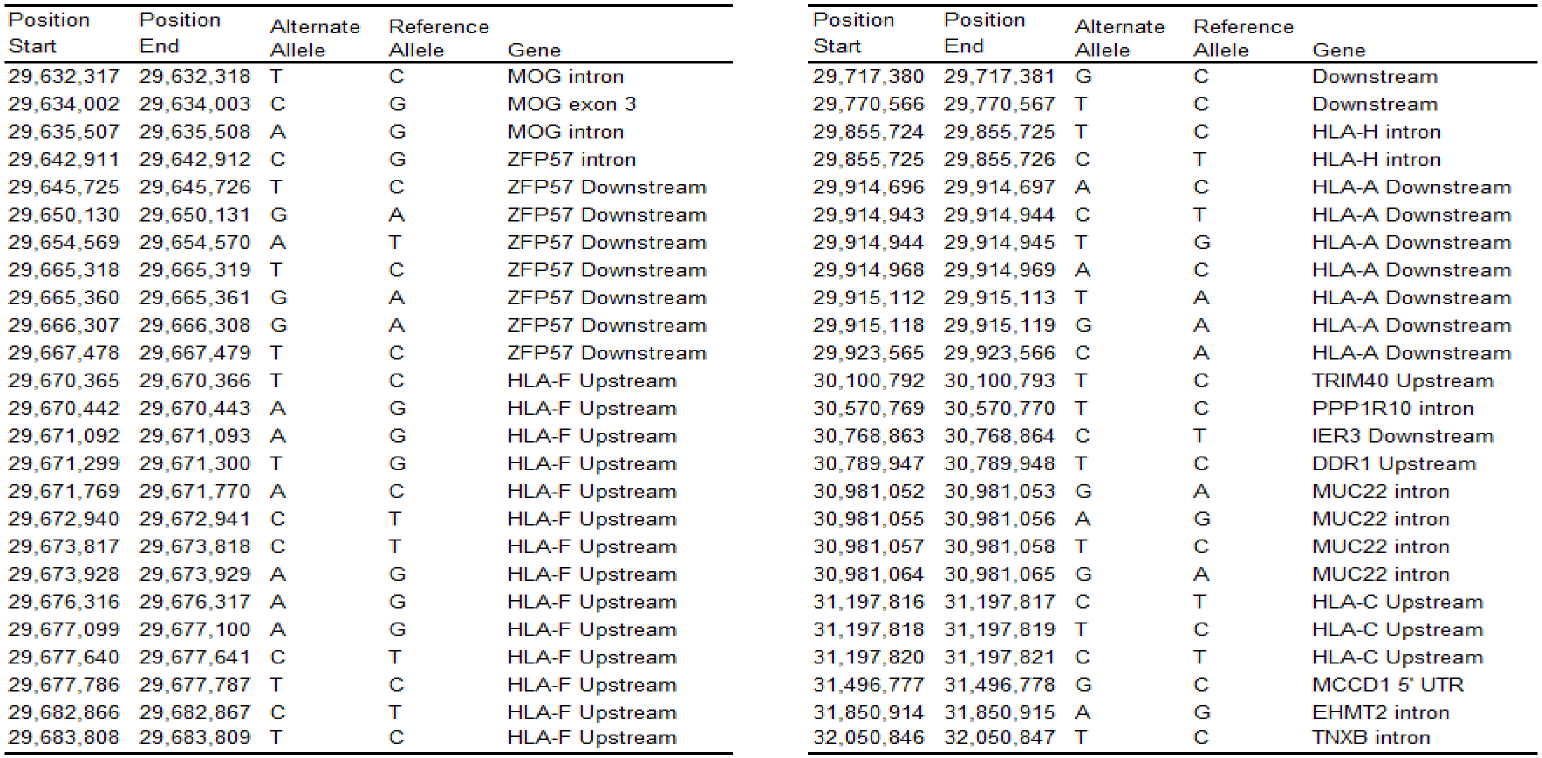
						B58AL		B58SC		B58CF	
Position Start	Position End	Region	VarType	Zygosity	Reference	CG Sequence	TA Cloning	CG Sequence	TA Cloning	CG Sequence	TA Cloning
29,912,652	29,912,657	Intron	SUB	Hmz	AGACC	CT/CT	CT/CT	CT/CT	CT/CT	CT/CT	CT/CT
29,912,667	29,912,669	Intron	SUB	Hmz	TA	GG/GG	GG/GG	GG/GG	GG/GG	GG/GG	GG/GG
29,912,753	29,912,754	Intron	SNV	Htz	G	G/A	G/A	G/A	G/A	G/A	G/A
29,912,765	29,912,766	Intron	SNV	Htz	A	A/T	A/T	A/T	A/T	A/T	A/T
29,912,766	29,912,767	Intron	SNV	Htz	G	G/C	G/C	G/C	G/C	G/C	G/C
29,912,829	29,912,830	Intron	SNV	Hmz	T	C/C	C/C	C/C	C/C	?/?	C/C
29,912,851	29,912,852	Exon 6	SNV	Hmz	T	C/C	C/C	C/C	C/C	C/C	C/C
29,912,855	29,912,856	Exon 6	SNV	Hmz	A	T/T	T/T	T/?	T/T	T/T	T/T
29,912,884	29,912,885	Intron	SNV	Hmz	C	A/A	A/A	A/A	A/A	A/A	A/A
29,912,889	29,912,893	Intron	SUB	Hmz	GCCT	CCAA/CCAA	CCAA/CCAA	CCAA/CCAA	CCAA/CCAA	CCAA/CCAA	CCAA/CCAA
29,912,895	29,912,897	Intron	SUB	Hmz	GG	AA/AA	AA/AA	AA/AA	AA/AA	AA/AA	AA/AA
29,912,899	29,912,901	Intron	SUB	Hmz	CT	AC/AC	AC/AC	AC/AC	AC/AC	AC/AC	AC/AC
29,912,917	29,912,918	Intron	SNV	Htz	G	?/?	G/A	G/A	G/A	?/?	G/A
29,912,922	29,912,923	Intron	SNV	Htz	C	?/?	C/A	C/A	C/A	?/?	C/A
29,912,924	29,912,925	Intron	SNV	Htz	A	?/?	A/G	A/G	A/G	?/?	A/G
29,912,926	29,912,927	Intron	SNV	Htz	G	?/?	G/T	G/T	G/T	?/?	G/T
29,912,934	29,912,935	Intron	SNV	Hmz	C	T/T	T/T	T/T	T/T	T/T	T/T
29,912,943	29,912,944	Intron	SNV	Hmz	T	C/C	C/C	C/C	C/C	C/C	C/C
29,912,960	29,912,961	Intron	SNV	Htz	G	?/?	G/T	G/T	G/T	G/?	G/T
29,913,000	29,913,001	Intron	SNV	Htz	T	T/C	T/C	T/C	T/C	T/C	T/C
29,913,036	29,913,037	Exon 7	SNV	Htz	G	G/A	G/A	G/A	G/A	G/A	G/A
29,913,041	29,913,042	Exon 7	SNV	Hmz	C	T/T	T/T	T/T	T/T	T/T	T/T
29,913,066	29,913,067	Intron	SNV	Hmz	T	C/C	C/C	C/C	C/C	C/C	C/C
29,913,069	29,913,070	Intron	SNV	Htz	G	G/C	G/C	G/C	G/C	G/C	G/C
29,913,073	29,913,074	Intron	SNV	Hmz	A	G/G	G/G	G/G	G/G	G/G	G/G
29,913,077	29,913,078	Intron	SNV	Hmz	A	G/G	G/G	G/G	G/G	G/G	G/G
29,913,097	29,913,098	Intron	SNV	Htz	C	C/A	C/A	C/A	C/A	C/A	C/A
29,913,110	29,913,111	Intron	SNV	Htz	A	A/G	A/G	A/G	A/G	A/G	A/G

Hmz : Homozygous; Htz: Heterozygous; SUB: substitution; SNV: Single nucleotide variants

Table 4.8 Novel SNVs in the A33-B58-DR3 haplotype.

Position Start	Position End	Alternate Allele	Reference Allele	Gene
28,840,256	28,840,257	C	A	LOC401242 Downstream
29,035,463	29,035,464	G	A	LOC100129636 intron
29,035,477	29,035,478	T	C	LOC100129636 intron
29,734,245	29,734,246	T	G	IFITM4P Downstream
29,747,779	29,747,780	A	C	IFITM4P Downstream
29,765,233	29,765,234	C	A	LOC554223 3' UTR
29,779,058	29,779,059	G	A	HLA-G Downstream
29,779,067	29,779,068	G	A	HLA-G Downstream
29,783,925	29,783,926	C	T	HLA-H Upstream
29,783,930	29,783,931	T	C	HLA-H Upstream
29,783,931	29,783,932	A	C	HLA-H Upstream
29,784,433	29,784,434	C	T	HLA-H Upstream
29,859,618	29,859,619	A	C	HLA-H Downstream
29,859,620	29,859,621	C	T	HLA-H Downstream
29,859,927	29,859,928	G	A	HLA-H Downstream
29,859,928	29,859,929	T	C	HLA-H Downstream
29,867,025	29,867,026	T	A	HLA-H Downstream
29,867,026	29,867,027	G	C	HLA-H Downstream
29,869,426	29,869,427	G	A	HLA-H Downstream
29,869,429	29,869,430	A	C	HLA-H Downstream
29,869,434	29,869,435	C	T	HLA-H Downstream
29,872,187	29,872,188	G	A	HCG4B Upstream
29,872,188	29,872,189	A	G	HCG4B Upstream
29,872,841	29,872,842	A	T	HCG4B Upstream
29,872,843	29,872,844	A	T	HCG4B Upstream
29,890,605	29,890,606	A	C	HCG4B Upstream
29,903,515	29,903,516	G	C	HLA-A Upstream
29,904,860	29,904,861	T	C	HLA-A Upstream
29,906,004	29,906,005	A	C	HLA-A Upstream
29,906,078	29,906,079	C	A	HLA-A Upstream
29,906,079	29,906,080	T	C	HLA-A Upstream
29,906,654	29,906,655	T	C	HLA-A Upstream
29,906,674	29,906,675	C	T	HLA-A Upstream
29,909,682	29,909,683	C	T	HLA-A Upstream
29,909,685	29,909,686	T	C	HLA-A Upstream
29,909,688	29,909,689	A	C	HLA-A Upstream
29,909,834	29,909,835	G	A	HLA-A Upstream
29,913,273	29,913,274	C	T	HLA-A 3' UTR
29,913,837	29,913,838	C	G	HLA-A Downstream
29,913,944	29,913,945	G	A	HLA-A Downstream
29,914,625	29,914,626	C	T	HLA-A Downstream
29,914,943	29,914,944	C	T	HLA-A Downstream
29,914,944	29,914,945	T	G	HLA-A Downstream
29,915,123	29,915,124	A	C	HLA-A Downstream
29,915,147	29,915,148	T	C	HLA-A Downstream
29,915,148	29,915,149	G	A	HLA-A Downstream
29,915,150	29,915,151	G	T	HLA-A Downstream
29,915,966	29,915,967	T	C	HLA-A Downstream
29,926,156	29,926,157	A	G	HCG9 Upstream
29,944,177	29,944,178	A	C	HCG9 intron
29,944,307	29,944,308	G	T	HCG9 intron
29,944,308	29,944,309	G	C	HCG9 intron
29,944,324	29,944,325	A	T	HCG9 intron
29,944,326	29,944,327	G	A	HCG9 intron
29,944,327	29,944,328	C	G	HCG9 intron
29,955,808	29,955,809	C	T	HCG9 Downstream
29,956,118	29,956,119	G	A	HCG9 Downstream
29,956,126	29,956,127	C	G	HCG9 Downstream
29,956,134	29,956,135	C	A	HCG9 Downstream
29,956,137	29,956,138	C	T	HCG9 Downstream
29,956,138	29,956,139	A	G	HCG9 Downstream
29,956,484	29,956,485	C	T	HCG9 Downstream
29,956,486	29,956,487	G	T	HCG9 Downstream
29,960,539	29,960,540	C	T	HLA-J Upstream
29,960,552	29,960,553	A	T	HLA-J Upstream
29,960,564	29,960,565	T	C	HLA-J Upstream
29,967,872	29,967,873	T	C	HLA-J Upstream
29,968,000	29,968,001	A	C	HLA-J Upstream
29,968,002	29,968,003	C	G	HLA-J Upstream
29,968,011	29,968,012	G	A	HLA-J Upstream
29,968,015	29,968,016	C	G	HLA-J Upstream
30,467,075	30,467,076	A	G	HLA-E Downstream
30,467,077	30,467,078	C	G	HLA-E Downstream
30,467,086	30,467,087	G	A	HLA-E Downstream
31,348,653	31,348,654	G	A	HLA-E Downstream
31,348,655	31,348,656	T	G	HLA-E Downstream
31,363,403	31,363,404	T	A	MICA Upstream

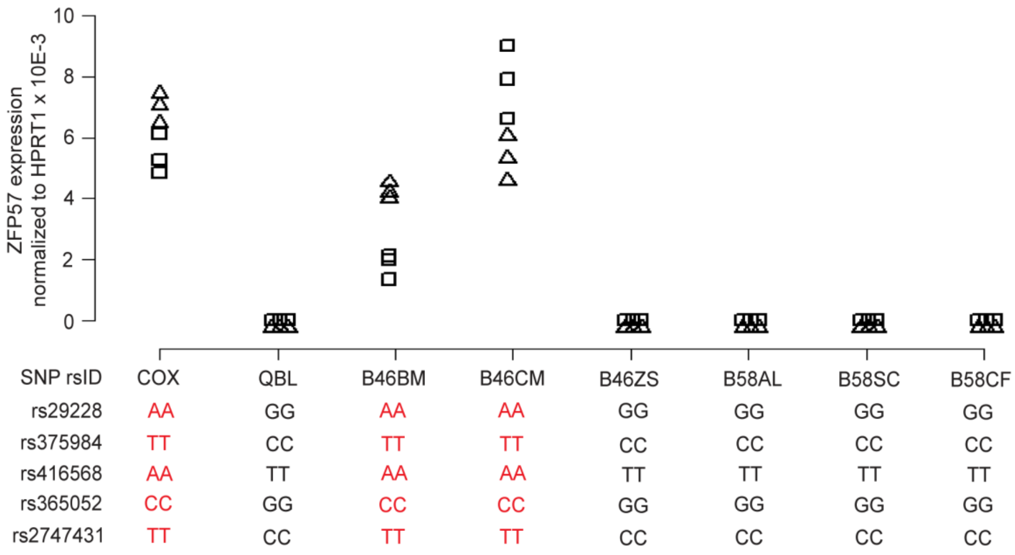
Table 4.9 Novel SNVs in the A2-B46-DR9 haplotype.



Position Start	Position End	Alternate Allele	Reference Allele	Gene
29,632,317	29,632,318	T	C	MOG intron
29,634,002	29,634,003	C	G	MOG exon 3
29,635,507	29,635,508	A	G	MOG intron
29,642,911	29,642,912	C	G	ZFP57 intron
29,645,725	29,645,726	T	C	ZFP57 Downstream
29,650,130	29,650,131	G	A	ZFP57 Downstream
29,654,569	29,654,570	A	T	ZFP57 Downstream
29,665,318	29,665,319	T	C	ZFP57 Downstream
29,665,360	29,665,361	G	A	ZFP57 Downstream
29,666,307	29,666,308	G	A	ZFP57 Downstream
29,667,478	29,667,479	T	C	ZFP57 Downstream
29,670,365	29,670,366	T	C	HLA-F Upstream
29,670,442	29,670,443	A	G	HLA-F Upstream
29,671,092	29,671,093	A	G	HLA-F Upstream
29,671,299	29,671,300	T	G	HLA-F Upstream
29,671,769	29,671,770	A	C	HLA-F Upstream
29,672,940	29,672,941	C	T	HLA-F Upstream
29,673,817	29,673,818	C	T	HLA-F Upstream
29,673,928	29,673,929	A	G	HLA-F Upstream
29,676,316	29,676,317	A	G	HLA-F Upstream
29,677,099	29,677,100	A	G	HLA-F Upstream
29,677,640	29,677,641	C	T	HLA-F Upstream
29,677,786	29,677,787	T	C	HLA-F Upstream
29,682,866	29,682,867	C	T	HLA-F Upstream
29,683,808	29,683,809	T	C	HLA-F Upstream
29,717,380	29,717,381	G	C	Downstream
29,770,566	29,770,567	T	C	Downstream
29,855,724	29,855,725	T	C	HLA-H intron
29,855,725	29,855,726	C	T	HLA-H intron
29,914,696	29,914,697	A	C	HLA-A Downstream
29,914,943	29,914,944	C	T	HLA-A Downstream
29,914,944	29,914,945	T	G	HLA-A Downstream
29,914,968	29,914,969	A	C	HLA-A Downstream
29,915,112	29,915,113	T	A	HLA-A Downstream
29,915,118	29,915,119	G	A	HLA-A Downstream
29,923,565	29,923,566	C	A	HLA-A Downstream
30,100,792	30,100,793	T	C	TRIM40 Upstream
30,570,769	30,570,770	T	C	PPP1R10 intron
30,768,863	30,768,864	C	T	IER3 Downstream
30,789,947	30,789,948	T	C	DDR1 Upstream
30,981,052	30,981,053	G	A	MUC22 intron
30,981,055	30,981,056	A	G	MUC22 intron
30,981,057	30,981,058	T	C	MUC22 intron
30,981,064	30,981,065	G	A	MUC22 intron
31,197,816	31,197,817	C	T	HLA-C Upstream
31,197,818	31,197,819	T	C	HLA-C Upstream
31,197,820	31,197,821	C	T	HLA-C Upstream
31,496,777	31,496,778	G	C	MCCD1 5' UTR
31,850,914	31,850,915	A	G	EHMT2 intron
32,050,846	32,050,847	T	C	TNXB intron

4.2.3 Intra-haplotypic Variations in A2-B46-DR9 Influence the Expression of ZFP57

Although the intra-haplotypic variants appear to be highly homologous, single nucleotide differences may exert functional consequences. To illustrate this mechanism, SNP rs29228 (chr6: 29,623,739) has been found to exert a cis-acting effect on the expression level of the Zinc Finger Protein 57 homolog (*ZFP57*) [179], located 16.43kb centromeric to rs29228. It was reported that carriers of the "AA" but not the "GG" genotype of rs29228 would support expression of *ZFP57*. Such locus where genetic variation is associated with the gene expression variation is commonly known as expression quantitative trait loci (eQTL) SNP [129]. A more recent study revealed four additional eQTL SNPs (chr6:29,644,502 – rs375984, chr6:29,647,628 – rs416568, chr6:29,648,398 – rs365052 and chr6:29,648,564 – rs2747431), located in the *ZFP57* introns and promoter region, were associated with the expression of *ZFP57* [180]. Interestingly, these SNP positions are intra-haplotypic variants in A2-B46-DR9 but not for A33-B58-DR3 haplotype. To learn whether the difference in nucleotide at these positions would affect the expression of *ZFP57*, reverse transcription quantitative PCR was performed to evaluate the ZFP57 mRNA levels in B58AL, B58SC, B58CF, B46BM, B46ZS, B46CM as well as the two European cell line COX and QBL. Noticeably, the COX, B46BM and B46CM cell line possess the "A" allele at chr6: 29,623,739, "T" allele at chr6: 29,644,502, "A" allele at chr6:29,647,628, "C" allele at chr6:29,648,398 and "T" allele at chr6:29,648,564 exhibited evidence of *ZFP57* expression while B46ZS, possessing the alternate allele at these positions, have no *ZFP57* expression (Figure 4.9).



SNP rsID	COX	QBL	B46BM	B46CM	B46ZS	B58AL	B58SC	B58CF
rs29228	AA	GG	AA	AA	GG	GG	GG	GG
rs375984	TT	CC	TT	TT	CC	CC	CC	CC
rs416568	AA	TT	AA	AA	TT	TT	TT	TT
rs365052	CC	GG	CC	CC	GG	GG	GG	GG
rs2747431	TT	CC	TT	TT	CC	CC	CC	CC

Figure 4.9 Association of A2-B46-DR9 intra-haplotypic variants with *ZFP57* expression. RT-qPCR was performed to determine the mRNA level of *ZFP57* in two biological replicates of COX, QBL, B46BM, B46ZS, B46CM, B58AL, B58SC and B58CF cell lines. Experiments were carried out in triplicates for each of the biological replicate. Triangle symbol indicates quantitative expression derived from biological replicate 1 while square symbol indicates quantitative expression derived from biological replicate 2.

Similarly, no *ZFP57* expression was observed in the A33-B58-DR3 and QBL cell lines that possess the alternate alleles. This result highlights the possibility that intra haplotypic variants would have an effect on the expression of genes within the MHC region among individuals carrying identical MHC CEHs.

The presence of eQTL SNPs associated with *ZFP57* indicates possible regulatory role for polymorphic sites in the genomic region proximate to these eQTL SNPs. To identify putative regulatory variants to *ZFP57*, we examined the intra-haplotype sequence variations in the A2-B46-DR9 cell lines in the 90kb genomic region (chr6:29,600,000 – 29,690,000) encompassing the *ZFP57* gene. A total of 202 A2-B46-DR9 intra-haplotypic SNVs were found within the genomic segment of interest and the majority of these variants were localised centromeric of the *ZFP57* (Figure 4.10). Interestingly, the nucleotide call of the 170 variants in the B46ZS cell line matched with the three cell lines carrying the A33-B58-DR3 haplotypes, suggesting these sites as potential regulatory variant candidates. Next, we determined the epigenetic landscape of this 90kb genomic region using the histone modifications data in BLCL obtained from the ENCODE project [181]. Sequences bearing H3K27ac, H3K4Me1, H3K4Me3 marks and DNase I hypersensitivity sites are reported to indicate the presence of enhancers and transcriptional activities [182-184]. Our analysis of the histone marks showed two elevated peaks of histone modification overlapped with a cluster of 25 intra-haplotypic variants at the intron 1 or promoter region of *ZFP57* (chr6: 29,645,000 – 29,650,000) (Figure 4.10). This provides suggestive evidence that polymorphic sites in this segment could have regulatory function.

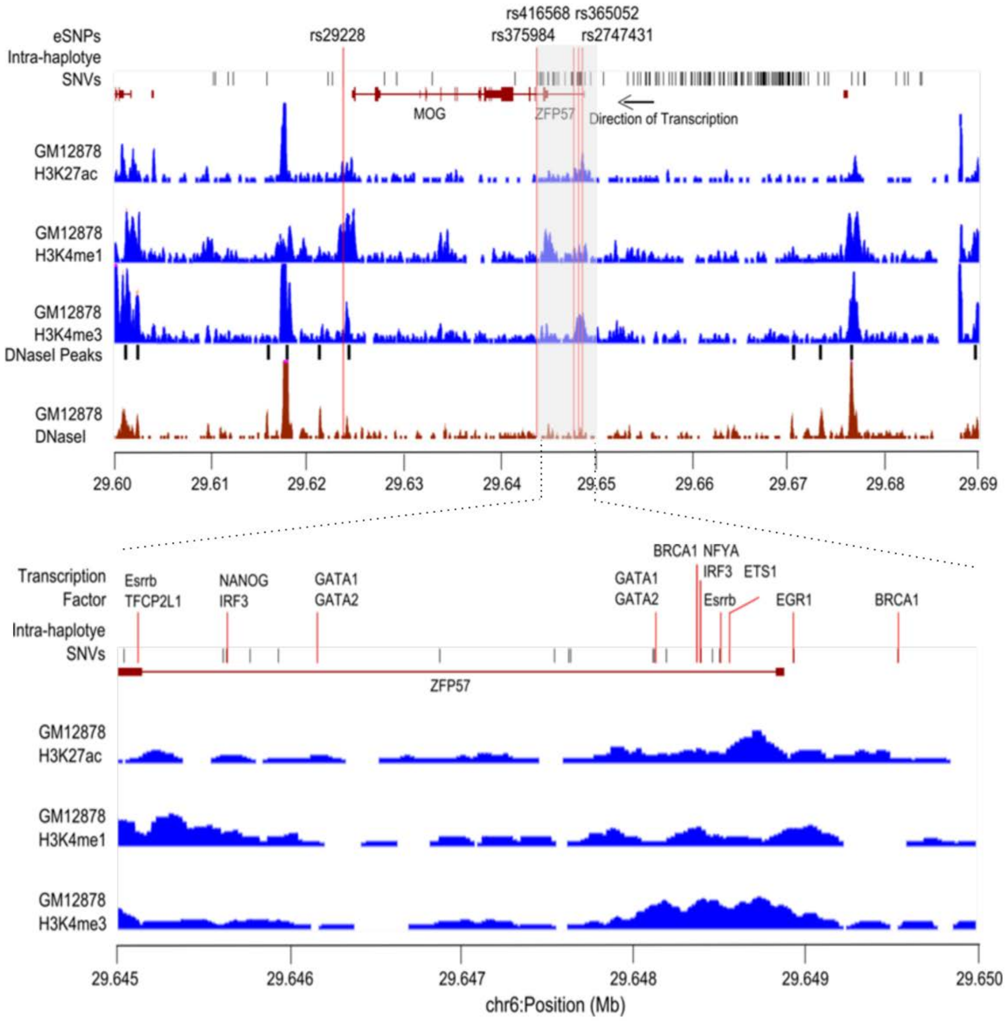


Figure 4.10 Mapping of putative regulatory variants for *ZFP57*. Epigenetic landscape in the 90kb genomic region. Histone modifications data from the ENCODE project based on the profiling of B-LCL (GM12878) were accessed through the UCSC Genome Browser (http://genome.ucsc.edu/). The lower panel represents the intron 1 or promoter region of *ZFP57*. Through *in-silico* analysis, 10 polymorphic sites in this region were identified to interact with transcription factors.

In-silico transcription factors binding sites prediction using Physbinder [185] revealed 10/25 of the polymorphic sites were potential binding sequence for transcription factors such as GATA1, GATA2, IRF3, NFYA, ETS1 and BRCA1 (Figure 4.10) and sites with no predicted transcription factors binding were mostly located at the region between the two histone modification peaks. These highlight that the 10 sites of intra-haplotypic variant have potential binding affinity to transcription factors.

4.2.4 Inter-haplotype Evaluation Reveals Non-random Genetic Variation across MHC Region

Next, the A33-B58-DR3 and A2-B46-DR9 CEH were compared against eight haplotypes of European origin: PGF, COX, QBL, APD, DBB, MANN, MCF and SSTO [14] to examine the degree of sequence variation or similarity. The derived consensus haploid sequences for the A33-B58-DR3 and A2-B46-DR9 haplotypes were aligned pairwise with each of the eight European MHC haploid data. For each pairwise comparison, sequence gaps and positions of no-calls in either haplotype compared were excluded, we observed distinct regions of increased variation at chromosomal regions around HLA-A (29.6-30.0Mb) and HLA-C – HLA-B (31.25-31.5Mb) (Figure 4.11). Since the various haplotypes compared have different HLA-A, -B and -C alleles, divergence detected at these loci are expected. In addition, even though the HLA loci alone are less than 12kb in length and the peaks of variation can stretch to more than 200kb in length.

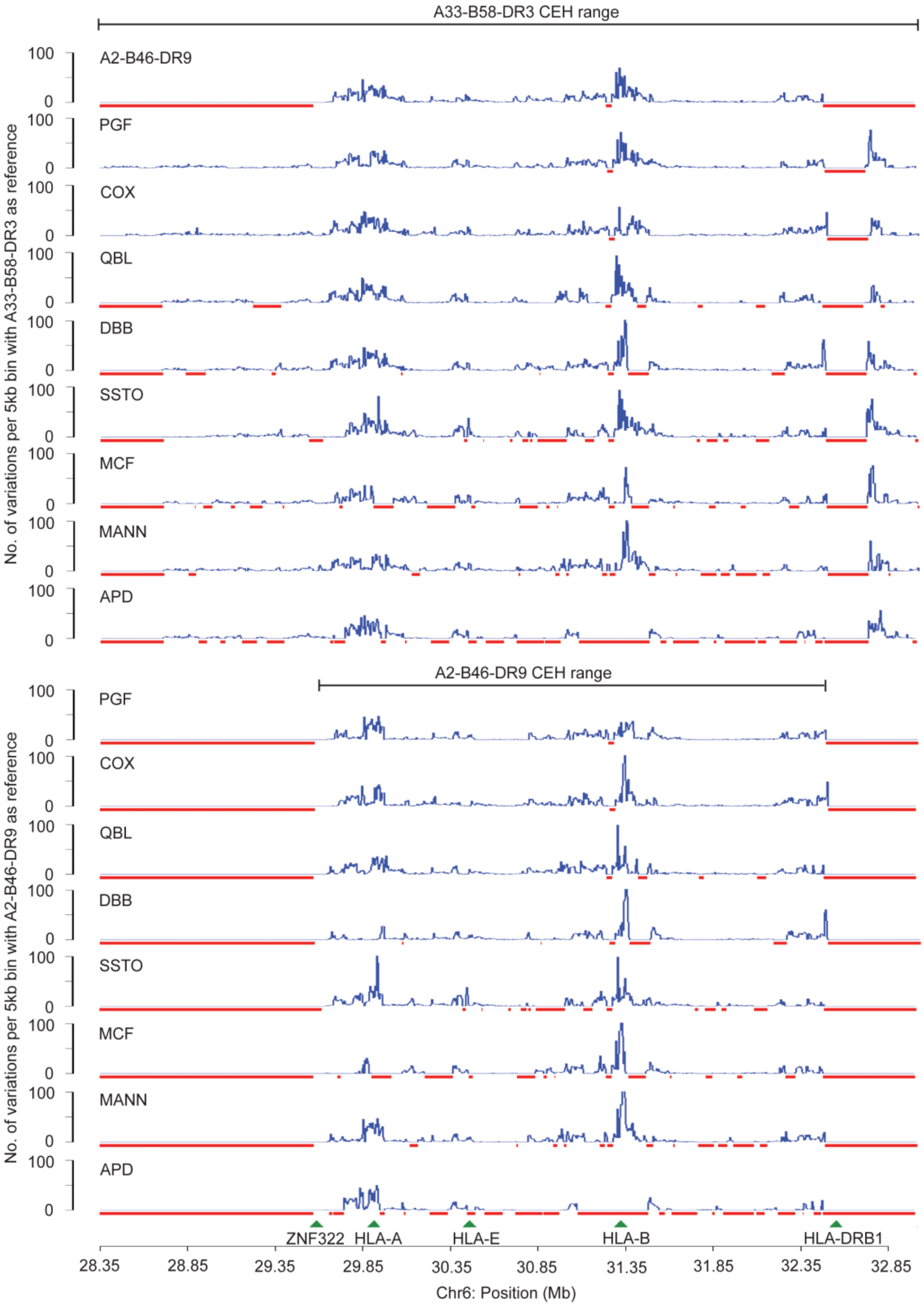


Figure 4.11 Pairwise inter-haplotypic variations across MHC region. A33-B58-DR3 and A2-B46-DR9 haplotype sequence were compared with each of the eight common European-descent MHC haplotype (PGF, COX, QBL, DBB, SSTO, MCF, MANN, APD). The variation counts were binned into non-overlapping 5kb windows. The red bars indicate gaps in the sequence of the European MHC haplotypes. Elevated sequence variation between haplotypes are localized to regions proximate to the HLA genes.

It has been reported that the flanking regions of the HLA genes extending up to thousands of kilobases can still be strongly linked with the associated HLA alleles [163, 174], and thus explains the extended increased in variation surrounding the HLA genes between haplotypes. We also performed functional characterization of the variants between haplotypes specifically the comparison of the 2 Asian CEH haplotypes with PGF and COX. Unsurprisingly, the number of inter-haplotype variation was significantly higher than the number of intra-haplotype variation (Table 4.6). Likewise, there was an elevated percentage (~2.2%) of inter-haplotype variation occurred in the gene coding region. The estimated π between the Asian haplotypes was 2.70 X 10^{-3}, and between the Asian and the European haplotypes the π values range from 2.11 X10^{-3} to 2.55 X 10^{-3}. These values indicate that the MHC region sequence differences between the Asian haplotypes are not significantly greater or lesser than between Asian and European.

To examine the patterns of inter-haplotype variation, we binned the 28.35-32.95Mb into windows of length 5kb, and a frequency histogram of the number of variations for each window was plotted (Figure 4.12A).The distribution observed was skewed left, with most windows had relatively low amounts of variation, albeit a few windows had extremely large amounts of variation. There is no window that had less than 1 variant, indicating that there is no large region of complete conservation across haplotypes within the extended MHC region. Figure 4.12B shows the range of the number of variations for each 10-percentile block. The top 10% of windows with the most number of variants have between 28-109 variants each, which are 5-20 times the MHC region-wide average of 5.39, implying that variation between the haplotypes across the MHC region is generally constant except in regions surrounding the HLA genes. Next, we identified regions of length >30kb with all windows containing below 3 variants or above 15 variants (Table 4.10) to mark out regions of low and high variation respectively.

Surprisingly, a cluster of low-variation regions was observed surrounding the RCCX region, which included the *C2* and *RAGE* loci. This may indicate conservation of these essential components of the innate immune system. The regions containing the highest amounts of variation were, as expected, the class I and class II loci and their neighboring regions. The amount of variation in these regions is remarkable, with on average 8 times the amount of variation compared to the MHC region-wide average.

Table 4.10 Regions of low and high variation between the sequenced MHC haplotypes

Regions with variations below genome average

Start	End	Length (kb)	Mean variant /5kb	Genes
28,350,000	28,415,000	65	0.61	ZSCAN12, ZSCAN23
28,430,000	28,495,000	65	1.96	GPX6
28,560,000	28,670,000	110	0.78	Downstream SCAND3, Upstream LOC401242
29,545,000	29,630,000	85	1.47	SNORD32B, OR2H2, GABBR1
30,545,000	30,680,000	135	1.49	PPP1R10, MRPS18B, ATAT1, C6orf136, DHX16, PPP1R18, NRM
31,690,000	31,805,000	115	1.66	DDAH2, CLIC1, MSH5, SAPCD1, VWA7, VARS, LSM2, HSPA1L, HSPA1A, HSPA1B
31,845,000	31,895,000	50	1.90	EHMT2, C2, ZBTB12
31,905,000	31,965,000	60	1.39	C2, CFB, NELFE, SKIV2L, DOM3Z, STK19, C4A, C4B
31,980,000	32,160,000	180	1.84	CYP21A2, TNXB, ATF6B, FKBPL, PRRT1, LOC100507547, PPT2, EGFL8, AGPAT1, RNF5, AGER,PBX2
32,855,000	32,895,000	40	1.35	LOC100294145

Regions with extreme variation

Start	End	Length (kb)	Mean variant /5kb	Genes
29,665,000	29,715,000	50	16.16	HLA-F
29,730,000	29,805,000	75	26.54	HCG4, LOC554223, HLA-G
29,820,000	29,975,000	155	36.93	HLA-H, HCG4B, HLA-A, HCG9
31,005,000	31,105,000	100	19.59	HCG22, C6orf15, PSORS1C1, CDSN
31,155,000	31,360,000	205	45.62	HCG27, HLA-C, HLA-B
31,375,000	31,415,000	40	22.38	MICA
32,190,000	32,220,000	30	18.97	Downstream NOTCH4
32,665,000	32,775,000	110	37.40	HLA-DQA2, HLA-DQB2

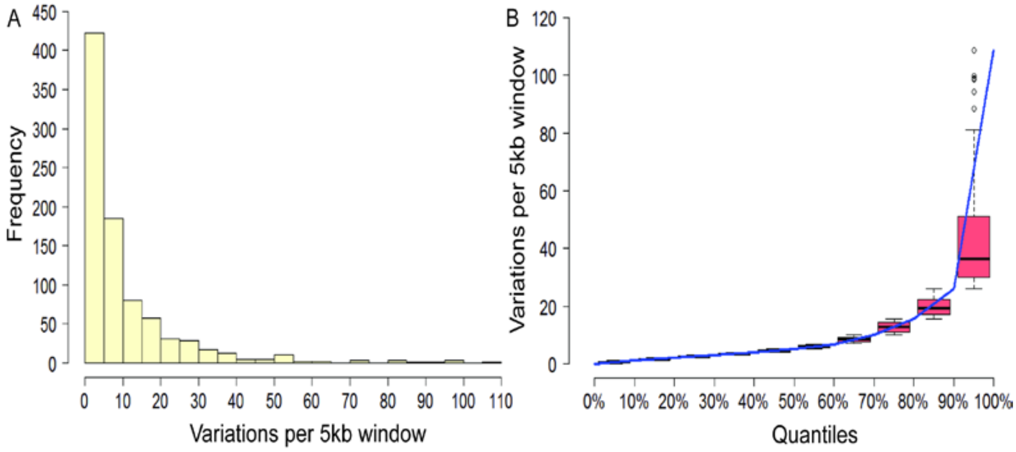


Figure 4.12 Distribution of inter-haplotypic variation. **(A)** Frequency histogram of the number of variations for each 5kb window. **(B)** Cumulative distribution for the number of variations per 5kb bin.

The RCCX region within chr6:31,939,646-32,077,151 is a common multi-allelic copy number variation locus. The number of modules and type of C4 complement genes within the RCCX region vary between individuals and the gene dosage of *C4A* and *C4B* have been associated with various disorders. For instance, lower levels of *C4A* have been associated with susceptibility for systemic lupus erythematosis [186], while lower levels of *C4B* have been associated with increased rates of acute myocardial infection and stroke [187]. To identify the number and type of RCCX modules associated with each of our haplotypes, we interrogated the RCCX region of each sample using a SYBR Green real-time PCR assay with primers specific for *C4A*, *C4B*, *C4L*, *C4S*, *TNXA*, and *RP1*. The total number of modules can be determined by three separate counts: (*C4A* + *C4B)/2*, (*C4L* + *C4S)/2*, and (*TNXA* + 2)/2.

In all samples, these counts gave a consistent total number of modules, thus showing internal validation of results. The A33-B58-DR3 haplotype was found to be monomodular, with one copy of *C4A* which was long. The A2-B46-DR9 haplotype was found to be bimodular, with 1 copy of *C4A* and 1 copy of *C4B*, one of which was long, and the other short (Table 4.11). We also report that APD, whose RCCX modular configuration is previously not determined, has 1 copy of *C4A* and 1 copy of *C4B*, both of which are long.

Table 4.11 RCCX modular structure in the 6 Singaporean Chinese cell lines and the APD cell line (European origin). The number of modules in each cell line was determined by the following counts: (*C4A* + *C4B)/2* or (*C4L* + *C4S)/2* or (*TNXA* + 2)/2. These 3 independent counts served as an internal verification for the number of RCCX modules carried by each cell line.

RCCX Structure	Sample	C4A	C4B	C4L	C4S	TNXA
Monomodular	B58SC	2	0	2	0	0
	B58AL	2	0	2	0	0
	B58CF	2	0	2	0	0
Bimodular	B46BM	2	2	2	2	2
	B46ZS	2	2	2	2	2
	B46CM	2	2	2	2	2
	APD	2	2	4	0	2

4.2.5 Conservation of HLA-DR Region between Asian and European Haplotypes

The availability of the Asian MHC haplotype sequences together with the sequences of the eight European haplotypes offers an excellent opportunity to study the MHC haplotypic relationship and provides insights into their recent evolutionary history. To do this, phylogenetic trees were derived from the SNP sequences of the MHC haplotypes and a total of four phylogenetic trees were built; representing the extended MHC region (29.65-33.0Mb), HLA-A region (27.0-30.2Mb), *HLA-B* region (31.1-31.6Mb) and HLA-DRB1 region (32.3-32.8Mb) (Figure 4.13A-D). The analysis showed that the trees were typically split into two main branches and the branching was not determined by the population ethnicity. In fact, the two Asian haplotypes were never found to form sister nodes with each other; instead, each Asian haplotype could be consistently found associating under the same clade with specific European haplotypes. For instance, the A33-B58-DR3 haplotype cell lines were found to be more closely related with the COX and QBL European haplotype at the HLA-DRB1 region than with the A2-B46-DR9 haplotype. This close relation of the Asian A33-B58-DR3 haplotype with COX and QBL is likely because these haplotypes carry the same HLA-DRB1*03:01 allele; the phylogenetic tree at the HLA-A and HLA-B region did not show such close association between the haplotypes. Likewise, the A2-B46-DR9 haplotype is more closely related to the DBB and MCF European haplotypes and this association is due to common possession of the HLA-A2 subtype allele. These analyses imply that the MHC haplotypic association is defined by the underlying HLA allelic typing rather than the population differences.

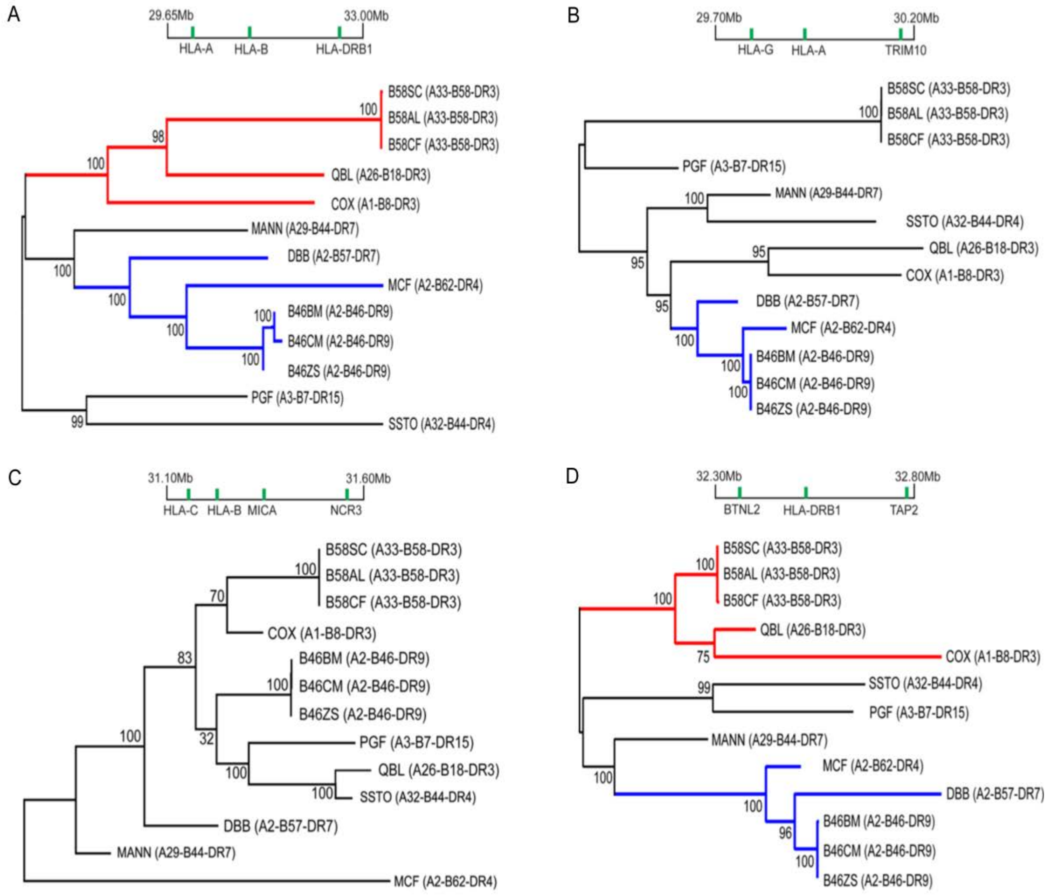


Figure 4.13 Phylogenetic relationships between Asian MHC haplotypes and European MHC haplotypes. Maximum likelihood (ML) tree derived from SNP sequence of the MHC haplotypes covering the **(A)** chr6:29.65 – 33.00Mb segment (18,781 SNPs), **(B)** chr6:29.70 – 30.20Mb segment (5,111 SNPs), **(C)** chr6:31.10 – 31.60Mb segment (3,617 SNPs) and **(D)** chr6:29.70 – 30.20Mb segment (5,237 SNPs)**.** The bootstrap value for each branch is indicated at the branching point. Generally most branches were able to achieve bootstrap values of >75 suggesting the reliability of the phylogenetic tree.

The COX and QBL cell lines were previously reported to have almost identical genomic sequences covering the *HLA-DRB1*, *-DQA1* and *DQB1* genes [15], and the phylogenetic analysis has demonstrated the close relationship between these haplotypes with the Asian A33-B58-DR3. To investigate the sequence relationship between the Asian and European genomes, we extracted 4441 consecutive SNPs derived from CG sequencing for the three cell lines carrying HLA-DRB1*03:01 allele; and compared with the COX and QBL nucleotide profiles. Given that the sequence length and genes composition are not entirely similar between the Human Reference Sequence Assembly 37.2 and the European haplotypes at the HLA-DR region, for this analysis, the COX sequence and genes annotation was used as reference. The selected 4441 SNPs spanned over 402,427bp of the COX sequence, encompassing the HLA-DRA, -DRB1, -DQB1, -DQA2, -DQB2 and -DOB genes. From the comparison, indeed, a 160kb segment enclosing the HLA-DR genes of the A33-B58-DR3 haplotype was almost identical to COX and QBL (Figure 4.14). Of the 1508 SNPs that fall within this 160kb segment, 1506 SNPs have nucleotide profiles that matched with COX and QBL, illustrating a remarkable conservation at the HLA-DR region among these haplotypes. The immediate centromeric end of this segment corresponds to a recombination hotspot providing strong evidence for haplotype break-up. Interestingly, the conservation region range between A33-B58-DR3 cell lines and QBL was even longer, extending up to almost 300kb. A smaller genomic segment of 58kb containing the HLA-DOB was detected to be almost identical to COX but not QBL. Again, this 58kb segment is flanked by recombination hotspots at both its telomeric and centromeric ends. These extreme conservations between the Asian and European haplotypes point to a shared recent common ancestor at the HLA-DR region.

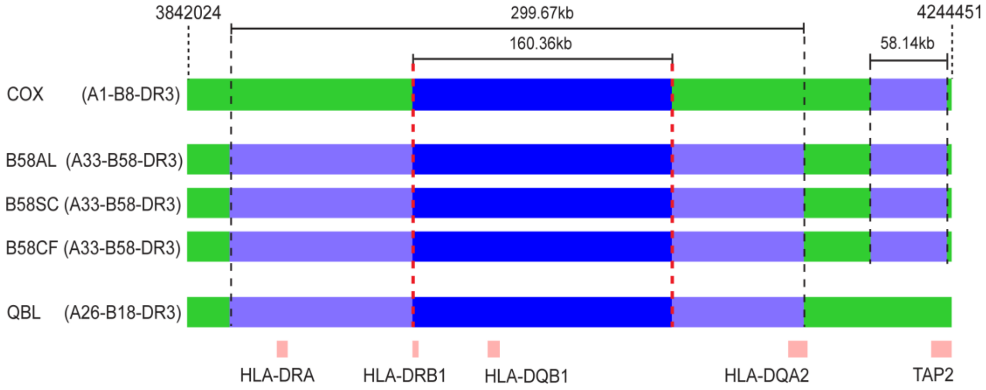


Figure 4.14 A33-B58-DR3 SNPs comparison with COX and QBL in the HLA-DRB region. COX sequence assembly and gene annotation (HSCHR6_MHC_COX_CTG1) was used as the reference for the comparison. The 4441 SNPs of B58AL, B58SC and B58CF, spanning over 402,427bp of the COX sequence, were included in this comparison. Dark blue bars indicate shared segment found in all samples while the light blue columns indicate shared segment found >2 samples but not in all samples.

4.3 Conclusion

In this study, with the use of HLA homozygous cell lines, we demonstrated extensive sequence conservation in two common MHC haplotypes of Asian ancestry - A33-B58-DR3 and A2-B46-DR9 by high throughput genome sequencing. We also for the first time described the extent of intra-haplotypic variation within the conserved boundaries of the MHC CEHs and revealed haplotype-specific novel variations. More significantly, we demonstrated that intra-haplotypic sequence variation in the cell lines carrying A2-B46-DR9 haplotype are associated with the expression of *ZFP57*; suggesting possible functional role in some of these polymorphic sites. Another major finding is that extreme sequence conservation extending up to 160kb at the HLA-DR region was found between the Asian A33-B58-DR3 haplotype and the European haplotypes (COX:A1-B8-DR3; QBL:A26-B18-DR3); implying individuals carrying these haplotypes shared a common ancestor. Overall, this approach has allowed us to assemble at least 90% phase-resolved MHC sequence representative of the A33-B58-DR3 and A2-B46-DR9 haplotype. The availability of these alternate Asian MHC sequences would complement the eight European MHC haplotype sequenced by the MHC Haplotype Project and provides a framework to study the MHC diversity and variations.

Chapter 5:

Transcriptome Landscape in MHC Conserved Extended Haplotypes

5.1 Introduction

Gene expression profile plays an important role in defining the phenotypic status in complex diseases. Description on gene transcripts, transcripts variability and isoform structure can provide insights on how differential gene expression leads to functional alteration that define phenotypic status. The MHC genomic region of approximate 4Mb has been associated with more than 100 diseases, including cancers, autoimmune diseases, infectious disease susceptibilities, neurodegenerative, cardiovascular, and metabolic disorders [6, 7]. Furthermore, many of these genetic associations are implicated by specific HLA haplotype marked by extensive LD most notably found in the common CEHs [13]. The ability to study the MHC transcription profile at haplotypic resolution can yield better understanding of the effects of HLA haplotypic differences on gene expression.

Currently, transcriptome characterization for the human MHC at haplotypic resolution has proven to be complicated. Firstly, the probes in the standard commercial expression microarray are annotated to the human genome reference sequence, thus are unable to account for the population MHC sequence and haplotype variation. The consequence of this is that individuals with different HLA haplotypic background and sequence profile from the reference sequence may not display expression of certain genes because the probes are unable to anneal to their unique gene sequences [188, 189]. This could distort the evaluation of gene expression and lead to erroneous conclusions. Secondly, majority of the individuals carry two distinct HLA haplotypes; hence in such circumstances, it is difficult to ascertain the haplotypic origin of a particular RNA transcript and complicates the analysis for association between HLA haplotypes and gene expression. To overcome these limitations, Vandiedonck and colleagues [179] used a hybrid microarray that includes alternate allele probes to account for known variation in gene sequences. These alternate allele probes were designed based on the annotated sequence variation (SNPs) and known segmental duplication in the MHC region. The customized MHC array was then applied to the PGF, COX and QBL MHC-homozygous LCLs of European-descent and the analysis revealed extensive haplotype-related transcriptional differences. Despite the impressive efforts to account for the sequence diversity in the MHC region, the customized MHC array could not have comprehensively covered all possible MHC variation across populations and is limited in revealing novel transcripts and splicing isoforms. More importantly, in the study, each MHC haplotype was represented by a single cell line; hence there is a possibility that the observed transcriptional differences might be attributed to the inherent cell lines variation but not the haplotypic variation.

In this chapter, we adopted the RNA-seq approach to interrogate the MHC transcription landscape of Asian CEH (A33-B58-DR3 and A2-B46-DR9) using multiple HLA homozygous LCLs for each of the haplotypes. The RNA-seq approach can allow us to annotate and quantification of all expressed transcripts at high level of sensitivity and accuracy [190, 191], accounting for the limitations in microarray-based methods. Here, we aim to assess MHC haplotype-related expression difference in the Asian CEHs as well as to perform a comparative transcriptomic analysis of the Asian and European CEHs.

5.2 Results

5.2.1 RNA-seq Experimental Design

RNA-seq datasets were generated from purified total RNA isolated from six selected B-LCLs – COX, QBL, B58AL, B58SC, B46BM and B46ZS. Prior to the isolation of RNAs, the cell lines were cultured independently in duplicates to five passages. The cell lines were then stimulated with 200nm PMA and 125nM ionomycin for six hours and harvested at approximate 1 X 10^6 cells per ml; the supernatant was then used for the ELISA experiement. DMSO were added to unstimulated cell cultures to act as controls. To ensure, all the cell lines were sufficiently stimulated and the replicates displayed similar profiles, ELISA was performed to quantify the levels of TNF-alpha and IL6 proteins in both unstimulated and stimulated cultures. TNF-alpha and IL6 were previously reported to be up-regulated in B-LCLs [192, 193] and hence were selected for the ELISA experiments. The ELISA experiments showed a clear increase in production of TNF-alpha and IL6 in stimulated samples when compared to the control samples; with an exception for B46ZS cell line where there was no or minimal production of IL6 in both stimulated and unstimulated culture (Figure 5.1). In addition, consistent proliferation patterns of TNF-alpha and IL6 were observed in both replicates across all the cell lines. Total RNA were then extracted from the cell pellets collected after stimulation and cytoplasmic ribosomal RNA were removed from the DNase-treated total RNA. The resulting RNA was then used to prepare RNA-seq libraries consisting individually-barcoded RNA fragments and these fragments were then sequenced using the Illumina Hi-Seq 2000 sequencing machine.

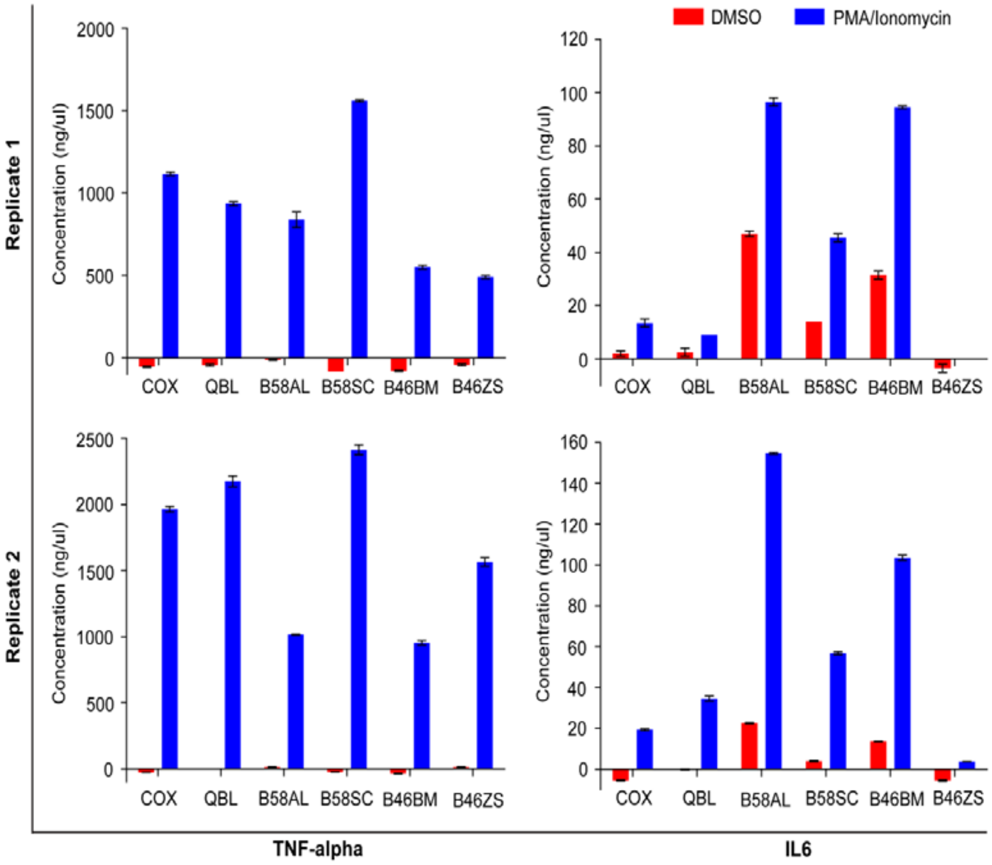


Figure 5.1 PMA and ionomycin stimulation determined by ELISA. Stimulation was performed for six hours and the supernatant was then used for the ELISA. DMSO was added in amount equal to the PMA and ionomycin into unstimulated cell cultures to act as controls. The ELISA was performed in triplicate for each B-LCL.

5.2.2 RNA-seq Data Processing, Reads Filtering and Mapping Results

We had outlined a systematic approach consisting of six phases to effectively handle and accurately infer biological implications from the large amount data generated by sequencing of the entire human transcriptome (Figure 5.2). The RNA-seq analysis in this study would follow closely to the workflow of this strategy.

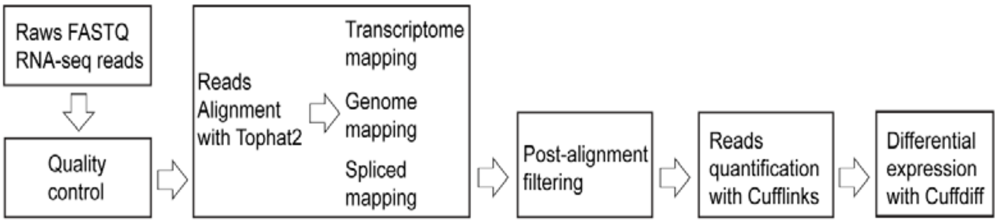


Figure 5.2 Schematic for RNA-seq workflow

For each sample, we obtained 78,890,079 ± 7,103,330 (mean ± standard deviation) 75bp paired-end reads from the RNA-sequencing (Table 5.1). Because of the inherent nature of the NGS platforms, there exists a sizeable number of sequencing errors and sequence biases in the raw reads [194]. Therefore, it is essential to perform quality control to filter out reads of low quality to avoid spurious reads alignment resulting in erroneous downstream analysis. Here, only reads with 70% of the base positions meet the Phred score cuff-off of 20 were retained for the downstream analysis. Phred score is an indication for the quality of the reads; a base position with a Phred score of 20 implies that there is a 1 out of 100 chances that this position is called incorrectly [195, 196]. Indeed, the quality of the sequenced reads improved after the quality control filtering; the percentage of sequence reads across all base positions having a Phred score range of 31-40 significantly increased by more than 10% while positions with Phred score of range 1-10 reduced to less than 5% after filtering (Figure 5.3) and more than 85% of the original raw reads were retained for further analysis (Table 5.1). In terms of the per base sequence content of the reads, equal A, T, C and G nucleotide compositions were observed at read positions greater than 10 and there was a distinctive nucleotide composition variation in the first nine positions of the reads across all sequenced samples (Figure 5.4). This distinct nucleotide composition variation is due to the use of the random hexamer primers during the synthesis of the double-stranded complementary DNA in the library preparation step [197] and this bias will be corrected and account for at the transcripts abundance quantification step.

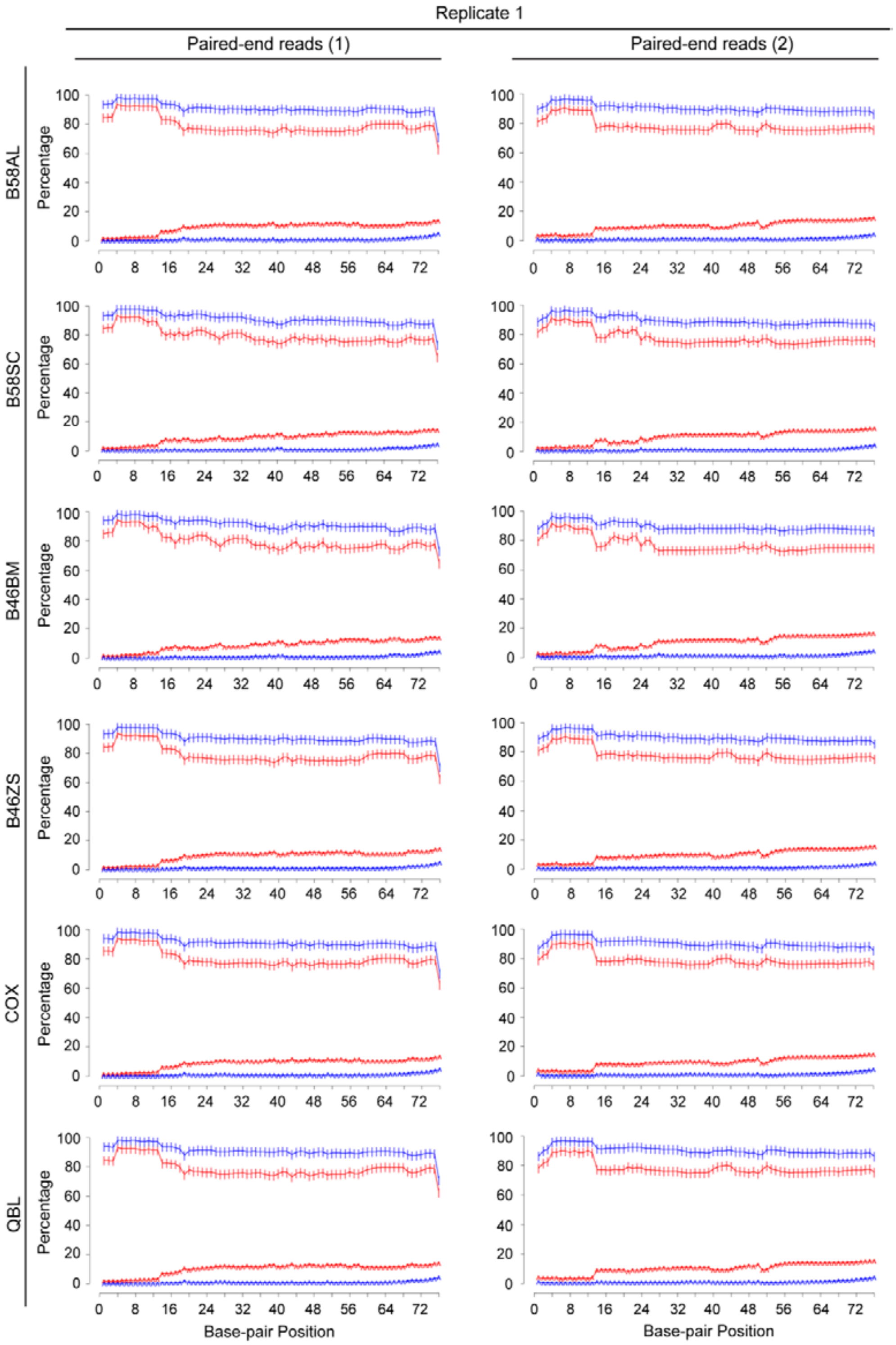


Figure 5.3 Paired-end reads base positions score before and after quality control filtering (Replicate 1). Red indicates base positions score before quality control while blue indicates score after quality control. "|" indicates percentage of reads with a Phred score range of 31-40 and "^" indicates percentage of reads with a Phred score range of 1-10.

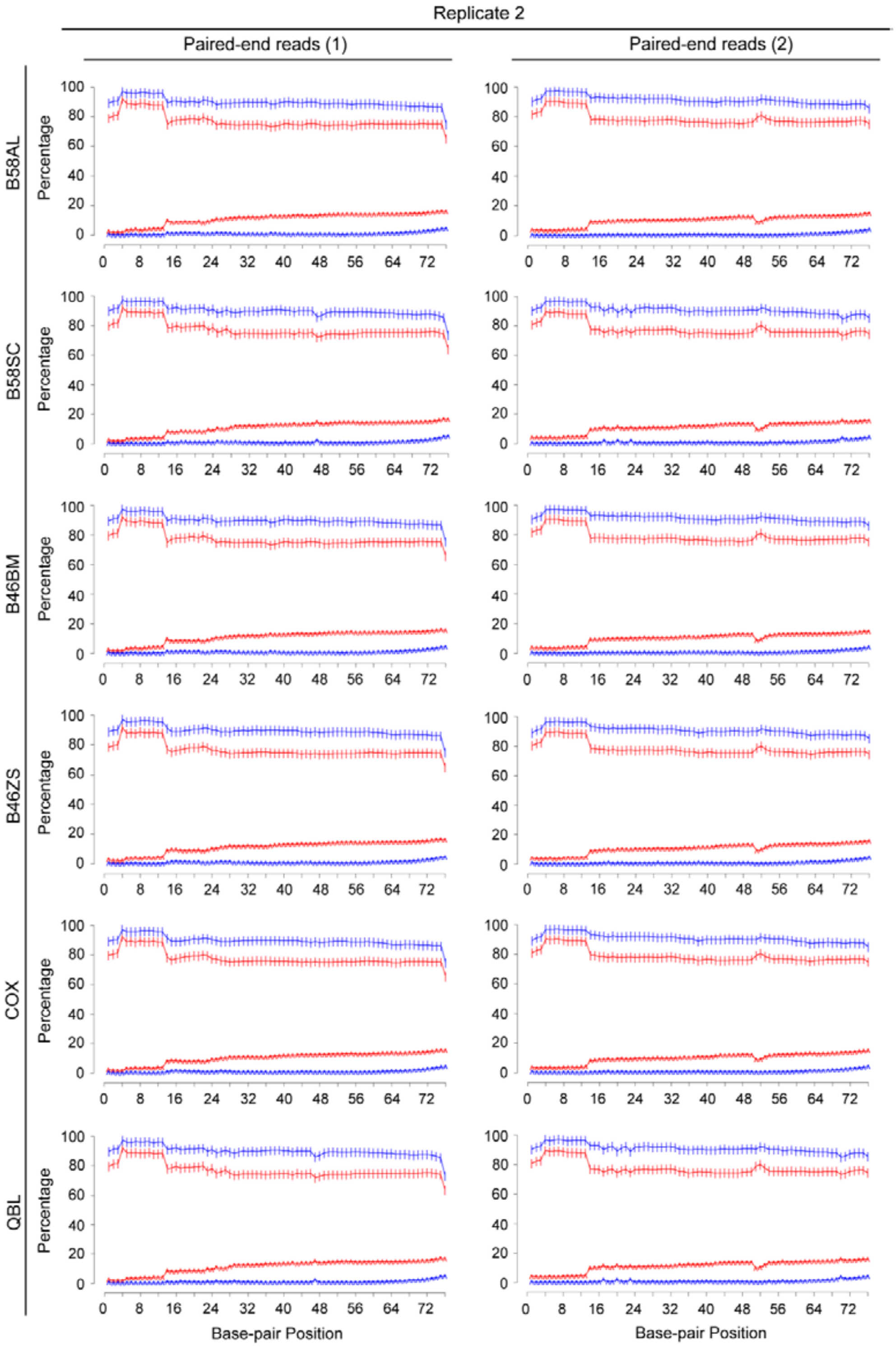


Figure 5.3 Paired-end reads base positions score before and after quality control filtering (Replicate 2). Red indicates base positions score before quality control while blue indicates score after quality control. "|" indicates percentage of reads with a Phred score range of 31-40 and "^" indicates percentage of reads with a Phred score range of 1-10.

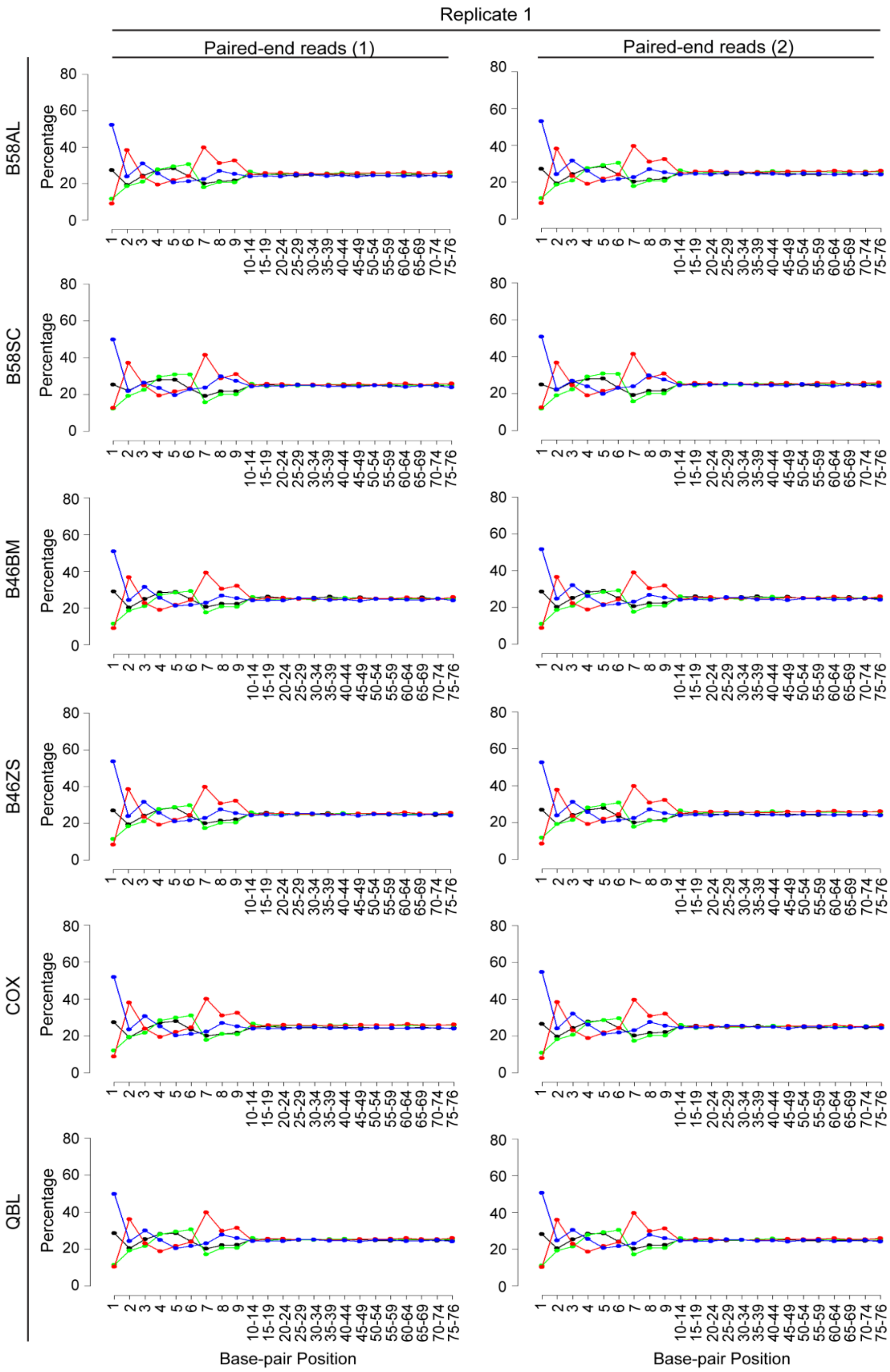


Figure 5.4 Nucleotide composition across all base positions in the sequenced reads (Replicate 1). Green indicates composition of “A” nucleotide, red indicates composition of “T” nucleotide, blue indicates composition of “C” nucleotide and black indicates composition of “G” nucleotide

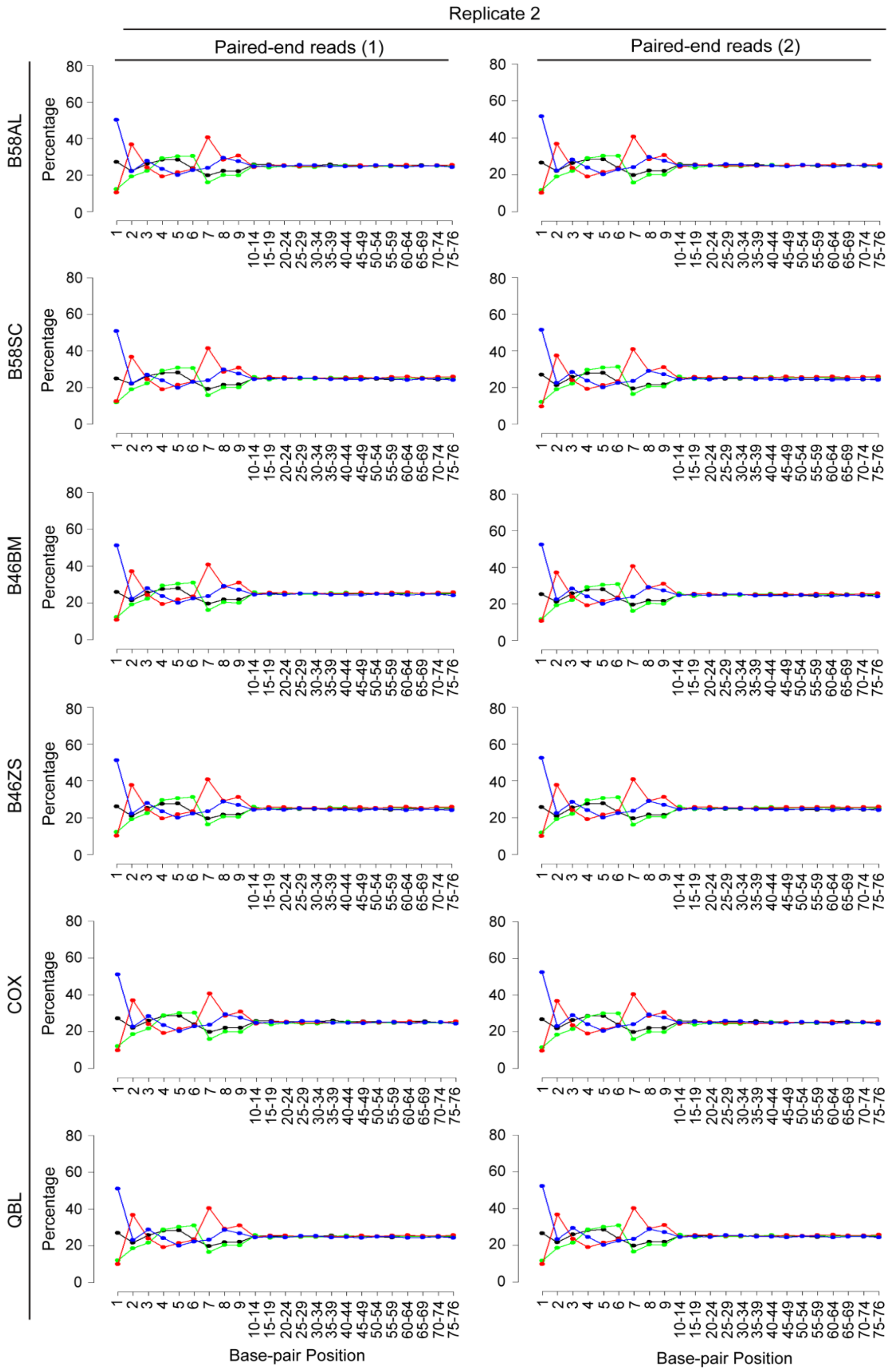


Figure 5.4 Nucleotide composition across all base positions in the sequenced reads (Replicate 2). Green indicates composition of “A” nucleotide, red indicates composition of “T” nucleotide, blue indicates composition of “C” nucleotide and black indicates composition of “G” nucleotide.

Table 5.1 Quality of raw reads and mapping metrics.

Replicate 1						
Cell line	Raw reads	QC reads	Reads mapped	% of reads mapped	Uniquely-mapped reads	% uniquely-mapped reads
B58AL	80,362,945	69,588,080	57,790,683	83.05	45,42,9741	78.61
B58SC	75,778,660	65,124,362	55,602,319	85.38	41,975,361	75.49
B46BM	67,751,606	57,936,456	48,472,507	83.66	36,299,920	74.89
B46ZS	73,156,838	63,331,245	53,152,992	83.93	42,867,424	80.65
COX	68,449,949	59,742,216	48,925,551	81.89	37,521,696	76.69
QBL	86,358,133	74,445,714	63,410,379	85.18	48,752,536	76.88

Replicate 2						
Cell line	Raw reads	QC reads	Reads mapped	% of reads mapped	Uniquely-mapped reads	% uniquely-mapped reads
B58AL	77,249,247	66,485,079	53,928,117	81.11	41,934,497	77.76
B58SC	86,174,592	74,038,168	59,540,848	80.42	48,681,488	81.76
B46BM	81,895,555	70,639,184	57,584,590	81.52	44,872,233	77.92
B46ZS	91,114,107	77,958,277	63,170,151	81.03	50,268,589	79.58
COX	81,075,411	70,199,584	58,532,731	83.38	43,702,252	74.66
QBL	77,313,902	66,249,897	53,906,838	81.37	41,783,599	77.51

The QC reads were then mapped to the human reference transcriptome (NCBI gene annotation Build 37.2) as well as to the human reference genome (NCBI Build 37.2 reference sequence) using Tophat2 [156]. To ensure accurate and high quality reads alignment, post-alignment filtering was performed to remove reads that were not mapped in proper pairs – paired-end reads that were incorrectly oriented with respect to each other and to filter aligned reads with template length that were deviate significantly from the expected template length. On a whole, at least 80% of the input reads were aligned to the human genome across all cell lines and their replicates, of which at least 74% were uniquely mapped to a single location of the genome.

Direct evaluation of the gene expression level from the mapped reads is difficult due to the variation in the number of reads generated from each independent sequencing runs and also the sequence biases introduced during the library preparation step. To account for these differences, normalization procedures are required in order to accurately quantify genes expression level. Here, we used the fragments per kilobase of exon per million mapped reads (FPKM) implemented in Cufflinks [157] to normalize and quantitate the relative gene expression from the assembled reads with the NCBI gene annotation Build 37.2 . Subsequently, genes expression (log2-transformed FPKM values) between the two biological replicates of each cell line were compared and evaluated using the square of pearson's correlation coefficient metric (R^2). High level of concordance in the genes log2-transformed FPKM values between the replicates ($R^2 = 0.96$) were observed across all the cell lines (Figure 5.5); indicating excellent reproducibility and minimal experimental errors during sample preparation.

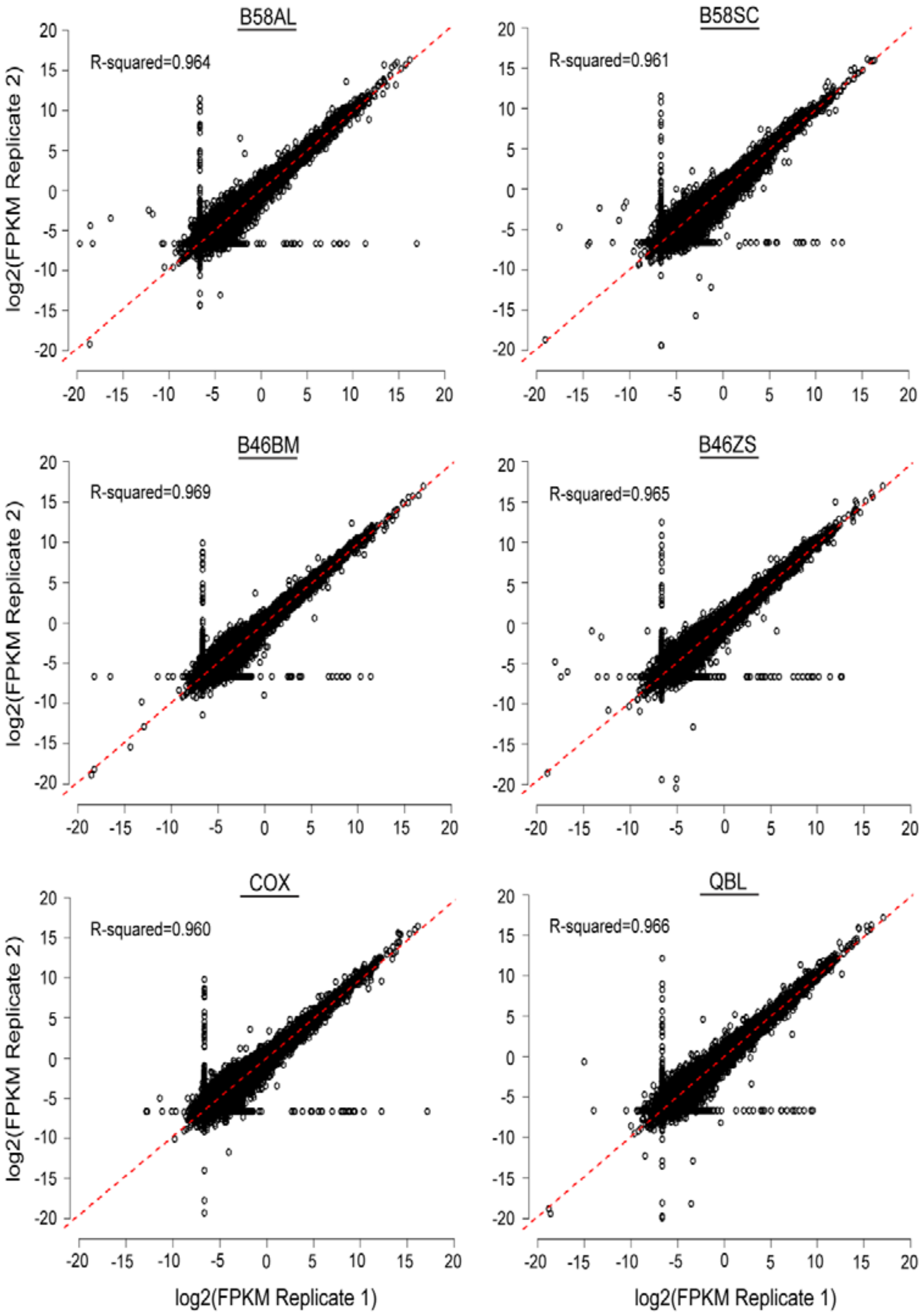
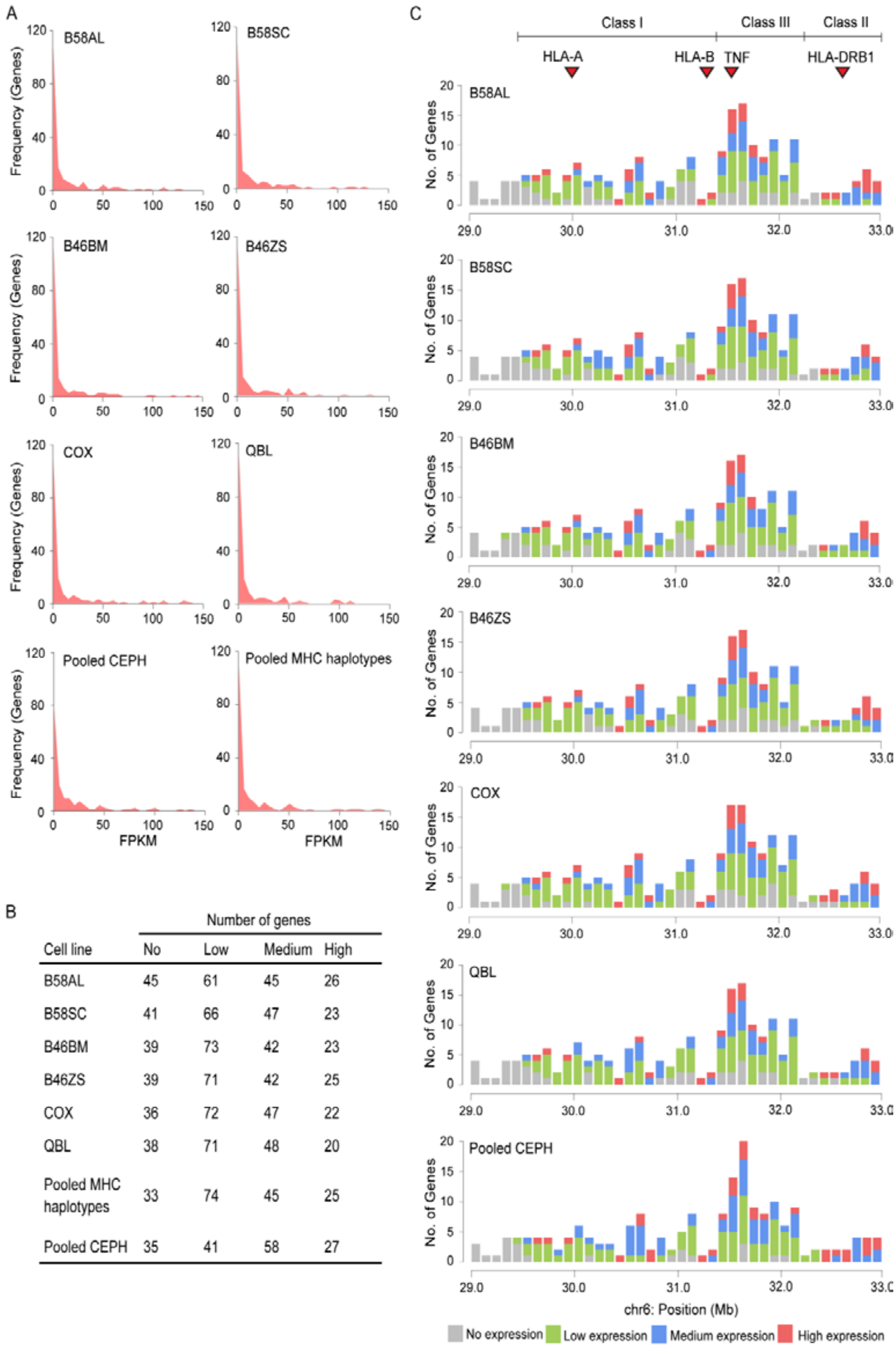


Figure 5.5 Gene expression comparison between cell line replicates. The expression of 25549 NCBI annotated loci in the entire genome was considered. High correlation in the genome-wide expression profiles between the replicates was observed for all the cell lines.

5.2.3 Landscape of Transcription in the MHC Region

The availability of this RNA-seq dataset allowed us to comprehensively analyze the transcription landscape of the MHC region in the BLCLs. There are 177 genes and pseudogenes annotated in the NCBI gene annotation Build 37.2 across the chr6:29.0 – 33.0Mb region and the calculated FPKM values of these 177 annotated genes were examined. The distribution of the FPKM values was skewed to the left (Figure 5.6A) and the median FPKM values range from 1.60 to 2.06 across all the six cell lines with more than 85% of the annotated genes having FPKM value less than 50. P-values (>0.05) derived from Kolmogorov-Smirnov test indicated there was no difference in the FPKM distribution across the cell lines. In addition, the FPKM distribution agrees with the RNA-seq dataset derived from the 20 unrelated individuals BLCLs in Centre d'Etude du Polymorphisme Humain (CEPH) collection [198] showing the reliability of our dataset. The slight aberration in the gene counts across FPKM bins observed between the two datasets are likely due to inherent cell lines divergence as well as the variation in the handling and treatment of the cell lines. Next we will like to determine whether a gene is expressed; evaluation of genes with FPKM value greater than zero will comprise of genes with FPKM values very close to zero and these are likely background noise due the erroneous mapping which cannot be completely eradicated. Hence, we set a FPKM value cutoff of 0.01 which is the 5th percentile of the genome-wide genes FPKM values across all the cell lines in our subsequent analysis. Based on the FPKM value, the level of gene expressions were categorized into no expression (FPKM<=0.01), low expression (0.01<FPKM<=5), medium expression (5<FPKM<=50) and high expression (FPKM>50). The number of genes expressed within the MHC region in each cell line agreed well with every other cell lines and the proportion of low, medium and high expression genes were also consistent across all the cell lines (Figure 5.6B). Likewise, the proportion on the different levels of gene expression in the CEPH dataset is comparable to the MHC haplotype cell lines.

The expression landscape across the MHC region (chr6: 29.0 – 33.0 Mb) was explored by examining the gene expression profiles within 100kb bin windows and the level of gene expression in each 100kb bin did not varied much across different cell lines (Figure 5.6C). Half of the non-expressing genes were found at the olfactory cluster (chr6:29.0- 29.60 Mb) which harbor numerous olfactory receptor genes while high level of transcription activities were observed at the Class II region. A closer look at the gene expression profiles based on the functional role revealed that genes coding for proteins involved in the antigen processing and presentation were highly expressed as compared to those involved in the stress response and regulation (Table 5.2). Indeed, the *HLA-A*, *HLA-B* and *HLA-DRA* were one of the most highly expressed genes in the MHC region. These suggest genes that are related by their functional role are expressed at similar levels.



	Number of genes			
Cell line	No	Low	Medium	High
B58AL	45	61	45	26
B58SC	41	66	47	23
B46BM	39	73	42	23
B46ZS	39	71	42	25
COX	36	72	47	22
QBL	38	71	48	20
Pooled MHC haplotypes	33	74	45	25
Pooled CEPH	35	41	58	27

Figure 5.6 Gene expression profiles within MHC region **(A)** Distribution of FPKM values for genes within the MHC region – chr6:29.0-33.0Mb **(B)** Number of no (FPKM<=0.01), low (0.01<FPKM<=5), medium (5<FPKM<=50) and high (FPKM>50) expressing genes in each cell line. **(C)** Landscape of gene expression within 100kb bin windows.

Table 5.2 Gene expression (FPKM) categorized by their functional role.

Function	Genes	B58AL	B58SC	B46BM	B46ZS	COX	QBL
Antigen presentation	HLA-A	469.31	327.63	439.58	449.52	344.78	254.70
	HLA-E	355.44	275.48	317.32	277.10	319.50	256.64
	HLA-C	236.47	192.29	164.48	163.14	527.82	96.76
	HLA-B	945.18	793.61	541.92	591.96	839.96	494.92
	HLA-DRA	1030.4	1286.1	918.9	856.2	1063.9	1119.8
	HLA-DRB1	250.63	223.30	39.68	38.42	130.13	154.44
	HLA-DQA1	21.74	25.91	2.07	2.54	7.44	6.66
	HLA-DQB1	5.95	5.49	3.21	4.11	2.77	3.23
	HLA-DOB	6.80	6.76	6.04	3.05	7.98	8.94
	HLA-DMB	60.06	23.60	49.96	51.26	35.37	60.22
	HLA-DMA	68.79	39.26	60.19	43.38	64.09	59.76
Antigen processing	TAP2	55.99	48.14	46.44	53.04	47.72	43.79
	PSMB8	111.04	74.61	148.38	130.73	93.79	103.87
	PSMB9	60.03	40.63	65.45	58.893	43.79	44.32
	TAP1	114.37	128.18	123.79	165.36	114.48	98.66
Inflammation	NFKBIL1	8.63	8.20	8.71	6.31	8.03	7.99
	LTA	211.32	185.85	189.19	198.41	210.57	245.51
	TNF	51.71	59.74	51.52	64.80	94.25	66.10
	LTB	8.58	3.19	6.12	14.54	10.90	21.51
	LST1	2.69	3.79	0.83	1.24	1.13	3.78
	NCR3	2.35	1.71	1.62	1.93	2.39	1.86
	AIF1	1.98	13.56	3.13	7.59	2.05	11.29
Stress response	HSPA1L	1.39	1.90	1.60	1.16	0.87	1.44
	HSPA1A	10.13	17.26	10.46	7.82	5.88	10.76
	HSPA1B	11.61	16.69	13.34	14.24	8.29	8.27
	MICA	3.56	3.47	6.48	8.77	6.49	5.25
	MICB	10.60	8.12	8.78	9.07	7.70	6.70
Regulatory receptors	AGER	1.21	1.60	1.56	1.04	1.24	0.80
	NOTCH4	0.15	0.18	0.14	0.12	0.10	0.11

Red = high expression (FPKM>50); blue = medium expression (5<FPKM<=50); green = low expression (0.01<FPKM<=5)

Subsequently, the alternative splicing activities in the MHC region were assessed. Of the 177 genes analyzed in the region, 51 genes were expressed with two or more isoforms and the relative abundance of each of the alternatively spliced gene transcripts were examined. The transcript with the highest FPKM value was regarded as the major isoform while all other isoforms were regarded as minor isoform and the proportion of the major isoform FPKM value to the aggregated FPKM value of all expressing transcripts of each gene was calculated across all the cell lines (Figure 5.7). From this analysis, it was noted that the isoforms were not expressed at equivalent level and in fact all the 51 genes had one dominant expressing isoform with the exception of *TAP2* where both of its isoforms were evenly expressed. In addition, there were 15 genes (constituting approximate 30% of the genes with multiple isoforms) where one or more cell lines do not share a common dominant expressing isoform.

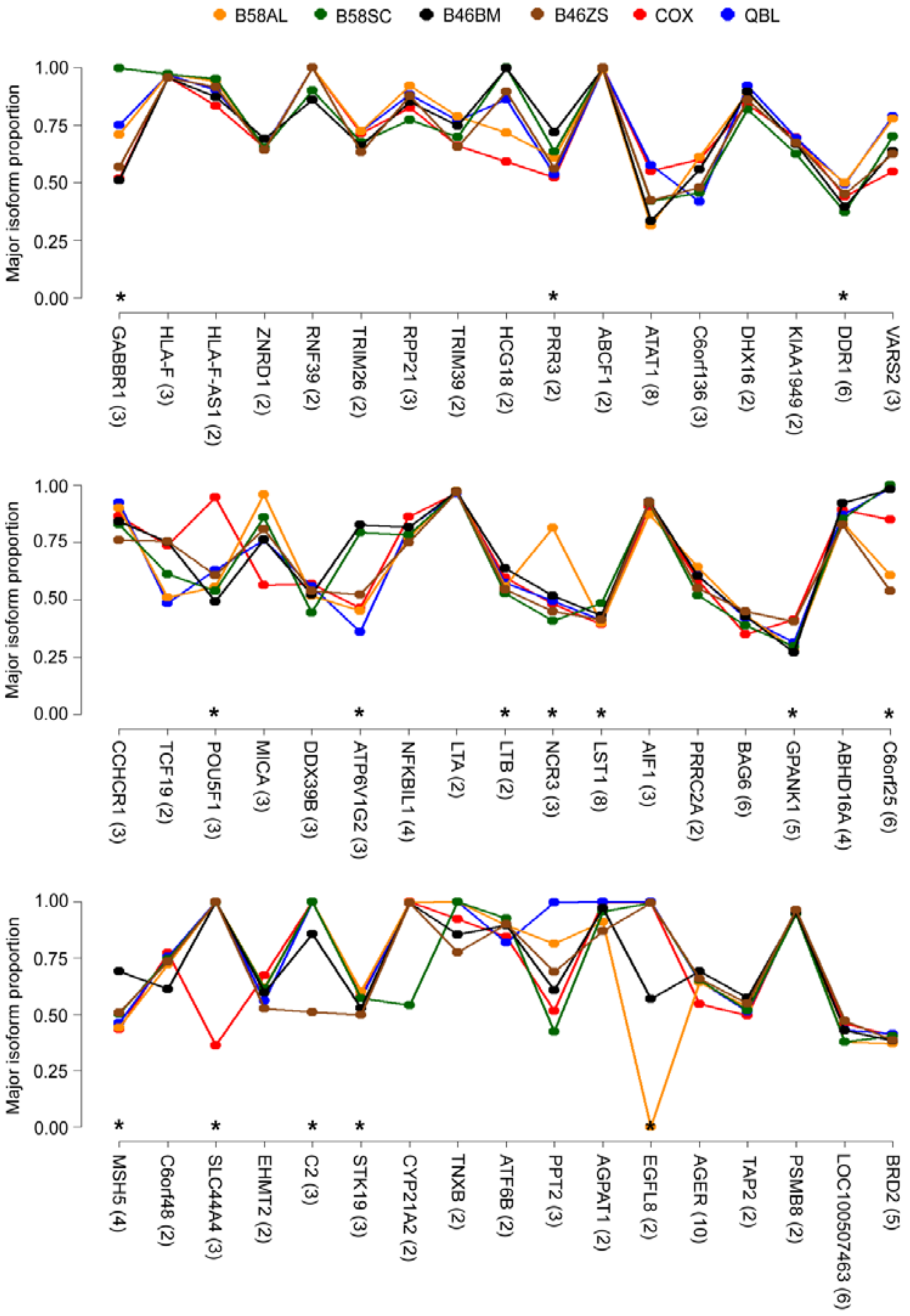


Figure 5.7 Major isoform proportion relative to all expressing isoforms of a gene across the six cell lines. Transcript with the highest FPKM value was considered as the major isoform. The "*" indicates that one or more cell lines do not shared identical major isoform for a particular gene. Bracket number indicates the number of known isoforms annotated in NCBI gene annotation Build 37.2.

5.2.4 MHC Haplotype-specific Gene Expression

The access of multiple HLA homozygous cell lines carrying identical MHC haplotype that displayed CEHs characteristics has offered a unique opportunity to investigate on haplotype-specific expression. In the previous chapter, we have showed that independent cell lines carrying identical MHC CEHs display high sequence invariant in the MHC region. Therefore it would be of interest to examine whether this sequence similarity would correspond to similar gene expression profiles among multiple cell lines carrying identical MHC haplotype (B58AL and B5SC – A33-B58-DR3; B46BM and B46ZS – A2-B46-DR9). To do this, principal component analysis (PCA) was applied to the expression data of the 177 MHC region genes represented by their FPKM values. The PCA results showed haplotype-specific clustering while surprising the QBL cell line had highly similar expression profile with the cell lines carrying A2-B46-DR9 haplotype (Figure 5.8). This data suggests that cell lines carrying identical Asian MHC CEH have probable correlation in their gene expression profile at the MHC region. Next, quantitative differences in gene expression levels between the MHC haplotypes were assessed by the grouping cell lines with identical MHC CEHs. Pairwise fold-change comparison of every possible MHC haplotype combinations was performed to provide a regional view of the transcription differences between the MHC haplotypes (Figure 5.9). Heighten transcriptional variation between haplotypes were found to be localized to regions proximate to the HLA genes as well as the genes cluster involved in inflammation, highlighting the influence of haplotype on the expression of immune related genes.

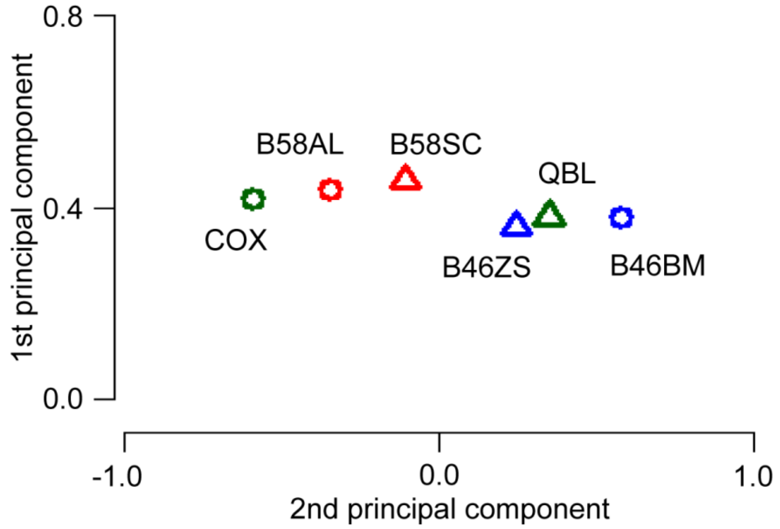


Figure 5.8 PCA analysis based on the relative abundance (expression) of 177 MHC region genes. Red indicates cell lines carrying A33-B58-DR3 MHC haplotype; blue indicates cell lines carrying A2-B46-DR9 MHC haplotype and green indicates cell lines of European origins.

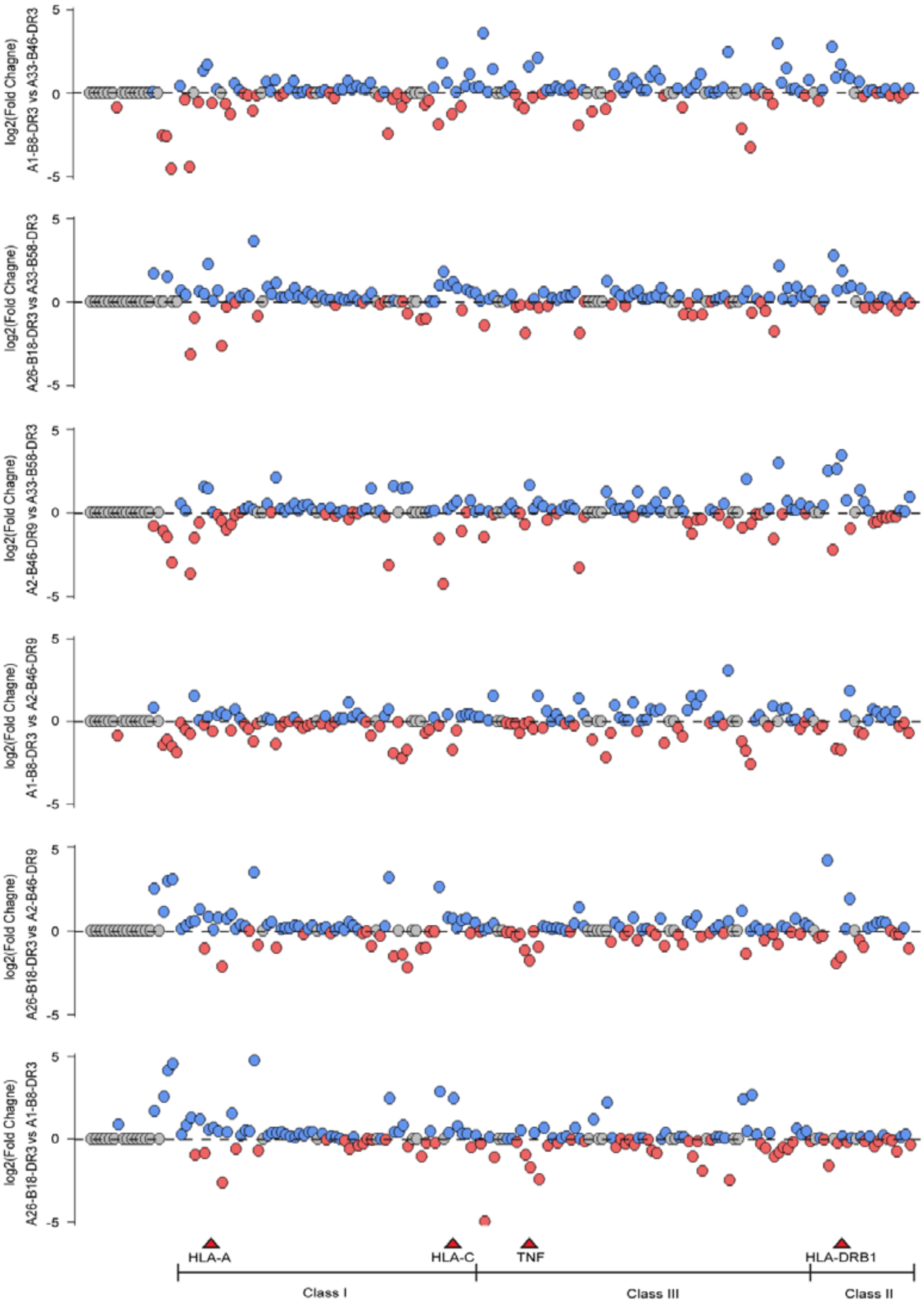


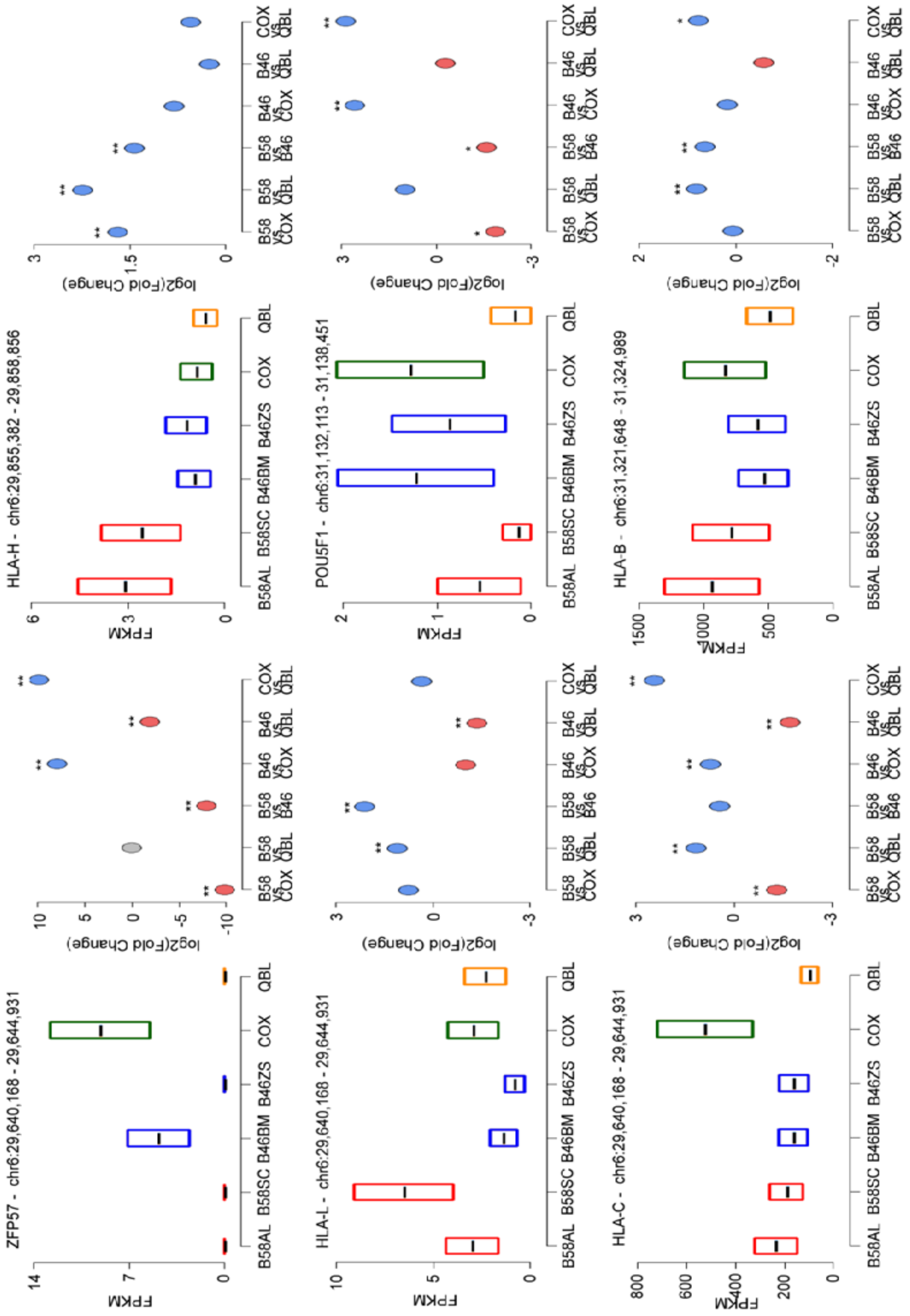
Figure 5.9 Pairwise comparisons of A33-B58-DR3 vs A1-B8-DR3 (COX), A33-B58-DR3 vs A26-B18-DR3 (QBL), A33-B58-DR3 vs A2-B46-DR9, A2-B46-DR9 vs A1-B8-DR3, A2-B46-DR9 vs A26-B18-DR3 and A1-B8-DR3 vs A26-B18-DR3. Fold change (FC) derived from the log2 ratio of gene FPKM values between two haplotypes. Blue circle indicates higher expression in the top haplotype than in bottom haplotype while red indicates otherwise.

Table 5.3 Variation of gene expression between MHC haplotypes.

	Log2(Fold Change)					
Gene	A33-B58-DR3 vs A1-B8-DR3	A33-B58-DR3 vs A26-B18-DR3	A33-B58-DR3 vs A2-B46-DR9	A2-B46-DR9 vs A26-B18-DR3	A2-B46-DR9 vs A1-B8-DR3	A1-B8-DR3 vs A26-B18-DR3
UBD	0.02 ^{NS}	1.70 **	-0.80 *	2.50 **	0.83 *	1.68 **
GABBR1	-2.61 **	1.50 ^{NS}	-1.47 ^{NS}	2.97 ^{NS}	-1.14 ^{NS}	4.11 **
ZFP57	-9.84 **	0.00 ^{NS}	-7.93 **	7.93 **	-1.91 **	9.84 **
HLA-H	1.00 **	2.23 **	1.43 **	0.80 ^{NS}	0.26 ^{NS}	0.54 ^{NS}
HLA-A	0.21 ^{NS}	0.65 *	-0.13 ^{NS}	0.78 **	0.34 ^{NS}	0.44 ^{NS}
HLA-L	0.75 ^{NS}	1.10 **	2.11 **	-1.01 ^{NS}	-1.36 **	0.35 ^{NS}
DDR1	0.59 ^{NS}	0.52 ^{NS}	1.44 **	-0.92 *	-0.85 ^{NS}	-0.07 ^{NS}
DPCR1	-0.37 ^{NS}	0.06 ^{NS}	1.59 **	-1.54 *	-1.96 **	0.42 ^{NS}
HCG22	-0.26 ^{NS}	-0.73 ^{NS}	1.47 **	-2.19 **	-1.73 **	-0.46 ^{NS}
POU5F1	-1.89 *	0.98 ^{NS}	-1.60 *	2.59 **	-0.28 ^{NS}	2.87 **
PSORS1C3	1.79 ^{NS}	1.79 ^{NS}	-4.32 ^{NS}	6.10 **	6.10 **	0.00 ^{NS}
HLA-C	-1.29 **	1.15 **	0.42 ^{NS}	0.73 **	-1.72 **	2.45 **
HLA-B	0.05 ^{NS}	0.82 **	0.64 **	0.17 ^{NS}	-0.59 ^{NS}	0.70 *
MICA	-0.82 ^{NS}	-0.51 ^{NS}	-1.14 **	0.63 ^{NS}	0.32 ^{NS}	0.30 ^{NS}
TNF	-0.73 **	-0.22 ^{NS}	-0.03 ^{NS}	-0.19 ^{NS}	-0.70 *	0.51 ^{NS}
LTB	-0.92 ^{NS}	-1.90 **	-0.74 ^{NS}	-1.16 **	-0.18 ^{NS}	-0.98 ^{NS}
AIF1	2.09 **	-0.37 ^{NS}	0.59 ^{NS}	-0.95 ^{NS}	1.51 ^{NS}	-2.46 **
HSPA1A	1.25 **	0.38 ^{NS}	0.62 ^{NS}	-0.24 ^{NS}	0.63 ^{NS}	-0.87 ^{NS}
C6orf48	-0.16 ^{NS}	0.21 ^{NS}	1.15 **	-0.93 **	-1.31 **	0.38 ^{NS}
NEU1	0.25 ^{NS}	0.36 ^{NS}	0.64 **	-0.27 ^{NS}	-0.39 ^{NS}	0.11 ^{NS}
HLA-DRB5	8.33 **	6.68 ^{NS}	2.51 **	4.18 ^{NS}	5.82 **	-1.64 ^{NS}
HLA-DRB1	0.90 **	0.65 *	2.59 **	-1.95 **	-1.70 **	-0.25 ^{NS}
HLA-DQA1	1.67 **	1.83 **	3.40 **	-1.57 **	-1.73 **	0.16 ^{NS}
HLA-DQA2	0.88 ^{NS}	0.90 ^{NS}	-0.96 **	1.87 **	1.84 **	0.03 ^{NS}
HLA-DQB2	0.66 ^{NS}	0.74 ^{NS}	1.31 **	-0.57 ^{NS}	-0.65 ^{NS}	0.09 ^{NS}
HLA-DOA	0.24 ^{NS}	-0.12 ^{NS}	0.93 **	-1.06 **	-0.69 ^{NS}	-0.36 ^{NS}

Genes showing significant differentially expressed in one or more haplotype-pairs comparison after Benjamini-Hochberg adjustment. “NS” denotes not significant haplotype-pair; “**” denotes haplotype-pair with adjusted P-value <0.05; “*” denotes haplotype-pair with adjusted P-value <0.1. COX cell line carries the A1-B8-DR3 haplotype while QBL cell line carries the A26-B18-DR3 haplotype.

To detect differentially expressed genes, the method implemented in Cuffdiff 2 was employed [157]. Overall, 26 genes were significantly differentially expressed (adjusted P-value<0.05) in at least one pair of haplotype comparison (Table 5.3). A number of these genes that are related to the antigen presentation (*HLA-A*, *HLA-C*, *HLA-B*, *HLA-DRB1*, *HLA-DQA1*, *HLA-DQA2*, *HLA-DQB2*) as well as the inflammatory genes (*TNF*, *LTB*, *AIF1*). As the variation in the gene expression levels could be due to individual cell line differences rather than haplotype-specific differences, we selected 12 genes that were found to be significantly differentially expressed in three or more haplotype-pairs comparison and inspect on the expression levels of these genes in each of the six individual cell lines. Indeed, with the exception of *ZFP57*, equivalent expression levels were observed in cell lines sharing identical MHC haplotype for the 11 selected genes (Figure 5.10). These provide strong evidence for the influence of haplotypic sequence variation on transcription activity in the MHC region.


ZFP57 - chr6:29,640,168 - 29,644,931
HLA-L - chr6:29,640,168 - 29,644,931
HLA-C - chr6:29,640,168 - 29,644,931
HLA-H - chr6:29,855,382 - 29,858,856
POU5F1 - chr6:31,132,113 - 31,138,451
HLA-B - chr6:31,321,648 - 31,324,989
FPKM
log2(Fold Change)
14
7
0
10
5
0
-5
-10
3
0
-3
800
600
400
200
6
1.5
2
1
1500
1000
500
-2
B58AL
B58SC
B46BM
B46ZS
COX
QBL
B58 vs COX
B58 vs QBL
B58 vs B46
B46 vs COX
B46 vs QBL
COX vs QBL
**
*

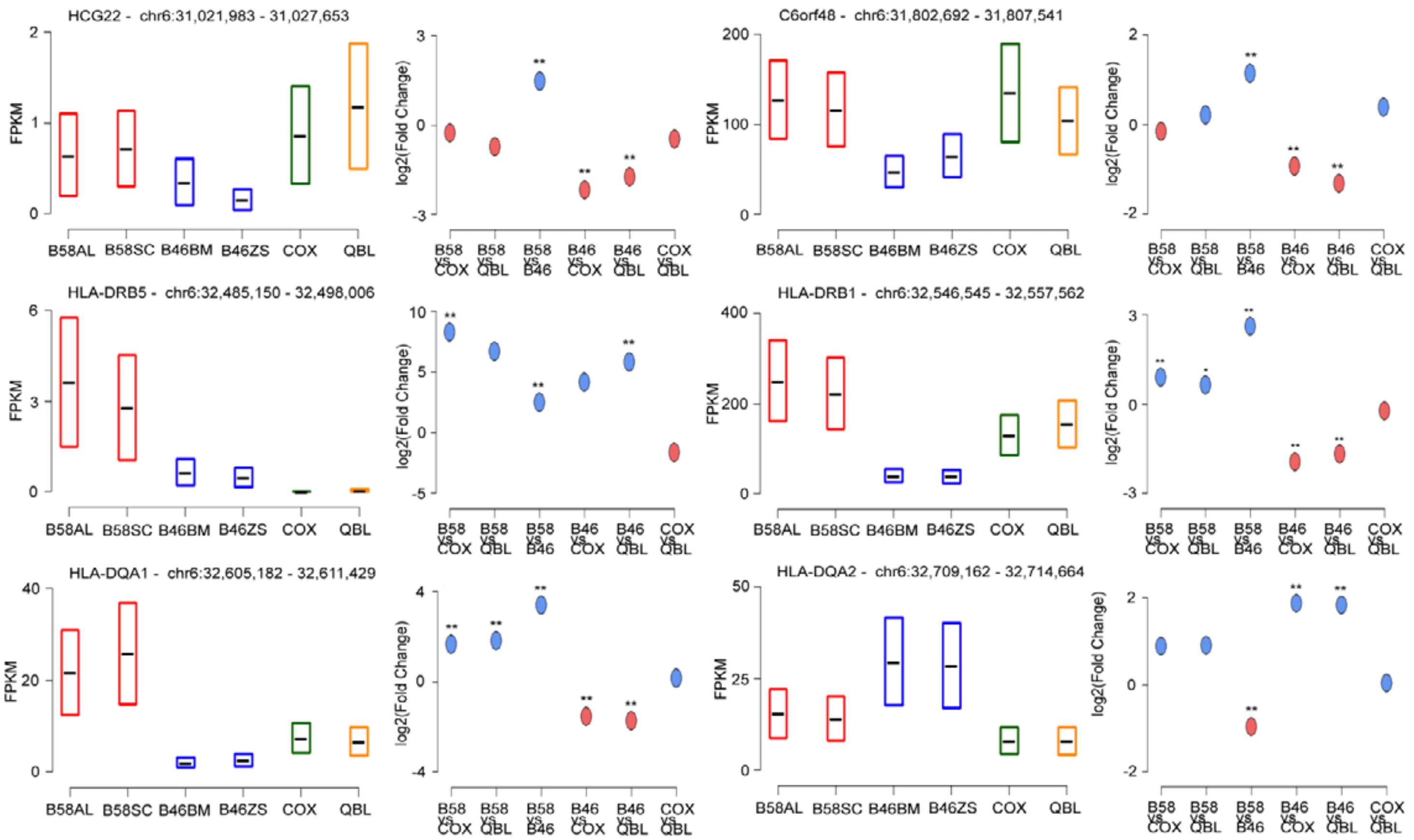


Figure 5.10 Individual cell line gene expression levels. These genes attained statistical significant in three or more haplotype-pairs comparisons. Red indicates cell lines carrying A33-B58-DR3 haplotype; blue indicates cell lines carrying A2-B46-DR9 haplotype; green indicates cell line carrying A1-B8-DR3 haplotype and orange indicates cell line carrying A26-B18-DR3 haplotype. ** denotes haplotype-pair with adjusted P-value <0.05; * denotes haplotype-pair with adjusted P-value <0.1. The boundaries of the rectangle indicate the 95% confidence intervals in expression.

5.2.5 ZFP57 Expression

The RNA-seq data allows us to verify the expression of *ZFP57* as well as assesses the possible isoforms expressed in the cell lines. Indeed, the RNA-seq data correlates with the RT-qPCR experimental results where expression of *ZFP57* was found in the B46BM and COX cell lines (Figure 5.10). Next, we also examined the possible expressing isoforms by plotting the sashimi plots which displayed the raw reads coverage mapped to the exons and splicing junctions (Figure 5.11). Currently there are two known *ZFP57* isoforms which is differed by the extension of the 5’ UTR. From the RNA-seq reads, it appears both isoforms are expressed in the cell lines. In addition, there is indication for the expression of a third shorter novel isoform which lack of the first exon of the previous isoforms. However this observation could be an artifact resulted from erroneous reads mapping and therefore more quantitative analysis and verification by experimental approaches are needed to confirm the presence of this putative isoform.

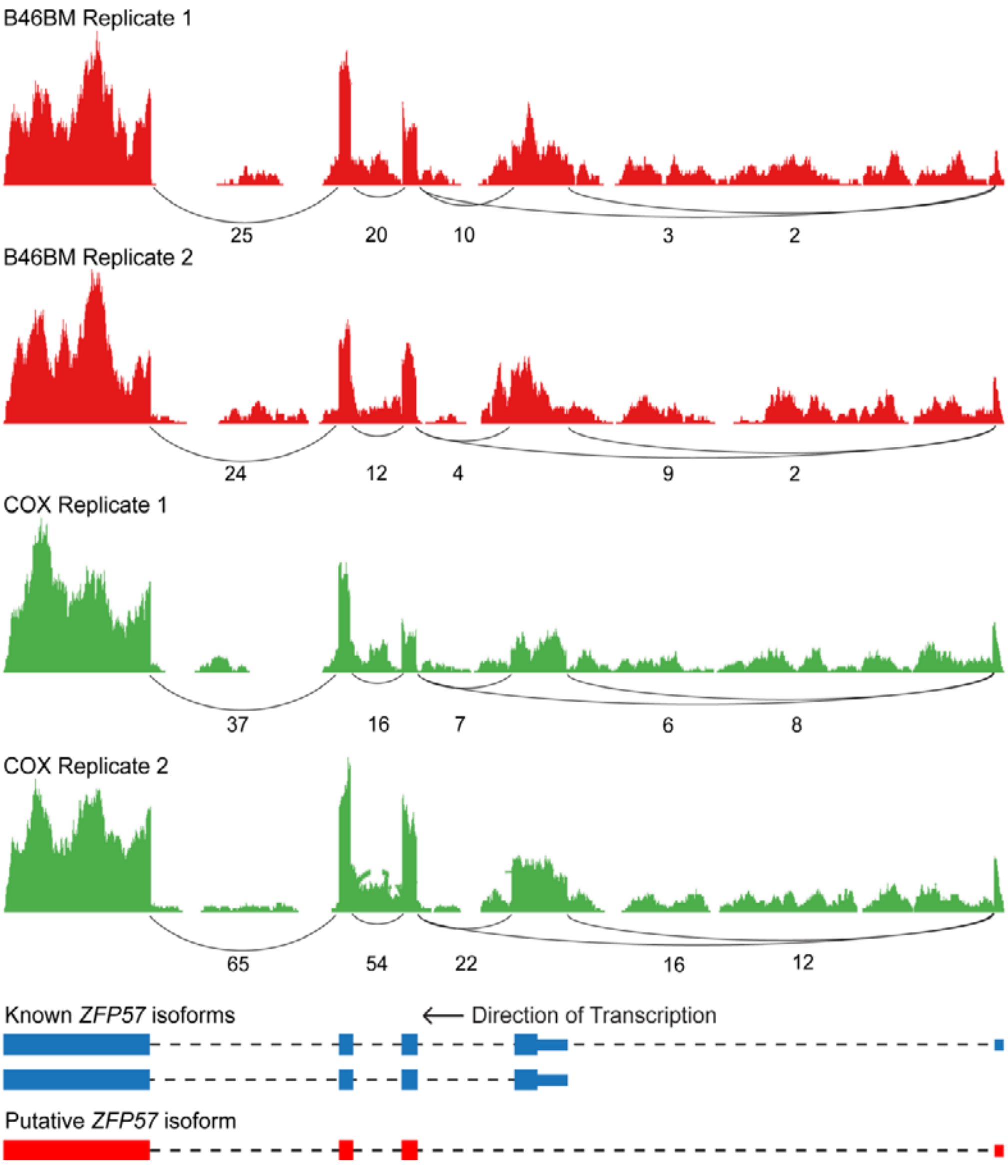


Figure 5.11 Sashimi plots for RNA-Seq analysis of *ZFP57* expression. Raw read densities mapped along the exons and splicing junctions of *ZFP57* in B46BM and COX cell lines. Spicing junctions are represented by the connecting lines and the number below each line indicates the number of reads spanning across each splicing junction. Below the sashimi plots are the known *ZFP57* isoforms from the NCBI gene annotation Build 37.2 and ensembl genes annotation as well as the putative isoform derived from the RNA-Seq mapping.

5.3 Conclusion

In this chapter, using RNA-seq technology on multiple HLA homozygous LCLs sharing identical HLA haplotype, we provided a comprehensive description on the transcriptional landscape in the MHC region (chr6:29.0 – 33.0Mb) for four common HLA haplotype (two with Asian ancestry – A33-B358-DR3 and A2-B46-DR9; two with European ancestry – A1-B8-DR3 and A26-B18-DR3). We observed elevated expression of the HLA class I / class II genes and genes related by their functions were expressed at equivalent levels. In addition, differences in gene expression as well as in alternate splicing events were influenced by the underlying MHC haplotypic structure. The transcription activities of 26 genes were found to be affected by the MHC haplotype diversity. This implies that underlying MHC haplotypic structure might have an effect on the transcription activity in the MHC region.

Chapter 6:

Discussion

6.1 Significance of the MHC Recombination Analysis in this Study

Current *in silico* approaches using of population genetics data is relatively successful in estimating the location of recombination sites on a genome-wide basis [199] which are showed to have good correlation with predigree-derived recombination maps and with known recombination sites. Most of these approaches are based on a population genetic model that does not account for natural selection, random mutation, human migration and variation of the effective population size [200]. Though it has proved to be robust with reasonable departure from those conditions, cautious has to be exercised when applying the model to infer recombination sites in the human MHC region given that this region is subjected to strong selection forces. The human MHC region is unique as it is characterized by diverse HLA haplotypic variation and structure. *In silico* approaches that used pooled global population data is unable to effectively account for the contribution of HLA allelic haplotypes towards recombination. The admixture of diverse HLA haplotypes complicates the inference process and obscures potential recombination sites. To address the problem, in this work, EHH plots using phased HLA haplotypes derived from CHSG were employed to identify MHC-residing recombination sites. This method was demonstrated to be an effective tool at locating recombination sites in the extended MHC region. The 69 putative recombination sites identified in this study correlate well with recombination segments determined by sperm typing [47] as well as with the majority of HapMap predicted sites [160]. We also uncovered an additional 37 sites that are not found in any of the previous studies. A review of recent reports [201, 202] based on European pedigree information, ascertains that our approach is far more sensitive at locating recombination events (69 sites vs 5 sites) within MHC and provides superior resolution (average 4.27kb vs 100kb intervals).

Earlier studies from Walsh *et al* [50] and Ahmad *et al* [203] have characterized the LD patterns of the Caucasian population. Though these studies highlighted the extent of LD within the MHC region, the resultant LD maps were limited in scope due to the relative small amount of polymorphic markers used to derive them and could not be used to infer recombination sites. A more recent study created a comprehensive haplotype map by investigating the LD between the SNPs and the HLA alleles across the extended MHC region (7.5Mb) in four populations [163]. This study effectively demonstrated that the extent of LD along the chromosome is dependent on the underlying HLA allelic haplotype and provided a panoramic view of the MHC genomic architecture. In comparison, our study provides a more detail and precise description on the variation of LD structure and breakages in distinctive HLA haplotypes localized to the HLA class I and class II gene regions where the LD breakpoints are interpreted as probable recombination sites.

6.2 Population-specific Recombination

The International HapMap Project reported over 33,000 genome-wide recombination sites derived from a pooled population comprising of CEU, YRI and an Asian population (Han Chinese, CHB and Japanese, JPT) [137, 204]. These HapMap recombination sites were classified only when two out of the three populations showed signal of recombination events, and hence, HapMap inferred recombination sites are commonly found across populations, not population-specific. Indeed, there is no extensive study to examine recombination variation across populations within MHC region. In the study by de Bakker *et al* [52], the recombination rates were separately estimated from individuals of a distinct population and coalesced the rates from four different populations into a single estimate for the MHC region, but did not delved further to evaluate the differences in recombination rates between populations. A subsequent study did a recombination rates correlation analysis on the LD-maps generated from three populations and found population recombination variation at the whole genome level [205]. However this study did not investigate on how the differences of the HLA haplotypes pool in each distinct population influence and affect the recombination variation in the MHC region. To provide a clearer insight into the population-specific recombination profile and the effect of underlying HLA haplotype on recombination variations, we applied the EHH approach coupled with the HLA information of the CHSG, CEU and YRI population. Our study shows low number of recombination sites overlap in multiple populations; in fact, > 50% of the identified recombination sites is specific to a single population. Recombination activities bring about the breakdown of LD and have a direct effect on the genome haplotype diversity [44]. The findings of vast number of unique recombination sites in a distinct population suggest that these population-specific sites could have a major role in the diversification of haplotypes in the MHC region. In contrast, in other parts of the human genome, the sites of LD decay are generally common across populations; resulting in extensive haplotype sharing among different populations [148, 161]. Given that the genomic region of the MHC can be influenced by a variety of evolutionary mechanisms such as genetic drift, demography and natural selection, the distribution of population recombination sites in the MHC region may not be a good refection on the recombination distribution elsewhere in the human genome.

Balancing selection through pathogen mediated selection are proposed to explain for the immense number of HLA alleles and haplotypes [206]. These selection forces favor new assortment of HLA allelic combinations across populations and lead to an increase in population frequency of these HLA combinations [15]. These mechanisms result in the occurrence of a given allele in two or more different HLA haplotype backgrounds. In this study, the A*02:01 allele is independently found in different HLA haplotype background in multiple populations and their recombination profiles varies across populations; indicating the combinatorial effects of underlying HLA haplotype and population background on recombination. On the whole, our data shows the significant role of HLA haplotypes on the patterns and occurrence of recombination events in the MHC region; and the discovery of unique recombination sites are possible only through single population analysis.

6.3 Evolutionary Conserved Recombination Sites

Previous section highlighted the differences of recombination sites between population groups; however, few sites are observed to be shared between populations (Table 3.3) and these are also of interest as this means that they are evolutionarily highly conserved and could therefore be of great importance. Though explanation for the mechanism behind the occurrence of these conserved recombination sites is not well-defined, one can study the recombination pathways to derive possible hypotheses. Initiation of recombination at a particular site starts when histone methyltransferase such as PR domain-containing 9 (Prdm9), locally acts and opens up the chromatin. This allows the topoisomerase sporulation-specific 11 (SPO11) to introduce double-strand breaks (DSBs) on the chromatid and the DSBs are then repaired through the process of homologous recombination resulting in cross-over or gene conversion events [207]. Therefore it could be possible that the local chromatin state at the conserved recombination sites is highly favorable for the recruitment of recombination-initiation machinery. Another probable hypothesis can be derived from the need to balance the two divergent mechanisms during meiosis where recombination is essential for proper chromosome separation but it also has to be controlled to minimize the breakage of important gene clusters and to maintain genome stability [44]. The occurrence of conserved recombination sites could suggest recombination activities at these genomic sites are conducive for the proper repair of DSBs; as such, would not disrupt the favorable linkage of gene clusters and would have little or no deleterious effect on the genome stability.

6.4 Sequencing MHC CEHs and Intra-haplotypic Variations

In this study, the presence of Singaporean Chinese CEHs in A33-B58-DR3 and A2-B46-DR9 HLA haplotypes were identified through SNP genotyping as well as deep sequencing of the MHC region (28.5 to 33.5 Mb). We have assembled at least 90% of the MHC sequence representative of the A33-B58-DR3 and A2-B46-DR9 haplotypes; and discovered that the sequences of these haplotypes are largely conserved from *HLA-F* to *HLA-DQA2* loci covering at least 3Mb of the MHC genomic region, proving the CEH nature of these haplotypes. The assembly of these common Singapore Chinese MHC sequences could act as an important framework for future disease studies on populations enriched with these haplotypes, which include Asian populations from Southern China and Taiwan [173, 174, 208].

Genome-wide association studies have identified more than 100 diseases that are implicated by the variants or genes within the MHC region [7]. Often, such genetic associations are not due to single specific variants, but by the underlying MHC haplotype structure marked by extensive LD [13]. These issues have complicated the identification of disease-causing variants or genes within the MHC region, especially in common CEHs, where extreme LD is often found. By deep sequencing of multiple individuals homozygous for the same haplotype, we were able to circumvent the assignment of haplotype chromosome and perform intra-haplotypic CEH comparison at the nucleotide level to determine the extent of variation within the conserved boundaries of each CEH. In total, more than 200 haplotype-specific SNVs were uncovered residing in each haplotype, up to a third of which are not annotated in any public archives for genetic polymorphism. In contrast, the use of the common SNP genotyping platform to interrogate the CEHs was unable to reveal the polymorphisms embedded within the conserved region. The ability to reveal variants that is enriched at the haplotype level but not at the population level allows fine discrimination within members containing the same CEH. This may allow us to better predict case-control status among individuals sharing the same risk-associated CEH, and also to define risk-associated variants that may be enhanced within a particular CEH.

The intra-haplotypic variations revealed in our study suggests that the pattern of variations in each CEH is unique. For example, a hyper-variable region was observed around the HLA-A region on the A33-B58-DR3 but not on A2-B46-DR9. Given that each MHC CEH has a distinct underlying HLA background and evolutionary history, and has been subjected to different environmental influences, such non-homogeneous variation distribution may be expected. The increase in variation around *HLA-A* in A33-B58-DR3 despite the apparent lack of recombination in this haplotype hints at the presence of non-random mutation in the region. The existence of non-random mutation rates across the genome has been recognized in organisms as diverse as bacteria and humans relatively recently, and there are still many unknowns regarding the biochemical mechanisms behind such phenomena [209-211]. It is tempting to speculate that specific genetic characteristics of *HLA-A* may increase its mutation rate, contributing to the increased global diversity observed at the locus in a manner independent but complementary to balancing selection. In fact, the occurrence of these intra-haplotypic variants where 99% of the A33-B58-DR3 and 23% of the A2-B46-DR9 variants exist in the heterozygote form can be well explained by an alternate model on MHC evolution called Associative Balancing Complex (ABC) [79]. This model states that recessive detrimental mutations are built up by the Muller's ratchet effect and are sheltered by the surrounding MHC genes through LD. Moreover, these mutations exist in the heterozygote forms and as such natural selection is not effective to select against these mutations. This leads to negative epistasis and the reduction of recombination events in the MHC. The ABC evolution model therefore accounts for the high number of heterozygous intra-haplotypic variants in the MHC CEHs and may provide a valid explanation for why numerous diseases are associated to MHC CEHs.

The identified intra- and inter-haplotypic variants in this work may be helpful in offering important clinical links to diseases. MHC-resident variants within the CEHs, apart from the HLA genes, have independent associations with diseases, even in autoimmune-related disorders such as systemic lupus erythematosus, Behcet's disease, graft-versus-host disease and rheumatoid arthritis [212-216]. These variants may exert *cis*-regulatory effects on the nearby genes and affect the expression of the target genes commonly known as *cis*-expression quantitative trait loci (cis-eQTLs). Recent studies have suggested that the MHC region is a prominent area for such cis-eQTL associations [133, 217, 218], and in our study, we also show that one of the A2-B46-DR9 intra-haplotypic variant regulates the expression of the *ZFP57* gene. The presence of such cis-eQTLs provides a possible mechanistic theory for the outstanding number of MHC disease associations yet to be explained, particularly the contribution of non-coding variants to disease phenotypes. In this work, up to 99% of the intra-haplotypic variations are located within the non-coding region; these variants can be incorporated into a reference panel to infer the effects of non-coding variants on diseases.

The numerous associations of MHC CEHs with various diseases signal a need to dissect CEHs to identify the true disease-causing variants among the large pool of benign variants in LD within the haplotype. Hence, it is essential to characterize the extent of variation within the MHC CEHs. The advent of next-generation sequencing technology offers an attractive option to assess such haplotypic variation. However, the short reads generated by NGS platforms are not well suited to accurately map the MHC region given its extreme sequence and structural variations. More importantly, it is difficult to resolve the phase of the haplotype from the short reads mapping. As seen in this study, the use of HLA homozygous cell lines avoids the need for phasing and improves the accuracy of the reads mapping. Third generation sequencing technologies with the ability to sequence reads of length greater than 1000bp [219, 220] have the potential to both phase and accurately map complex region such within the MHC and will eventually allow large scale haplotype comparative analysis.

6.5 Origin and Age of Conserved Extended Haplotype (CEH)

The MHC CEHs can span over several megabases contain numerous genes that are fundamental for human immune functions. The mechanism behind the maintenance of such extensive LD in this important and highly variable region of the human genome is unclear. There are two plausible explanations for the generation of MHC CEHs. The first is that these long sweeps of conserved sequence haplotypes may have been driven to high frequency by positive selection over a relatively short period of time and have yet to be disrupted by recombination events [10, 144]. A single gene or a combination of genes within the conserved stretch would be adequate to drive the haplotype expansion in the population. Another possibility is that given the almost non-existent of recombination events on haplotypes carrying a specific HLA allelic combination, these extensive conserved segments are exposed to allele-specific recombination suppression preventing haplotype breakdown [203]. The age of the A1-B8-DR3 haplotype has been estimated to be about 23,500 years [21]. The relative young age of the A1-B8-DR3 suggests that the extensive LD observed between the HLA alleles is more likely due to recent expansion and that recombination forces have yet to act on this haplotype. Assuming that the human mutation rate per nucleotide per generation is $1.1X10^{-8}$ [154] and a given generation is 20 years, the age of the Asian A33-B58-DR3 and A2-B46-DR9 works out to be about 21,460 and 26,500 years respectively. These values are compatible with the age of the A1-B8-DR3 haplotype lending support to the theory that these MHC CEHs are likely to be resulted from recent expansion. In addition, it has been suggested that the expansion of the CEHs is possibly because of a single HLA allele under strong positive selection rather than epistatic selection of a specific HLA allelic combination [10]. This study reports the occurrence of sequence similarity at the HLA-DRB region between two common CEHs (A33-B58-DR3 and A1-B8-DR3) which belong to different ethnic backgrounds but were both driven to high frequency in their respective populations. It is conceivable that this ancestral DR3 segment was introduced into different MHC haplotype background whereby the selection for this segment drove the expansion of these MHC haplotypes.

Where is the origin of this DR3 ancestral segment? Given that the A33-B58-DR3 and A1-B8-DR haplotypes are found in populations that are geographically distant to each other, it is unlikely that the DR3 segment is derived from either of the two haplotypes. It is more likely that this segment could originate from another population group and diverge into the European and Chinese population. The study of the human migration pattern might able to provide insight to this hypothesis. Based on the Out of Africa theory, human first migrated out of Africa into Middle Asia then spread to South Asia by 50,000 years ago, and from South Asia human slowly spread to China, South East Asia and then finally reached Europe by 40,000 years ago [221, 222]. Literature search on the MHC haplotype distribution has revealed an enrichment of DR3 in the modern South Asian population (A24-B8-DR3 - 4.8% & A26-B8-DR3 - 6.2%) [170], hence it is plausible that the DR3 segment could origin from the South Asian population and then independently expands in the European and Chinese population through human migration. To validate this model, further work has to be conducted to ensure DR3 sequence similarity can be found between the South Asian DR3 and the Chinese/European DR3 segment and preferably supported by the age of the South Asian MHC haplotypes.

6.6 Influence of Haplotype-defined Nucleotide Sequence Variation on MHC Gene Regulation

There are many instances where the underlying HLA haplotype rather than genotype of a specific locus plays a significant role in disease susceptibility [13, 223, 224]. In this thesis, we have unambiguously identified haplotype-specific sequence variations. Though the knowledge of these variations is important for the explanation of the occurrence of the disease, it could not have revealed the functional mechanisms that lead to disease progression. Therefore, the ability to map the transcriptomic landscape of the MHC region at haplotypic resolution is an important step in understanding the molecular basis for a number of MHC associations to diseases. In this study, the MHC transcriptome profiles of multiple HLA homozygous B-LCLs with identical Singaporean Chinese HLA haplotype were examined using RNA-seq. This approach not only averts the confounding effect in the evaluation of diploid genome but will also account for transcription variability between cell lines. Overall, our study has revealed that the expressions of 26 genes are attributed to haplotypic effects. In comparison, the report by Vandiedonck *et al* [179] which was based on the analysis of three common European MHC haplotypes identified 96 genes whose expression levels are associated to haplotypic differences. There are two likely reasons for this discrepancy. The first is the difference in the MHC haplotypes used in context between the two studies. Secondly, in the work by Vandiedonck *et al*, each MHC haplotype was represented by only a single cell line; hence some of the transcription differences reported could be due to individual cell lines variations rather than haplotype-related transcriptional differences. In contrast, our analysis involved two independent cell lines carrying identical MHC haplotype and we showed that haplotype-identical cell lines have equivalent expression in 25 out 26 genes. This provides strong evidence that the differential gene expression observed is indeed due to haplotypic effects and dismisses the notion of individual cell lines variations. Nevertheless, 15/26 differentially expressed genes identified in this thesis overlapped with the genes listed in Vandiedonck's work. This suggests the MHC transcriptional landscape is likely to vary in a context-specific manner; dependent on the relevant haplotype of interest, the cell types and conditions applied. Together, the identification of genes whose expression levels are implicated by haplotypic differences allow us to shortlist candidate genes to consider for diseases linked to the haplotype in context.

Haplotype-related transcriptional differences signify the possible effect of haplotype-defined nucleotide sequence variations on MHC gene regulation. Haplotype-specific sequence variations in cis–acting regulatory promoter elements or even distal trans-acting regulatory elements could affect DNA methylation or chromatin accessibility and hence are critical to the regulation of gene expression. Indeed, sequence variations in the enhancer and the interferon-stimulated response element in the MHC class I promoter bring about differential promoter activation among various MHC class I loci [120, 225]. Currently, much of the MHC epigenetic studies are directed at the regulation of transcription initiation. The complex interplay between cris- /trans- regulatory elements and the epigenetic mechanisms that modulate the expression of MHC class I and class II genes are well established [118, 226, 227]. However, the effect of sequence diversity, in particular sequence variants defined by the MHC haplotypic structure, on epigenetic modifications is not well understood. Haplotype-specific sequence diversity can influence the entire transcription as well as translation process and also defines the packing of the chromatin that coordinates gene expression at local and global level. Hence, the establishment and mapping chromatin modification at haplotypic resolution could well provide the mechanistic explanation on how transcription is modulated by underlying sequence variation. The ability to connect MHC sequence variation, chromatin modification and the subsequent transcription process has enormous potential to provide insight into the functional basis for many complex diseases linked to the MHC region.

6.7 Future Directions

In this thesis, we found that certain individuals showed expression of *ZFP57*, a Kruppel-associated box (KRAB) containing zinc-finger protein, in the B-LCLs; dependent on the nucleotide sequence configuration at multiple positions proximate to the gene. These positions signify possible regulatory sites for the transcription of *ZFP57*; therefore further work is required to ascertain which of these site(s) truly regulate the expression of the gene. Approach such as cloning the genomic segments bearing these sites into an expression vector and transfect them into a non-expressing *ZFP57* mammalian cell can be used to identify the possible regulatory region. This identification of *ZFP57* regulatory region(s) can facilitate the finding of the transcription factors or other co-factors that may involve in the modulation of the *ZFP57* transcription. In embryonic stem (ES) cells*,* signal transducer and activator of transcription *3 (*STAT3) and octamer-binding transcription factor -3/4 (Oct-3/4) are the transcription activators for *ZFP57* [228], it would be of interest to examine whether the same transcription factors or other novel factors are utilized to regulate the transcription *ZFP57* in adult cell lines.

To date, ZFP57 is involved in the maintenance of DNA methylation whereby ZFP57 acts as an anchor for the binding of KAP1 and the recruitment of other epigenetic regulators at imprinting control regions in the ES cells [229]. However, functional significance of the expression of *ZFP57* and its role in the adult cells is not known. The KRAB-ZNF family genes have various functions such as cell-cycle regulation [230], specification of meiotic recombination hotspots [231] and more importantly some of the KRAB-ZNF genes exhibit tumor suppression properties [230]. In addition, ZFP57's co-factor KAP1 is found to be associated with tumor developments in several studies [232-234]. Hence further work is needed to investigate on the role of ZFP57 in epigenetic regulation in the context of cancers or infection.

The A2-B46-DR9 and A33-B58-DR3 HLA haplotypes are implicated in multiple diseases. Most notably, several studies have reported the association of these haplotypes with nasopharyngeal carcinoma [235-237] and a recent GWAS study has identified 3 new susceptibility loci with extremely strong statistical significance within the MHC region [82]; but the exact location of the disease causative variant/element is yet to be determined. The findings of A2-B46-DR9 and A33-B58-DR3 haplotype-specific sequence variations in this study, in particular the novel variants not reported in any prior studies, offer an excellent opportunity to revisit these association studies. These haplotype-specific variations might be able to distinguish disease-affected haplotype carriers from unaffected haplotype carriers and eventually facilitate the mapping of the disease causative variant/element.

Increasingly in recent years, it is apparent that a significant fraction of the human genome expresses non-coding RNAs (ncRNAs) and some these ncRNAs are crucial for normal development and physiology [238]. Hence, it is not surprising that many of them are implicated in numerous diseases. Many studies have showed that the dysregulation of these ncRNAs contribute to the development and progression of many human conditions particularly in cancers [239-241]. It would be of interested to investigate the presence of ncRNAs in the MHC region and more significantly to examine whether the underlying haplotype structure would influence the expression of ncRNAs. Through de novo analysis, the RNA-seq data used in this study can be employed to inspect the expression of the ncRNAs in the Singaporean Chinese HLA haplotype and provides insights for the above mentioned questions.

References

1. Dausset J: The major histocompatibility complex in man. *Science* 1981, 213:1469-1474.
2. Snell GD: Methods for the study of histocompatibility genes. *J Genet* 1948, 49:87-108.
3. Bodmer WF: Evolutionary significance of the HL-A system. *Nature* 1972, 237:139-145 passim.
4. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, Lush MJ, Povey S, Talbot CC, Jr., Wright MW, et al: Gene map of the extended human MHC. *Nat Rev Genet* 2004, 5:889-899.
5. Shiina T, Hosomichi K, Inoko H, Kulski JK: The HLA genomic loci map: expression, interaction, diversity and disease. *J Hum Genet* 2009, 54:15-39.
6. Shiina T, Inoko H, Kulski JK: An update of the HLA genomic region, locus information and disease associations: 2004. *Tissue Antigens* 2004, 64:631-649.
7. Trowsdale J: The MHC, disease and selection. *Immunol Lett* 2011, 137:1-8.
8. Buhler S, Sanchez-Mazas A: HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events. *PLoS One* 2011, 6:e14643.
9. Spurgin LG, Richardson DS: How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proc Biol Sci* 2010, 277:979-988.
10. Traherne JA: Human MHC architecture and evolution: implications for disease association studies. *Int J Immunogenet* 2008, 35:179-192.
11. Robinson J, Halliwell JA, McWilliam H, Lopez R, Parham P, Marsh SG: The IMGT/HLA database. *Nucleic Acids Res* 2013, 41:D1222-1227.
12. Trowsdale J, Knight JC: Major Histocompatibility Complex Genomics and Human Disease. *Annu Rev Genomics Hum Genet* 2013.
13. Vandiedonck C, Knight JC: The human Major Histocompatibility Complex as a paradigm in genomics research. *Brief Funct Genomic Proteomic* 2009, 8:379-394.
14. Horton R, Gibson R, Coggill P, Miretti M, Allcock RJ, Almeida J, Forbes S, Gilbert JG, Halls K, Harrow JL, et al: Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. *Immunogenetics* 2008, 60:1-18.
15. Traherne JA, Horton R, Roberts AN, Miretti MM, Hurles ME, Stewart CA, Ashurst JL, Atrazhev AM, Coggill P, Palmer S, et al: Genetic analysis of completely sequenced disease-associated MHC haplotypes identifies shuffling of segments in recent human history. *PLoS Genet* 2006, 2:e9.
16. Stewart CA, Horton R, Allcock RJ, Ashurst JL, Atrazhev AM, Coggill P, Dunham I, Forbes S, Halls K, Howson JM, et al: Complete MHC haplotype sequencing for common disease gene mapping. *Genome Res* 2004, 14:1176-1187.
17. Yeo TW, De Jager PL, Gregory SG, Barcellos LF, Walton A, Goris A, Fenoglio C, Ban M, Taylor CJ, Goodman RS, et al: A second major histocompatibility complex susceptibility locus for multiple sclerosis. *Ann Neurol* 2007, 61:228-236.
18. Nejentsev S, Howson JM, Walker NM, Szeszko J, Field SF, Stevens HE, Reynolds P, Hardy M, King E, Masters J, et al: Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. *Nature* 2007, 450:887-892.

19. Proll J, Danzer M, Stabentheiner S, Niklas N, Hackl C, Hofer K, Atzmuller S, Hufnagl P, Gully C, Hauser H, et al: Sequence capture and next generation resequencing of the MHC region highlights potential transplantation determinants in HLA identical haematopoietic stem cell transplantation. *DNA Res* 2011, 18:201-210.

20. Cao H, Wu J, Wang Y, Jiang H, Zhang T, Liu X, Xu Y, Liang D, Gao P, Sun Y, et al: An Integrated Tool to Study MHC Region: Accurate SNV Detection and HLA Genes Typing in Human MHC Region Using Targeted High-Throughput Sequencing. *PLoS One* 2013, 8:e69388.

21. Smith WP, Vu Q, Li SS, Hansen JA, Zhao LP, Geraghty DE: Toward understanding MHC disease associations: partial resequencing of 46 distinct HLA haplotypes. *Genomics* 2006, 87:561-571.

22. Andersson G: Evolution of the human HLA-DR region. *Front Biosci* 1998, 3:d739-745.

23. Marsh SGE, Parham P, Barber LD: The HLA factsbook. In *Factsbook series*. pp. xiii, 398 p. ill. 324 cm. San Diego: Academic Press,; 2000:xiii, 398 p. ill. 324 cm.

24. Vincent R, Louis P, Gongora C, Papa I, Clot J, Eliaou JF: Quantitative analysis of the expression of the HLA-DRB genes at the transcriptional level by competitive polymerase chain reaction. *J Immunol* 1996, 156:603-610.

25. Kerlan-Candon S, Combe B, Vincent R, Clot J, Pinet V, Eliaou JF: HLA-DRB1 gene transcripts in rheumatoid arthritis. *Clin Exp Immunol* 2001, 124:142-149.

26. Chung EK, Yang Y, Rennebohm RM, Lokki ML, Higgins GC, Jones KN, Zhou B, Blanchong CA, Yu CY: Genetic sophistication of human complement components C4A and C4B and RP-C4-CYP21-TNX (RCCX) modules in the major histocompatibility complex. *Am J Hum Genet* 2002, 71:823-837.

27. Blanchong CA, Zhou B, Rupert KL, Chung EK, Jones KN, Sotos JF, Zipf WB, Rennebohm RM, Yung Yu C: Deficiencies of human complement component C4A and C4B and heterozygosity in length variants of RP-C4-CYP21-TNX (RCCX) modules in caucasians. The load of RCCX genetic diversity on major histocompatibility complex-associated disease. *J Exp Med* 2000, 191:2183-2196.

28. Li L, Chow SC, Smith W: Cross-validation for linear model with unequal variances in genomic analysis. *J Biopharm Stat* 2004, 14:723-739.

29. Yang Z, Mendoza AR, Welch TR, Zipf WB, Yu CY: Modular variations of the human major histocompatibility complex class III genes for serine/threonine kinase RP, complement component C4, steroid 21-hydroxylase CYP21, and tenascin TNX (the RCCX module). A mechanism for gene deletions and disease associations. *J Biol Chem* 1999, 274:12147-12156.

30. Jaatinen T, Eholuoto M, Laitinen T, Lokki ML: Characterization of a de novo conversion in human complement C4 gene producing a C4B5-like protein. *J Immunol* 2002, 168:5652-5658.

31. Yang Y, Chung EK, Wu YL, Savelli SL, Nagaraja HN, Zhou B, Hebert M, Jones KN, Shu Y, Kitzmiller K, et al: Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *Am J Hum Genet* 2007, 80:1037-1054.

32. Sweeten TL, Odell DW, Odell JD, Torres AR: C4B null alleles are not associated with genetic polymorphisms in the adjacent gene CYP21A2 in autism. *BMC Med Genet* 2008, 9:1.

33. Amos B, Ward FE, Zmijewski CM, Hattler BG, Seigler HF: Graft donor selection based upon single locus (haplotype) analysis within families. *Transplantation* 1968, 6:524-534.

34. Alper CA, Raum D, Karp S, Awdeh ZL, Yunis EJ: Serum complement 'supergenes' of the major histocompatibility complex in man (complotypes). *Vox Sang* 1983, 45:62-67.

35. Dawkins R, Leelayuwat C, Gaudieri S, Tay G, Hui J, Cattley S, Martinez P, Kulski J: Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease. *Immunol Rev* 1999, 167:275-304.

36. Williams TM: Human leukocyte antigen gene polymorphism and the histocompatibility laboratory. *J Mol Diagn* 2001, 3:98-104.

37. Yunis EJ, Larsen CE, Fernandez-Vina M, Awdeh ZL, Romero T, Hansen JA, Alper CA: Inheritable variable sizes of DNA stretches in the human MHC: conserved extended haplotypes and their fragments or blocks. *Tissue Antigens* 2003, 62:1-20.

38. Szilagyi A, Banlaki Z, Pozsonyi E, Yunis EJ, Awdeh ZL, Hosso A, Rajczy K, Larsen CE, Fici DA, Alper CA, Fust G: Frequent occurrence of conserved extended haplotypes (CEHs) in two Caucasian populations. *Mol Immunol* 2010, 47:1899-1904.

39. Degli-Esposti MA, Leaver AL, Christiansen FT, Witt CS, Abraham LJ, Dawkins RL: Ancestral haplotypes: conserved population MHC haplotypes. *Hum Immunol* 1992, 34:242-252.

40. Price P, Witt C, Allcock R, Sayer D, Garlepp M, Kok CC, French M, Mallal S, Christiansen F: The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases. *Immunol Rev* 1999, 167:257-274.

41. Candore G, Lio D, Colonna Romano G, Caruso C: Pathogenesis of autoimmune diseases associated with 8.1 ancestral haplotype: effect of multiple gene interactions. *Autoimmun Rev* 2002, 1:29-35.

42. Jawaheer D, Li W, Graham RR, Chen W, Damle A, Xiao X, Monteiro J, Khalili H, Lee A, Lundsten R, et al: Dissecting the genetic complexity of the association between human leukocyte antigens and rheumatoid arthritis. *Am J Hum Genet* 2002, 71:585-594.

43. Carrington M: Recombination within the human MHC. *Immunol Rev* 1999, 167:245-256.

44. Kauppi L, Jeffreys AJ, Keeney S: Where the crossovers are: recombination distributions in mammals. *Nat Rev Genet* 2004, 5:413-424.

45. Kauppi L, Sajantila A, Jeffreys AJ: Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. *Hum Mol Genet* 2003, 12:33-40.

46. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, et al: A high-resolution recombination map of the human genome. *Nat Genet* 2002, 31:241-247.

47. Cullen M, Perfetto SP, Klitz W, Nelson G, Carrington M: High-resolution patterns of meiotic recombination across the human major histocompatibility complex. *Am J Hum Genet* 2002, 71:759-776.

48. Jeffreys AJ, Kauppi L, Neumann R: Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* 2001, 29:217-222.

49. Miretti MM, Walsh EC, Ke X, Delgado M, Griffiths M, Hunt S, Morrison J, Whittaker P, Lander ES, Cardon LR, et al: A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. *Am J Hum Genet* 2005, 76:634-646.

50. Walsh EC, Mather KA, Schaffner SF, Farwell L, Daly MJ, Patterson N, Cullen M, Carrington M, Bugawan TL, Erlich H, et al: An integrated haplotype map of the human major histocompatibility complex. *Am J Hum Genet* 2003, 73:580-590.

51. Stenzel A, Lu T, Koch WA, Hampe J, Guenther SM, De La Vega FM, Krawczak M, Schreiber S: Patterns of linkage disequilibrium in the MHC region on human chromosome 6p. *Hum Genet* 2004, 114:377-385.

52. de Bakker PI, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, Ke X, Monsuur AJ, Whittaker P, Delgado M, et al: A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat Genet* 2006, 38:1166-1172.

53. Crawford DC, Bhangale T, Li N, Hellenthal G, Rieder MJ, Nickerson DA, Stephens M: Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat Genet* 2004, 36:700-706.

54. Ke X, Hunt S, Tapper W, Lawrence R, Stavrides G, Ghori J, Whittaker P, Collins A, Morris AP, Bentley D, et al: The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum Mol Genet* 2004, 13:577-588.

55. Neefjes J, Jongsma ML, Paul P, Bakke O: Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat Rev Immunol* 2011, 11:823-836.

56. Hoglund P, Brodin P: Current perspectives of natural killer cell education by MHC class I molecules. *Nat Rev Immunol* 2010, 10:724-734.

57. Kurts C, Robinson BW, Knolle PA: Cross-priming in health and disease. *Nat Rev Immunol* 2010, 10:403-414.

58. Wooldridge L, Ekeruche-Makinde J, van den Berg HA, Skowera A, Miles JJ, Tan MP, Dolton G, Clement M, Llewellyn-Lacey S, Price DA, et al: A single autoimmune T cell receptor recognizes more than a million different peptides. *J Biol Chem* 2012, 287:1168-1177.

59. O'Sullivan D, Arrhenius T, Sidney J, Del Guercio MF, Albertson M, Wall M, Oseroff C, Southwood S, Colon SM, Gaeta FC, et al.: On the interaction of promiscuous antigenic peptides with different DR alleles. Identification of common structural motifs. *J Immunol* 1991, 147:2663-2669.

60. Frahm N, Yusim K, Suscovich TJ, Adams S, Sidney J, Hraber P, Hewitt HS, Linde CH, Kavanagh DG, Woodberry T, et al: Extensive HLA class I allele promiscuity among viral CTL epitopes. *Eur J Immunol* 2007, 37:2419-2433.

61. Rao X, Hoof I, Costa AI, van Baarle D, Kesmir C: HLA class I allele promiscuity revisited. *Immunogenetics* 2011, 63:691-701.

62. Wearsch PA, Cresswell P: The quality control of MHC class I peptide loading. *Curr Opin Cell Biol* 2008, 20:624-631.

63. Raulet DH: Roles of the NKG2D immunoreceptor and its ligands. *Nat Rev Immunol* 2003, 3:781-790.

64. Gonzalez S, Lopez-Soto A, Suarez-Alvarez B, Lopez-Vazquez A, Lopez-Larrea C: NKG2D ligands: key targets of the immune response. *Trends Immunol* 2008, 29:397-403.

65. Hajeer AH, Hutchinson IV: TNF-alpha gene polymorphism: clinical and biological implications. *Microsc Res Tech* 2000, 50:216-228.

66. Qidwai T, Khan F: Tumour necrosis factor gene polymorphism and disease prevalence. *Scand J Immunol* 2011, 74:522-547.

67. Doherty PC, Zinkernagel RM: Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. *Nature* 1975, 256:50-52.

68. Bernatchez L, Landry C: MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? *J Evol Biol* 2003, 16:363-377.

69. Slade RW, McCallum HI: Overdominant vs. frequency-dependent selection at MHC loci. *Genetics* 1992, 132:861-864.

70. Hedrick PW: Pathogen resistance and genetic variation at MHC loci. *Evolution* 2002, 56:1902-1908.

71. Hughes AL, Nei M: Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 1988, 335:167-170.

72. Wegner KM, Kalbe M, Kurtz J, Reusch TB, Milinski M: Parasite selection for immunogenetic optimality. *Science* 2003, 301:1343.

73. Kalbe M, Eizaguirre C, Dankert I, Reusch TB, Sommerfeld RD, Wegner KM, Milinski M: Lifetime reproductive success is maximized with optimal major histocompatibility complex diversity. *Proc Biol Sci* 2009, 276:925-934.

74. Nowak MA, Tarczy-Hornoch K, Austyn JM: The optimal number of major histocompatibility complex molecules in an individual. *Proc Natl Acad Sci USA* 1992, 89:10896-10899.

75. Takahata N, Nei M: Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics* 1990, 124:967-978.

76. Apanius V, Penn D, Slev PR, Ruff LR, Potts WK: The nature of selection on the major histocompatibility complex. *Crit Rev Immunol* 1997, 17:179-224.

77. Martinsohn JT, Sousa AB, Guethlein LA, Howard JC: The gene conversion hypothesis of MHC evolution: a review. *Immunogenetics* 1999, 50:168-200.

78. Zemmour J, Gumperz JE, Hildebrand WH, Ward FE, Marsh SG, Williams RC, Parham P: The molecular basis for reactivity of anti-Cw1 and anti-Cw3 alloantisera with HLA-B46 haplotypes. *Tissue Antigens* 1992, 39:249-257.

79. van Oosterhout C: A new theory of MHC evolution: beyond selection on the immune genes. *Proc Biol Sci* 2009, 276:657-665.

80. McDevitt H: The discovery of linkage between the MHC and genetic control of the immune response. *Immunol Rev* 2002, 185:78-85.

81. De Jager PL, Jia X, Wang J, de Bakker PI, Ottoboni L, Aggarwal NT, Piccio L, Raychaudhuri S, Tran D, Aubin C, et al: Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat Genet* 2009, 41:776-782.

82. Bei JX, Li Y, Jia WH, Feng BJ, Zhou G, Chen LZ, Feng QS, Low HQ, Zhang H, He F, et al: A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci. *Nat Genet* 2010, 42:599-603.

83. Eleftherohorinou H, Hoggart CJ, Wright VJ, Levin M, Coin LJ: Pathway-driven gene stability selection of two rheumatoid arthritis GWAS identifies

and validates new susceptibility genes in receptor mediated signalling pathways. *Hum Mol Genet* 2011, 20:3494-3506.

84. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 2009, 106:9362-9367.

85. Moore CB, John M, James IR, Christiansen FT, Witt CS, Mallal SA: Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science* 2002, 296:1439-1443.

86. Carlson JM, Listgarten J, Pfeifer N, Tan V, Kadie C, Walker BD, Ndung'u T, Shapiro R, Frater J, Brumme ZL, et al: Widespread impact of HLA restriction on immune control and escape pathways of HIV-1. *J Virol* 2012, 86:5230-5243.

87. Tomiyama H, Miwa K, Shiga H, Moore YI, Oka S, Iwamoto A, Kaneko Y, Takiguchi M: Evidence of presentation of multiple HIV-1 cytotoxic T lymphocyte epitopes by HLA-B*3501 molecules that are associated with the accelerated progression of AIDS. *J Immunol* 1997, 158:5026-5034.

88. Kulkarni S, Savan R, Qi Y, Gao X, Yuki Y, Bass SE, Martin MP, Hunt P, Deeks SG, Telenti A, et al: Differential microRNA regulation of HLA-C expression and its association with HIV control. *Nature* 2011, 472:495-498.

89. Thursz MR, Thomas HC, Greenwood BM, Hill AV: Heterozygote advantage for HLA class-II type in hepatitis B virus infection. *Nat Genet* 1997, 17:11-12.

90. Carrington M, Nelson GW, Martin MP, Kissner T, Vlahov D, Goedert JJ, Kaslow R, Buchbinder S, Hoots K, O'Brien SJ: HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. *Science* 1999, 283:1748-1752.

91. Godkin A, Davenport M, Hill AV: Molecular analysis of HLA class II associations with hepatitis B virus clearance and vaccine nonresponsiveness. *Hepatology* 2005, 41:1383-1390.

92. Liu Y, Helms C, Liao W, Zaba LC, Duan S, Gardner J, Wise C, Miner A, Malloy MJ, Pullinger CR, et al: A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci. *PLoS Genet* 2008, 4:e1000041.

93. Aly TA, Ide A, Jahromi MM, Barker JM, Fernando MS, Babu SR, Yu L, Miao D, Erlich HA, Fain PR, et al: Extreme genetic risk for type 1A diabetes. *Proc Natl Acad Sci USA* 2006, 103:14074-14079.

94. International Consortium for Systemic Lupus Erythematosus G, Harley JB, Alarcon-Riquelme ME, Criswell LA, Jacob CO, Kimberly RP, Moser KL, Tsao BP, Vyse TJ, Langefeld CD, et al: Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. *Nat Genet* 2008, 40:204-210.

95. Fernando MM, Stevens CR, Walsh EC, De Jager PL, Goyette P, Plenge RM, Vyse TJ, Rioux JD: Defining the role of the MHC in autoimmunity: a review and pooled analysis. *PLoS Genet* 2008, 4:e1000024.

96. International MHC, Autoimmunity Genetics N, Rioux JD, Goyette P, Vyse TJ, Hammarstrom L, Fernando MM, Green T, De Jager PL, Foisy S, et al: Mapping of multiple susceptibility variants within the MHC region for 7 immune-mediated diseases. *Proc Natl Acad Sci USA* 2009, 106:18680-18685.

97. Nair RP, Stuart PE, Nistor I, Hiremagalore R, Chia NV, Jenisch S, Weichenthal M, Abecasis GR, Lim HW, Christophers E, et al: Sequence and

haplotype analysis supports HLA-C as the psoriasis susceptibility 1 gene. *Am J Hum Genet* 2006, 78:827-851.

98. Feng BJ, Sun LD, Soltani-Arabshahi R, Bowcock AM, Nair RP, Stuart P, Elder JT, Schrodi SJ, Begovich AB, Abecasis GR, et al: Multiple Loci within the major histocompatibility complex confer risk of psoriasis. *PLoS Genet* 2009, 5:e1000606.

99. Leone P, Shin EC, Perosa F, Vacca A, Dammacco F, Racanelli V: MHC class I antigen processing and presenting machinery: organization, function, and defects in tumor cells. *J Natl Cancer Inst* 2013, 105:1172-1187.

100. de Miranda NF, Nielsen M, Pereira D, van Puijenbroek M, Vasen HF, Hes FJ, van Wezel T, Morreau H: MUTYH-associated polyposis carcinomas frequently lose HLA class I expression - a common event amongst DNA-repair-deficient colorectal cancers. *J Pathol* 2009, 219:69-76.

101. Benitez R, Godelaine D, Lopez-Nevot MA, Brasseur F, Jimenez P, Marchand M, Oliva MR, van Baren N, Cabrera T, Andry G, et al: Mutations of the beta2-microglobulin gene result in a lack of HLA class I molecules on melanoma cells of two patients immunized with MAGE peptides. *Tissue Antigens* 1998, 52:520-529.

102. Hasim A, Abudula M, Aimiduo R, Ma JQ, Jiao Z, Akula G, Wang T, Abudula A: Post-transcriptional and epigenetic regulation of antigen processing machinery (APM) components and HLA-I in cervical cancers from Uighur women. *PLoS One* 2012, 7:e44952.

103. Meissner M, Reichert TE, Kunkel M, Gooding W, Whiteside TL, Ferrone S, Seliger B: Defects in the human leukocyte antigen class I antigen processing machinery in head and neck squamous cell carcinoma: association with clinical outcome. *Clin Cancer Res* 2005, 11:2552-2560.

104. Kaklamanis L, Leek R, Koukourakis M, Gatter KC, Harris AL: Loss of transporter in antigen processing 1 transport protein and major histocompatibility complex class I molecules in metastatic versus primary breast cancer. *Cancer Res* 1995, 55:5191-5194.

105. Mallal S, Nolan D, Witt C, Masel G, Martin AM, Moore C, Sayer D, Castley A, Mamotte C, Maxwell D, et al: Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *Lancet* 2002, 359:727-732.

106. Hung SI, Chung WH, Jee SH, Chen WC, Chang YT, Lee WR, Hu SL, Wu MT, Chen GS, Wong TW, et al: Genetic susceptibility to carbamazepine-induced cutaneous adverse drug reactions. *Pharmacogenet Genomics* 2006, 16:297-306.

107. Illing PT, Vivian JP, Dudek NL, Kostenko L, Chen Z, Bharadwaj M, Miles JJ, Kjer-Nielsen L, Gras S, Williamson NA, et al: Immune self-reactivity triggered by drug-modified HLA-peptide repertoire. *Nature* 2012, 486:554-558.

108. Wei CY, Chung WH, Huang HW, Chen YT, Hung SI: Direct interaction between HLA-B and carbamazepine activates T cells in patients with Stevens-Johnson syndrome. *J Allergy Clin Immunol* 2012, 129:1562-1569 e1565.

109. Allcock RJ, de la Concha EG, Fernandez-Arquero M, Vigil P, Conejero L, Arroyo R, Price P: Susceptibility to multiple sclerosis mediated by HLA-DRB1 is influenced by a second gene telomeric of the TNF cluster. *Hum Immunol* 1999, 60:1266-1273.

110. Gregersen JW, Kranc KR, Ke X, Svendsen P, Madsen LS, Thomsen AR, Cardon LR, Bell JI, Fugger L: Functional epistasis on a common MHC haplotype associated with multiple sclerosis. *Nature* 2006, 443:574-577.

111. Begovich AB, McClure GR, Suraj VC, Helmuth RC, Fildes N, Bugawan TL, Erlich HA, Klitz W: Polymorphism, recombination, and linkage disequilibrium within the HLA class II region. *J Immunol* 1992, 148:249-258.

112. Trachtenberg E, Bhattacharya T, Ladner M, Phair J, Erlich H, Wolinsky S: The HLA-B/-C haplotype block contains major determinants for host control of HIV. *Genes Immun* 2009, 10:673-677.

113. Koeleman BP, Lie BA, Undlien DE, Dudbridge F, Thorsby E, de Vries RR, Cucca F, Roep BO, Giphart MJ, Todd JA: Genotype effects and epistasis in type 1 diabetes and HLA-DQ trans dimer associations with disease. *Genes Immun* 2004, 5:381-388.

114. Parham P: MHC class I molecules and KIRs in human history, health and survival. *Nat Rev Immunol* 2005, 5:201-214.

115. Martin MP, Qi Y, Gao X, Yamada E, Martin JN, Pereyra F, Colombo S, Brown EE, Shupert WL, Phair J, et al: Innate partnership of HLA-B and KIR3DL1 subtypes against HIV-1. *Nat Genet* 2007, 39:733-740.

116. Luszczek W, Manczak M, Cislo M, Nockowski P, Wisniewski A, Jasek M, Kusnierczyk P: Gene for the activating natural killer cell receptor, KIR2DS1, is associated with susceptibility to psoriasis vulgaris. *Hum Immunol* 2004, 65:758-766.

117. Martin MP, Nelson G, Lee JH, Pellett F, Gao X, Wade J, Wilson MJ, Trowsdale J, Gladman D, Carrington M: Cutting edge: susceptibility to psoriatic arthritis: influence of activating killer Ig-like receptor genes in the absence of specific HLA-C alleles. *J Immunol* 2002, 169:2818-2822.

118. van den Elsen PJ: Expression regulation of major histocompatibility complex class I and class II encoding genes. *Front Immunol* 2011, 2:48.

119. Gobin SJ, Keijsers V, van Zutphen M, van den Elsen PJ: The role of enhancer A in the locus-specific transactivation of classical and nonclassical HLA class I genes by nuclear factor kappa B. *J Immunol* 1998, 161:2276-2283.

120. Johnson DR: Locus-specific constitutive and cytokine-induced HLA class I gene expression. *J Immunol* 2003, 170:1894-1902.

121. Gobin SJ, van Zutphen M, Westerheide SD, Boss JM, van den Elsen PJ: The MHC-specific enhanceosome and its role in MHC class I and beta(2)-microglobulin gene transactivation. *J Immunol* 2001, 167:5175-5184.

122. Gobin SJ, Biesta P, de Steenwinkel JE, Datema G, van den Elsen PJ: HLA-G transactivation by cAMP-response element-binding protein (CREB). An alternative transactivation pathway to the conserved major histocompatibility complex (MHC) class I regulatory routes. *J Biol Chem* 2002, 277:39525-39531.

123. Flajollet S, Poras I, Carosella ED, Moreau P: RREB-1 is a transcriptional repressor of HLA-G. *J Immunol* 2009, 183:6948-6959.

124. Wiendl H, Mitsdoerffer M, Hofmeister V, Wischhusen J, Bornemann A, Meyermann R, Weiss EH, Melms A, Weller M: A functional role of HLA-G expression in human gliomas: an alternative strategy of immune escape. *J Immunol* 2002, 168:4772-4780.

125. Cabestre FA, Lefebvre S, Moreau P, Rouas-Friess N, Dausset J, Carosella ED, Paul P: HLA-G expression: immune privilege for tumour cells? *Semin Cancer Biol* 1999, 9:27-36.

126. Steimle V, Otten LA, Zufferey M, Mach B: Complementation cloning of an MHC class II transactivator mutated in hereditary MHC class II deficiency (or bare lymphocyte syndrome). *Cell* 1993, 75:135-146.

127. Chou SD, Tomasi TB: Spatial distribution of histone methylation during MHC class II expression. *Mol Immunol* 2008, 45:971-980.

128. Majumder P, Gomez JA, Chadwick BP, Boss JM: The insulator factor CTCF controls MHC class II gene expression and is required for the formation of long-distance chromatin interactions. *J Exp Med* 2008, 205:785-798.

129. Gilad Y, Rifkin SA, Pritchard JK: Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet* 2008, 24:408-415.

130. Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, Dilthey A, Ellis P, Langford C, Vannberg FO, Knight JC: Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat Genet* 2012, 44:502-510.

131. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, et al: A genome-wide association study of global gene expression. *Nat Genet* 2007, 39:1202-1207.

132. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, et al: Genetics of gene expression and its effect on disease. *Nature* 2008, 452:423-428.

133. Thomas R, Apps R, Qi Y, Gao X, Male V, O'HUigin C, O'Connor G, Ge D, Fellay J, Martin JN, et al: HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. *Nat Genet* 2009, 41:1290-1294.

134. Yu HX, Chia JM, Bourque G, Wong MV, Chan SH, Ren EC: A population-based LD map of the human chromosome 6p. *Immunogenetics* 2005, 57:559-565.

135. Sayer D, Whidborne R, De Santis D, Rozemuller E, Christiansen F, Tilanus M: A multicenter international evaluation of single-tube amplification protocols for sequencing-based typing of HLA-DRB1 and HLA-DRB3,4,5. *Tissue Antigens* 2004, 63:412-423.

136. Wu YL, Savelli SL, Yang Y, Zhou B, Rovin BH, Birmingham DJ, Nagaraja HN, Hebert LA, Yu CY: Sensitive and specific real-time polymerase chain reaction assays to accurately determine copy number variations (CNVs) of human complement C4A, C4B, C4-long, C4-short, and RCCX modules: elucidation of C4 CNVs in 50 consanguineous subjects with defined HLA genotypes. *J Immunol* 2007, 179:3012-3025.

137. International HapMap C, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, et al: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007, 449:851-861.

138. Stephens M, Smith NJ, Donnelly P: A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001, 68:978-989.

139. Stephens M, Scheet P: Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet* 2005, 76:449-462.

140. McVean GA, Cardin NJ: Approximating the coalescent with recombination. *Philos Trans R Soc Lond B Biol Sci* 2005, 360:1387-1393.

141. Marchini J, Cutler D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, Qin ZS, Munro HM, Abecasis GR, et al: A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet* 2006, 78:437-450.

142. Browning SR, Browning BL: Haplotype phasing: existing methods and new developments. *Nat Rev Genet* 2011, 12:703-714.

143. Erlich RL, Jia X, Anderson S, Banks E, Gao X, Carrington M, Gupta N, DePristo MA, Henn MR, Lennon NJ, de Bakker PI: Next-generation sequencing for HLA typing of class I loci. *BMC Genomics* 2011, 12:42.

144. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, et al: Detecting recent positive selection in the human genome from haplotype structure. *Nature* 2002, 419:832-837.

145. McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P: The fine-scale structure of recombination rate variation in the human genome. *Science* 2004, 304:581-584.

146. Erlich HA, Bergstrom TF, Stoneking M, Gyllensten U: HLA Sequence Polymorphism and the Origin of Humans. *Science* 1996, 274:1552b-1554b.

147. Takahata N: Allelic genealogy and human evolution. *Mol Biol Evol* 1993, 10:2-22.

148. Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, Pritchard JK: A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* 2006, 38:1251-1260.

149. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007, 81:559-575.

150. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, et al: Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 2010, 327:78-81.

151. Carnevali P, Baccash J, Halpern AL, Nazarenko I, Nilsen GB, Pant KP, Ebert JC, Brownley A, Morenzoni M, Karpinchyk V, et al: Computational techniques for human genome resequencing using mated gapped reads. *J Comput Biol* 2012, 19:279-292.

152. Wang K, Li M, Hakonarson H: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010, 38:e164.

153. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 2011, 28:2731-2739.

154. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, et al: Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 2010, 328:636-639.

155. Patel RK, Jain M: NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* 2012, 7:e30619.

156. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 2013, 14:R36.

157. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* 2013, 31:46-53.

158. Gonzalez-Galarza FF, Christmas S, Middleton D, Jones AR: Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. *Nucleic Acids Res* 2011, 39:D913-919.

159. Thorvaldsdottir H, Robinson JT, Mesirov JP: Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013, 14:178-192.

160. Consortium IH: A haplotype map of the human genome. *Nature* 2005, 437:1299-1320.

161. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES: High-resolution haplotype structure in the human genome. *Nat Genet* 2001, 29:229-232.

162. Blomhoff A, Olsson M, Johansson S, Akselsen HE, Pociot F, Nerup J, Kockum I, Cambon-Thomsen A, Thorsby E, Undlien DE, Lie BA: Linkage disequilibrium and haplotype blocks in the MHC vary in an HLA haplotype specific manner assessed mainly by DRB1*03 and DRB1*04 haplotypes. *Genes Immun* 2006, 7:130-140.

163. de Bakker PIW, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, Ke X, Monsuur AJ, Whittaker P, Delgado M, et al: A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat Genet* 2006, 38:1166-1172.

164. Aly TA, Eller E, Ide A, Gowan K, Babu SR, Erlich HA, Rewers MJ, Eisenbarth GS, Fain PR: Multi-SNP analysis of MHC region: remarkable conservation of HLA-A1-B8-DR3 haplotype. *Diabetes* 2006, 55:1265-1269.

165. Morishima S, Ogawa S, Matsubara A, Kawase T, Nannya Y, Kashiwase K, Satake M, Saji H, Inoko H, Kato S, et al: Impact of highly conserved HLA haplotype on acute graft-versus-host disease. *Blood* 2010, 115:4664-4670.

166. Alper CA, Larsen CE, Dubey DP, Awdeh ZL, Fici DA, Yunis EJ: The haplotype structure of the human major histocompatibility complex. *Hum Immunol* 2006, 67:73-84.

167. Lu CC, Chen JC, Jin YT, Yang HB, Chan SH, Tsai ST: Genetic susceptibility to nasopharyngeal carcinoma within the HLA-A locus in Taiwanese. *Int J Cancer* 2003, 103:745-751.

168. Goldsmith DB, West TM, Morton R: HLA associations with nasopharyngeal carcinoma in Southern Chinese: a meta-analysis. *Clin Otolaryngol Allied Sci* 2002, 27:61-67.

169. Chan SH, Tan CB, Lin YN, Wee GB, Degli-Esposti MA, Dawkins RL: HLA and Singaporean Chinese myasthenia gravis. *Int Arch Allergy Immunol* 1993, 101:119-125.

170. Kumar N, Kaur G, Tandon N, Kanga U, Mehra NK: Genomic evaluation of HLA-DR3+ haplotypes associated with type 1 diabetes. *Ann N Y Acad Sci* 2013, 1283:91-96.

171. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA: Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* 2009, 324:387-389.

172. Stankiewicz P, Lupski JR: Structural variation in the human genome and its role in disease. *Annu Rev Med* 2010, 61:437-455.

173. Lai MJ, Wen SH, Lin YH, Shyr MH, Lin PY, Yang KL: Distributions of human leukocyte antigen-A, -B, and -DRB1 alleles and haplotypes based on 46,915 Taiwanese donors. *Hum Immunol* 2010, 71:777-782.

174. Lam TH, Shen M, Chia JM, Chan SH, Ren EC: Population-specific recombination sites within the human MHC region. *Heredity (Edinb)* 2013.

175. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, et al: Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998, 280:1077-1082.

176. Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Lander ES: An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 2000, 407:513-516.

177. Kumar P, Henikoff S, Ng PC: Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009, 4:1073-1081.

178. Adzhubei I, Jordan DM, Sunyaev SR: Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* 2013, Chapter 7:Unit7 20.

179. Vandiedonck C, Taylor MS, Lockstone HE, Plant K, Taylor JM, Durrant C, Broxholme J, Fairfax BP, Knight JC: Pervasive haplotypic variation in the spliceo-transcriptome of the human major histocompatibility complex. *Genome Res* 2011, 21:1042-1054.

180. Plant K, Fairfax BP, Makino S, Vandiedonck C, Radhakrishnan J, Knight JC: Fine mapping genetic determinants of the highly variably expressed MHC gene ZFP57. *Eur J Hum Genet* 2013.

181. Consortium EP, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M: An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012, 489:57-74.

182. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al: Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 2007, 39:311-318.

183. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al: Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci USA* 2010, 107:21931-21936.

184. Zhang Z, Zhang MQ: Histone modification profiles are predictive for tissue/cell-type specific expression of both protein-coding and microRNA genes. *BMC Bioinformatics* 2011, 12:155.

185. Broos S, Soete A, Hooghe B, Moran R, van Roy F, De Bleser P: PhysBinder: Improving the prediction of transcription factor binding sites by flexible inclusion of biophysical properties. *Nucleic Acids Res* 2013, 41:W531-534.

186. Yang Y, Chung EK, Zhou B, Lhotta K, Hebert LA, Birmingham DJ, Rovin BH, Yu CY: The intricate role of complement component C4 in human systemic lupus erythematosus. *Curr Dir Autoimmun* 2004, 7:98-132.

187. Szilagyi A, Fust G: Diseases associated with the low copy number of the C4B gene encoding C4, the fourth component of complement. *Cytogenet Genome Res* 2008, 123:118-130.

188. Walter NA, McWeeney SK, Peters ST, Belknap JK, Hitzemann R, Buck KJ: SNPs matter: impact on detection of differential expression. *Nat Methods* 2007, 4:679-680.

189. Zhang W, Duan S, Kistner EO, Bleibel WK, Huang RS, Clark TA, Chen TX, Schweitzer AC, Blume JE, Cox NJ, Dolan ME: Evaluation of genetic variation contributing to differences in gene expression between populations. *Am J Hum Genet* 2008, 82:631-640.

190. Ozsolak F, Milos PM: RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* 2011, 12:87-98.
191. Marguerat S, Bahler J: RNA-seq: from technology to biology. *Cell Mol Life Sci* 2010, 67:569-579.
192. Yokoi T, Miyawaki T, Yachie A, Kato K, Kasahara Y, Taniguchi N: Epstein-Barr virus-immortalized B cells produce IL-6 as an autocrine growth factor. *Immunology* 1990, 70:100-105.
193. Rochford R, Cannon MJ, Sabbe RE, Adusumilli K, Picchio G, Glynn JM, Noonan DJ, Mosier DE, Hobbs MV: Common and idiosyncratic patterns of cytokine gene expression by Epstein-Barr virus transformed human B cell lines. *Viral Immunol* 1997, 10:183-195.
194. Minoche AE, Dohm JC, Himmelbauer H: Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol* 2011, 12:R112.
195. Ewing B, Green P: Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 1998, 8:186-194.
196. Ewing B, Hillier L, Wendl MC, Green P: Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 1998, 8:175-185.
197. Hansen KD, Brenner SE, Dudoit S: Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* 2010, 38:e131.
198. Toung JM, Morley M, Li M, Cheung VG: RNA-sequence analysis of human B-cells. *Genome Res* 2011, 21:991-998.
199. Stumpf MPH, McVean GAT: Estimating recombination rates from population-genetic data. *Nat Rev Genet* 2003, 4:959-968.
200. Clark AG, Wang X, Matise T: Contrasting methods of quantifying fine structure of human recombination. *Annu Rev Genomics Hum Genet* 2010, 11:45-64.
201. Coop G, Wen X, Ober C, Pritchard JK, Przeworski M: High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* 2008, 319:1395-1398.
202. Khil PP, Camerini-Otero RD: Genetic crossovers are predicted accurately by the computed human recombination map. *PLoS Genet* 2010, 6:e1000831.
203. Ahmad T, Neville M, Marshall SE, Armuzzi A, Mulcahy-Hawes K, Crawshaw J, Sato H, Ling KL, Barnardo M, Goldthorpe S, et al: Haplotype-specific linkage disequilibrium patterns define the genetic topography of the human MHC. *Hum Mol Genet* 2003, 12:647-656.
204. International HapMap C: A haplotype map of the human genome. *Nature* 2005, 437:1299-1320.
205. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Jonasdottir A, Gylfason A, Kristinsson KT, et al: Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* 2010, 467:1099-1103.
206. Piertney SB, Oliver MK: The evolutionary ecology of the major histocompatibility complex. *Heredity (Edinb)* 2006, 96:7-21.
207. Allers T, Lichten M: Differential timing and control of noncrossover and crossover recombination during meiosis. *Cell* 2001, 106:47-57.
208. Yang KL, Chen SP, Shyr MH, Lin PY: High-resolution human leukocyte antigen (HLA) haplotypes and linkage disequilibrium of HLA-B and -C and

HLA-DRB1 and -DQB1 alleles in a Taiwanese population. *Hum Immunol* 2009, 70:269-276.

209. Hodgkinson A, Eyre-Walker A: Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* 2011, 12:756-766.

210. Martincorena I, Seshasayee AS, Luscombe NM: Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature* 2012, 485:95-98.

211. Martincorena I, Luscombe NM: Non-random mutation: the evolution of targeted hypermutation and hypomutation. *Bioessays* 2013, 35:123-130.

212. Petersdorf EW, Malkki M, Gooley TA, Martin PJ, Guo Z: MHC haplotype matching for unrelated hematopoietic cell transplantation. *PLoS Med* 2007, 4:e8.

213. Petersdorf EW, Malkki M, Gooley TA, Spellman SR, Haagenson MD, Horowitz MM, Wang T: MHC-resident variation affects risks after unrelated donor hematopoietic cell transplantation. *Sci Transl Med* 2012, 4:144ra101.

214. Morris DL, Taylor KE, Fernando MM, Nititham J, Alarcon-Riquelme ME, Barcellos LF, Behrens TW, Cotsapas C, Gaffney PM, Graham RR, et al: Unraveling multiple MHC gene associations with systemic lupus erythematosus: model choice indicates a role for HLA alleles and non-HLA genes in Europeans. *Am J Hum Genet* 2012, 91:778-793.

215. Mitsunaga S, Hosomichi K, Okudaira Y, Nakaoka H, Kunii N, Suzuki Y, Kuwana M, Sato S, Kaneko Y, Homma Y, et al: Exome sequencing identifies novel rheumatoid arthritis-susceptible variants in the BTNL2. *J Hum Genet* 2013, 58:210-215.

216. Montes-Cano M, Conde-Jaldon M, Garcia-Lozano J, Ortiz-Fernandez L, Ortego-Centeno N, Castillo-Palma M, Espinosa G, Grana-Gil G, Gonzalez-Gay M, Barnosi-Marin A, et al: HLA and non-HLA genes in Behcet's disease: a multicentric study in the Spanish population. *Arthritis Res Ther* 2013, 15:R145.

217. Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, Cheung VG: Common genetic variants account for differences in gene expression among ethnic groups. *Nat Genet* 2007, 39:226-231.

218. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT: Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 2005, 437:1365-1369.

219. Timp W, Mirsaidov UM, Wang D, Comer J, Aksimentiev A, Timp G: Nanopore Sequencing: Electrical Measurements of the Code of Life. *IEEE Trans Nanotechnol* 2010, 9:281-294.

220. Peters BA, Kermani BG, Sparks AB, Alferov O, Hong P, Alexeev A, Jiang Y, Dahl F, Tang YT, Haas J, et al: Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* 2012, 487:190-195.

221. Armitage SJ, Jasim SA, Marks AE, Parker AG, Usik VI, Uerpmann HP: The southern route "out of Africa": evidence for an early expansion of modern humans into Arabia. *Science* 2011, 331:453-456.

222. Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, Blackburn J, Semino O, Scozzari R, Cruciani F, et al: Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science* 2005, 308:1034-1036.

223. Kagnoff MF: Celiac disease: pathogenesis of a model immunogenetic disease. *J Clin Invest* 2007, 117:41-49.

224. Todd JA, Bell JI, McDevitt HO: HLA-DQ beta gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. *Nature* 1987, 329:599-604.

225. Girdlestone J: Synergistic induction of HLA class I expression by RelA and CIITA. *Blood* 2000, 95:3804-3808.

226. Choi NM, Majumder P, Boss JM: Regulation of major histocompatibility complex class II genes. *Curr Opin Immunol* 2011, 23:81-87.

227. Wright KL, Ting JP: Epigenetic regulation of MHC-II and CIITA genes. *Trends Immunol* 2006, 27:405-412.

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

229. Quenneville S, Verde G, Corsinotti A, Kapopoulou A, Jakobsson J, Offner S, Baglivo I, Pedone PV, Grimaldi G, Riccio A, Trono D: In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. *Mol Cell* 2011, 44:361-372.

230. Huang C, Jia Y, Yang S, Chen B, Sun H, Shen F, Wang Y: Characterization of ZNF23, a KRAB-containing protein that is downregulated in human cancers and inhibits cell cycle progression. *Exp Cell Res* 2007, 313:254-263.

231. Segurel L, Leffler EM, Przeworski M: The case of the fickle fingers: how the PRDM9 zinc finger protein specifies meiotic recombination hotspots in humans. *PLoS Biol* 2011, 9:e1001211.

232. Silva FP, Hamamoto R, Furukawa Y, Nakamura Y: TIPUH1 encodes a novel KRAB zinc-finger protein highly expressed in human hepatocellular carcinomas. *Oncogene* 2006, 25:5063-5070.

233. Yokoe T, Toiyama Y, Okugawa Y, Tanaka K, Ohi M, Inoue Y, Mohri Y, Miki C, Kusunoki M: KAP1 is associated with peritoneal carcinomatosis in gastric cancer. *Ann Surg Oncol* 2010, 17:821-828.

234. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, et al: Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 2002, 8:816-824.

235. Tang M, Zeng Y, Poisson A, Marti D, Guan L, Zheng Y, Deng H, Liao J, Guo X, Sun S, et al: Haplotype-dependent HLA susceptibility to nasopharyngeal carcinoma in a Southern Chinese population. *Genes Immun* 2010, 11:334-342.

236. Yu KJ, Gao X, Chen CJ, Yang XR, Diehl SR, Goldstein A, Hsu WL, Liang XS, Marti D, Liu MY, et al: Association of human leukocyte antigens with nasopharyngeal carcinoma in high-risk multiplex families in Taiwan. *Hum Immunol* 2009, 70:910-914.

237. Hildesheim A, Apple RJ, Chen C-J, Wang SS, Cheng Y-J, Klitz W, Mack SJ, Chen I-H, Hsu M-M, Yang C-S, et al: Association of HLA class I and II alleles and extended haplotypes with nasopharyngeal carcinoma in Taiwan. *J Natl Cancer Inst* 2002, 94:1780-1789.

238. Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB: Annotating non-coding regions of the genome. *Nat Rev Genet* 2010, 11:559-571.

239. Manikandan J, Aarthi JJ, Kumar SD, Pushparaj PN: Oncomirs: the potential role of non-coding microRNAs in understanding cancer. *Bioinformation* 2008, 2:330-334.

240. Croce CM: Causes and consequences of microRNA dysregulation in cancer. *Nat Rev Genet* 2009, 10:704-714.

241. Wapinski O, Chang HY: Long noncoding RNAs and human disease. *Trends Cell Biol* 2011, 21:354-361.

List of Publications

- **Lam TH**, Shen M, Chia JM, Chan SH, Ren EC. Population-specific recombination sites within the human MHC region. *Heredity*, 11(2):131-8, 2013.
- Wang B, Niu D, **Lam TH**, Xiao Z, Ren EC. Mapping the p53 transcriptome universe using p53 natural polymorphs. *Cell Death Differ*, 2013.
- **Lam TH,** Tay M, Wang B, Xiao ZW, Ren EC. Fine genomic variation within conserved extended haplotypes of human MHC detected by whole genome sequencing. (Submitted)