What has GWAS Done for HLA and Disease Associations?

Amy E. Kennedy [1], Umut Ozbek [2,3], Mehmet T. Dorak [4]
[1] Center for Research Strategy, National Cancer Institute, National Institutes of Health, Bethesda, MD, U.S.A.
[2] Tisch Cancer Institute, Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY, U.S.A.
[3] Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY, U.S.A.
[4] School School of Life Sciences, Pharmacy & Chemistry, Kingston University London, Kingston KT1 2EE, U.K.

Correspondence:
Prof. M. T. Dorak
School of Life Sciences, Pharmacy & Chemistry
Kingston University London
Penrhyn Road
Kingston KT1 2EE
U.K.
e-mail: m.dorak@kingston.ac.uk

Abstract

The major histocompatibility complex (MHC) is located in chromosome 6p21 and contains crucial regulators of immune response, including human leukocyte antigen (HLA) genes, alongside other genes with non-immunological roles. More recently, a repertoire of non-coding RNA genes, including expressed pseudogenes, has also been identified. The MHC is the most gene-dense and most polymorphic part of the human genome. The region exhibits haplotype-specific linkage disequilibrium patterns, contains the strongest cis- and trans-eQTLs/meQTLs in the genome, and is known as a hot spot for disease associations. Another layer of complexity is provided to the region by the extreme structural variation and copy number variations. While the *HLA-B* gene has the highest number of alleles, the HLA-DR/DQ subregion is structurally most variable, and shows the highest number of disease associations. Reliance on a single reference sequence has complicated the design, execution and analysis of GWAS for the MHC region and not infrequently, the MHC region has even been excluded from the analysis of GWAS data. Here, we contrast features of the MHC region with the rest of the genome, and highlight its complexities, including its functional polymorphisms beyond those determined by single nucleotide polymorphisms or single amino acid residues. One of the several issues with customary GWAS analysis is that it does not address this additional layer of polymorphisms unique to the MHC region. We highlight alternative approaches that may assist with the analysis of GWAS data from the MHC region and unravel associations with all functional polymorphisms beyond single SNPs. We suggest that despite already showing the highest number of disease associations, the true extent of the involvement of the MHC region in disease genetics may not have been uncovered.

Introduction

The major histocompatibility complex (MHC) is present in all mammals, but was first discovered in tumour transplantation studies in mice (Gorer, 1937). The human MHC, which is called human leukocyte antigens (HLA) complex, was discovered independently by Dausset, van Rood, and Payne & Bodmer during studies of antibodies against leukocytes in multiparous women (Dausset, 1981). Like blood group antigens (Aird, Bentall, & Roberts, 1953), HLA antigens are among the first genetic markers examined for disease associations (Amiel, 1967). While the emphasis was initially on the histocompatibility products (HLA-A, -B, -C, -DR, -DQ, -DP), the Human Genome Project unravelled the true content of the MHC region, and over the last decade, genome-wide association studies (GWAS) have unravelled a large number of disease associations with MHC region variants.

The MHC is the most gene-dense region of the human genome (**Table 1**), containing a diversity of genes involved in major physiologic phenomena (Roger Horton et al., 2004; Shiina, Hosomichi, Inoko, & Kulski, 2009; Vandiedonck & Knight, 2009; T. Xie et al., 2003). The region is clearly enriched for genes encoding molecules participating in immune and inflammatory pathways (R. Horton et al., 2008; Roger Horton et al., 2004; Shiina et al., 2009; Traherne, 2008; Trowsdale & Knight, 2013) (**Table 2**), and about 60% of the gene content is involved in non-immunological roles (Shiina et al., 2009; Vandiedonck & Knight, 2009). Of the total 677 genes, there are still 60 without sufficient characterization (open reading frames or uncharacterised loci). The extended MHC (xMHC) contains 1.5% of the genes in OMIM and 6.4% of genome-wide significant single nucleotide polymorphism (SNP) associations in the NHGRI/EBI GWAS catalog (Ripke et al., 2013). Before the GWAS era, the list of traits and diseases that show MHC associations in mammals was already very long and included a variety of conditions from reproductive issues (Kostyu, 1994; Lerner & Finch, 1991), to cancer (Chaudhuri et al., 2000; de Jong et al., 2003; DeWolf, Lange, Einarson, & Yunis, 1979; Diepstra et al., 2005; M. T. Dorak et al., 1999; Klitz, Aldrich, Fildes, Horning, & Begovich, 1994; Lu et al., 1990; Magnusson et al., 2001) and longevity (Ivanova et al., 1998). Of particular interest is the mapping of breast cancer (de Jong et al., 2003) and Hodgkin lymphoma susceptibility (Diepstra et al., 2005) to the MHC class III region, which is devoid of classical HLA genes. During the GWAS era, disease associations with MHC region variants have drastically increased (Lenz, Spirin, Jordan, & Sunyaev, 2016; Ripke et al., 2013; Vandiedonck & Knight, 2009) (**Figure 1**). GWAS have identified SNP level associations for most robustly validated HLA associations (**Table 3**). Autoimmune disorders have always shown strong and consistent MHC associations (Matzaraki, Kumar, Wijmenga, & Zhernakova, 2017) some of which have been even narrowed down to amino acid level (Achkar et al., 2012; Miyadera, Ohashi, Lernmark, Kitamura, & Tokunaga, 2015; Raychaudhuri et al., 2012), but detailed studies also show independent associations with non-HLA polymorphisms in the MHC region (Handunnetthi, Ramagopalan, Ebers, & Knight, 2010; Rioux et al., 2009).

Progress has been made in understanding the mechanisms of several HLA-associated diseases (Caillat-Zucman, 2009; Sollid, Pos, & Wucherpfennig, 2014), especially in rheumatoid arthritis (Klareskog, Catrina, & Paget, 2009), type 1 diabetes and Celiac disease (Busch et al., 2012), as well as drug hypersensitivities (Illing, Vivian, Purcell, Rossjohn, & McCluskey, 2013). However, given the number of disease associations with MHC region variants, only a small number of potential mechanisms have been uncovered (Howell, 2014).

Unique Features of the Extended MHC Region

The HLA region has many unique features which distinguish it from the rest of the genome (**Box 1**). The two most gene-dense regions of the human genome are the MHC class III region and the histone gene supercluster within the xMHC region (T. Xie et al., 2003). Besides being the most gene-dense, xMHC is also the most polymorphic region in the genome, which is compounded by a complex linkage disequilibrium (LD) structure. In the xMHC, individual SNPs and haplotypes are as relevant as elsewhere, additionally, constellations of them make up HLA alleles, groups of HLA alleles form functional supertypes or ancestral supertypes, and mark evolutionary lineages. GWAS analysis of more recent studies included HLA imputation and analysis by HLA type or amino acid residues (A. Dilthey et al., 2013; A. T. Dilthey, Moutsianas, Leslie, & McVean, 2011; Jia et al., 2013; M. Xie, Li, & Jiang, 2010; X. C. Zhang, Li, Wang, Hansen, & Zhao, 2011), but no GWAS to date has specifically examined the MHC region for functional polymorphisms such as highly functional epitopes (including HLA-Bw4/Bw6 or HLA-C1/C2) that have previously shown important disease associations (discussed later).

An important difference in the xMHC from the rest of the genome concerns the structural variation and the presence of closely related genes in these regions. Two such regions contain the complement (*C2*, *C4A*, *C4B* in a CNV region) and *HLA-DRB* (B1-B9) genes (a segmental duplication region). The polymorphisms in these regions are particularly difficult to genotype using high-throughput methods, mainly because they violate Hardy-Weinberg equilibrium (HWE) due to the presence of paralogs and CNV. An analysis of the whole genome sequencing data from the 1000 Genomes Project revealed that the MHC region shows the highest Hardy-Weinberg disequilibrium levels (Graffelman, Jain, & Weir, 2017). Most of this must be due to the genomic features of this region. SNPs that do not conform to HWE are excluded from GWAS chips at the time of quality control steps during the production phase. At the time of the first wave of GWAS, the Illumina MHC SNP Panel was a dedicated SNP typing platform for 2360 SNPs in the xMHC. The first-generation GWAS chips typically covered the xMHC, at most, as extensively as this panel. The two segmental duplication regions of the MHC were grossly underrepresented and the HapMap phase I data contained genotype data for considerably fewer SNPs in these regions: (i) complement subregion: only four SNPs in the 78.5kb subregion covering the *C4A*, *C4B* and *CYP21A2* genes within the class III region; (ii) HLA-DR subregion: just four SNPs in the 114.5kb subregion covering the area flanked by *HLA-DRA* and *-DRB1* genes. Extreme variation of this region, including structural variation and presence of paralogs, creates difficulties for inclusion of many SNPs in GWAS chips, but alternative genotyping methods exist (M.T. Dorak, 2007). Imputation of the SNPs in the subregions that are not covered sufficiently in GWAS chips may be thought of as a solution. The success of imputation, however, depends on the reference panel used, and even 1KG data may not be ideal for these structural variation regions, which also suffers from SNP genotype call difficulties for the MHC (Brandt et al., 2015).

Box 1. Unique features of the xMHC relevant to GWAS:

- Most gene dense in the genome (T. Xie et al., 2003)
- Paralog regions and genes (one-third of the genes residing in the MHC have paralogous copies) (Endo, Imanishi, Gojobori, & Inoko, 1997; Roger Horton et al., 2004; Kasahara, 1999a, 1999b; Kasahara et al., 1996; Katsanis, Fitzgibbon, & Fisher, 1996; Shiina et al., 2001)
- Clustering of functionally related genes (Roger Horton et al., 2004; Trowsdale & Knight, 2013)
- Strongest trans-eQTLs (Fairfax et al., 2012; Fehrmann et al., 2011; Westra et al., 2013) and meQTLs (van Dongen et al., 2016) in the genome as well as an exceptional number of connected components in genotype networks (Dall'Olio, Bertranpetit, Wagner, & Laayouni, 2014)
- CNV and structural variation (Andersson, 1998; Blanchong et al., 2000; Y. B. Zhang, Li, Zhang, Wang, & Yu, 2012)
- Extremely polymorphic at the nucleotide level (Durbin et al., 2010; Gaudieri, Dawkins, Habara, Kulski, & Gojobori, 2000; Vandiedonck & Knight, 2009)
- Highest trait-associated variant density even by standard analysis of GWAS data (treating the xMHC as anywhere else in the genome) (Lenz et al., 2016; Ripke et al., 2013; Vandiedonck & Knight, 2009)
- Non-HLA genes throughout xMHC carry deleterious variants at high frequencies (more than two orders of magnitude above the genome-wide average for some of them) (Lenz et al., 2016).
- Very high linkage disequilibrium over very long range resulting from conserved extended haplotypes (Ahmad et al., 2003; Aly et al., 2006; Blomhoff et al., 2006; T. M. S. Consortium, 1999) due to lower recombinational rates than the rest of the genome (3-fold lower than 1.2cM/Mb) (de Bakker et al., 2006)
- Higher than average rates of alternative splicing as a manifestation of DNA sequence diversity (Vandiedonck et al., 2011)

- Highest gene expression levels across the genome *; highest heritability of gene expression levels (Wright et al., 2014); and trans-generational inheritance of methylation patterns (McRae et al., 2014)

  * The GTEx database lists *HLA-B* (9th), *HLA-C* (15th), *HLA-E* (19th) and *HLA-A* (37th) in the top 50 genes for expression levels (the beta2-microglobulin gene *B2M* is 19th).

CNV within the MHC may not have received sufficient attention. HLA-B/C and HLA-DR/DQ subregions are within CNVs of rather large fragments. The Database of Genetic Variants currently (March 2017) lists six CNVs within xMHC larger than 1Mbp, and 41 CNVs between 0.2 and 1Mbp. However, their correlations with disease susceptibility are not well studied. CNVs correlate with transcription levels of genes within the CNV region, including MHC genes (Schlattl, Anders, Waszak, Huber, & Korbel, 2011). The HLA class II region contains CNVs at appreciable population frequencies and with sizes reaching up to 421,697bp (Conrad et al., 2010; de Smith et al., 2007). The largest class II region CNV (NCBI 36.1 ID: Variation_64476) reported by Conrad *et al* in the HapMap population as a gain with a frequency of 16.1% spans a region containing the genes *HLA-DRA*, *-DRB1, -DRB5, -DRB6*, *-DQA1*, *-DQB1*, *-DQA2*, and *-DQB2* (Conrad et al., 2010).

Structural variation in the MHC is not restricted to CNVs. Segmental duplications are neighbouring duplicated genomic segments that are large (at least >1 kb in length), and that show more than 90% sequence identity. The class III region has a typical segmental duplication called the RCCX module (*RP1/2-C4A/B-CYP21A1P/A2-TNXA/B*) and contains the *C4* and *CYP21* genes, and may be present in more than one copy on each chromosome (Blanchong et al., 2000). The RCCX module copy number, each module's size and gene content, and each C4 gene's size on each HLA haplotype is variable (Blanchong et al., 2000; Collier et al., 1989; Y. L. Wu et al., 2007; Y. L. Wu et al., 2008). GWAS is unable to detect any of these variations, and as has been pointed out elsewhere (Traherne, 2008), there is no known proxy SNP for any of these alterations. The only complementary study that has been done to follow up any GWAS specifically for this region is in schizophrenia and identified copy number variation of *C4* genes as the causal variation (Sekar et al., 2016).

The frequency of a partial C4A or C4B protein deficiency in the Caucasian populations is between 25.5 and 33.5% (Hauptmann, Tappeiner, & Schifferli, 1988) making partial C4 deficiency the most common immune protein deficiency in humans. *C4A* deletion is also an established risk marker for systemic lupus erythematosus (Y. L. Wu et al., 2008) as well as other autoimmune disorders and infectious diseases (Hauptmann et al., 1988). Still, no GWAS chip contains a single marker for *C4A* deletion, or, in fact, no more than a few markers from the whole *C4A/C4B* genes. A *C4A* deletion does not necessarily mean physical deletion of the gene. Nevertheless, no form of *C4A* deletion is represented by any polymorphisms on current GWAS chips.

The complement proteins have common and useful polymorphisms that were used to define complotypes as constellations of complement components C2, factor B, C4A and C4B. The complotypes were considered as the most informative molecular markers defining the common HLA haplotypes in the 1980s and 1990s (Simon et al., 1997; Whitehead et al., 1984). Polymorphisms of these genes were used to determine HLA haplotype identity in HLA-matched transplant pairs [Dorak, 1993 #2467. However, once PCR-based genotypings took over, interest in complotypes faded. Since 2010, only nine papers have been published with complotype included in their titles or abstracts. It appears that the usefulness of complotypes disappeared along with our inability to genotype them by high-throughput methods. High-throughput methods may be superior on average, but their deficiency for polymorphisms in genes that have paralogs and are in CNV regions is obvious (M. Li, Li, & Guan, 2008).

Another major structural variation in the MHC concerns the HLA-DR/DQ region. This region always contains the *HLA-DRA* and *-DRB1* genes encoding the alpha and beta chains of the HLA-DR molecule. However, on most haplotypes, there is a second expressed HLA-DRB gene, which may be *-DRB3*, *-DRB4*, or *-DRB5* (Andersson, 1998). These second expressed DRB genes, however, are mutually exclusive and only one of them can be on a haplotype. The pseudogene *HLA-DRB9*, which is a duplication copy of the ancestral DRB gene, is present on all haplotypes, but other pseudogenes such as *DRB2*, *DRB7* and *DRB8* only exist on certain haplotypes. The reference sequence of the MHC has derived from the PGF cell line (*HLA-DRB1*1501*) used in the MHC Haplotype Project (R. Horton et al., 2008). Its haplotype contains the *HLA-DRB5* gene (encoding DR51 serotype). The *DRB5* gene is missing in more than 80% of chromosomes in European origin populations, but it is featured in HLA-DR region maps as if constantly present. Until now, any SNP mapping to the coordinates of the *HLA-DRB5* gene has been considered a SNP in this gene, despite that the individual may even be missing this gene, and may have *-DRB3* or *-DRB4* in its place. Another implication is that the *-DRB3* or *-DRB4* genes are themselves polymorphic genes (Robbins et al., 1997), but their variants are not included in GWAS chips as they do not map to the reference sequence.

The structural variation in this region is not currently considered in the analysis of data. Two problems arise that may result in the loss of otherwise useful information. First is the presence of duplication products of the ancestral DRB gene confounds genotyping, and secondly, the presence of a SNP in the paralogous position (i.e., a pseudoSNP) may result in excess heterozygosity and violation of HWE (Leal, 2005), as has been specifically documented for the xMHC SNPs in the 1KG data (Brandt et al., 2015). These difficulties for genotyping are in addition to the extreme polymorphism of the region which makes it very difficult to design typing assays.

The number of closely related DRB genes is six (DRB1/4/6/7/8/9) on the *HLA-DRB1*0401* haplotype. The overall structural variation creates an anomalous situation in that some SNP positions may not even be present in some haplotypes. The presence/absence polymorphisms may result in low rates of genotype calls and subsequently exclusion of polymorphism data when in fact the missing genotype is the perfectly natural consequence of a missing gene. This is not taken into account in the analysis of data; in fact, such data do not exist as SNPs of this type would be excluded from the microarrays at the quality control step. The current NCBI SNP Database lists almost 4,000 SNPs mapping to the *HLA-DRB5* gene, which is included in the reference sequence, but none have shown any disease associations. At least through LD, some of these would be expected to be associated with disease if they were included in GWAS chips and passed the quality control steps. Due to not taking into the structural variation, these SNPs would have violated HWE, and would be excluded from chips. An inspection of the ImmunoChip SNP content reveals a total lack of SNPs in a region that is larger than 50kb corresponding to the second expressed DRB gene region.

The xMHC region is also very rich in paralog genes as a result of genomic duplications in the past, which are common events (Abi-Rached, Gilles, Shiina, Pontarotti, & Inoko, 2002; Flajnik & Kasahara, 2010; Kasahara, 1999a; Katsanis et al., 1996). Nearly one-third of the genes residing in the MHC have paralogous copies in at least one of the three regions established to be paralogous to MHC on 9q33-q34, 1q21-q25/1p11-p32, and 19p13.1-p13.3 (Roger Horton et al., 2004; Kasahara, 1999a; Shiina et al., 2001). An example of paralogy within the MHC is the *CYP21A2* gene, which is adjacent to its pseudogene *CYP21A1P*. Very high sequence similarity between these two paralogs complicates genotyping efforts. *CYP21A2* encodes 21-hydroxylase and is the cause of the most common autosomal recessive condition of childhood, congenital adrenal hyperplasia (CAH; OMIM 201910). 21-hydroxylase is involved in adrenal sex steroid biosynthesis and is likely to play a role in hormonally mediated conditions, which may include breast cancer (Woolcott et al., 2010; X. Zhang, Tworoger, Eliassen, & Hankinson, 2013). No GWAS on any condition has ever examined any polymorphism of *CYP21A2* and any data on polymorphisms of this gene have been generated by conventional methods as is routinely done in medical genetics laboratories. The most common mutation of *CYP21A2* that is involved in late-onset CAH is V282L (rs6471), which is listed in dbSNP with some data showing the mutant allele frequency up to 0.540 in some populations obviously due to genotyping error. The problem with *CYP21A2* genotyping by high-throughput methods is specifically due to the interference by its pseudogene *CYP21A1P* that lies adjacent to the active gene.

Another example of paralogy is the heat shock protein (HSP) genes *HSPA1A*, *HSPA1B* and *HSPA1L*. These three genes are extremely similar in their sequences, and part of a large HSP superfamily (Calderwood & Ciocca, 2008). As a result, their genotyping is extremely difficult

(Contreras-Sesvold, Sambuughin, Blokhin, & Deuster, 2010) and almost impossible with high-throughput methods. This must be why the *HSPA1B* SNP rs1061581 has never been examined in GWAS despite its replicated associations with susceptibility to cancer in candidate gene studies (Guo et al., 2011; Ucisik-Akkaya, Davis, Gorodezky, Alaez, & Dorak, 2010).

xMHC Region Associations in GWAS

Although the earliest serological associations with autoimmune diseases stood the test of time (Brewerton et al., 1973; Stastny, 1978), due to the presence of many inconclusive and inconsistent reports, pre-GWAS era HLA-disease associations were often met with some scepticism. It is unfortunate that most of those HLA association studies indeed did not conform to the current standards of genetic epidemiological research, and may have suffered from small sample size, methodological imperfections including HLA typing errors, disregard of population structure, and lack of replication.

GWAS have unravelled many unsuspected susceptibility markers for many traits (Manolio, 2013). GWAS have achieved much more than candidate gene studies in terms of identifying genotype-phenotype correlations. However, there is still a degree of disappointment with the cumulative results; only a modest amount of disease heritability has been explained, even after multiple studies targeting the same disease (Maher, 2008; Manolio et al., 2009).

It is generally assumed that GWAS provide approximately uniform representation of the entire genome. However, the xMHC, which accounts for a disproportional number of disease associations, is underrepresented in GWAS chips. Still, GWAS have reported many top hits within the xMHC in a variety of disorders and traits with or without an immune basis. Most notably, the strongest markers for drug hypersensitivities have been located within the MHC, and several have been FDA approved for clinical use (Profaizer & Eckels, 2012).

Cancer susceptibility is historically linked to the histocompatibility loci. The earliest disease susceptibility study in animals examining MHC effects highlighted its role in leukaemia in mice (Lilly, Boyse, & Old, 1964) followed by other cancers (Oomen, Van der Valk, & Den Engelse, 1983) including breast (Dux & Demant, 1987; Muhlbock & Dux, 1974; Ropcke, Moen, Hart, & Demant, 1990) and lung cancer (Demant, Oomen, & Oudshoorn-Snoek, 1989; Oomen et al., 1983; Snoek et al., 2000). Those studies were not limited to virally-induced leukaemia and mammary tumours, but also examined spontaneous, chemically- and hormonally-induced tumors. Until the GWAS era, replicated associations were few and far in between. In the GWAS era, robust associations have emerged in lung cancer (Broderick et al., 2009; Guo et al., 2011; Y. Wang et al., 2008), breast cancer (Michailidou et al., 2015), prostate cancer (Kote-Jarai et al., 2011), testicular germ cell tumour (Rapley et al., 2009), liver cancer (Kumar et al., 2011), multiple myeloma (Chubb et al., 2013), Hodgkin lymphoma (Moutsianas et al., 2011; Urayama et al., 2012), follicular lymphoma (Conde et al., 2010), nasopharyngeal carcinoma (Tse et al., 2009), cervical cancer (Chen et al., 2013), and glioma (Bethke et al., 2008). As in other diseases, with increasing use of rare variants, much larger sample sizes and meta-analysis approaches in association studies, more associations are being reported (Fitzgerald et al., 2013; Haiman et al., 2013; Kuchenbaecker et al., 2015; Timofeeva et al., 2012; C. Wu et al., 2014).

GWAS have shown associations of xMHC variants not only with autoimmune disorders and infectious diseases (Chapman & Hill, 2012; Handunnetthi et al., 2010; Rioux et al., 2009) as expected, but also with a diverse set of other diseases such as Barrett esophagus (Su et al., 2012), metabolic disorders (Chasman et al., 2009), obesity (Thorleifsson et al., 2009), schizophrenia (S. W. G. o. t. P. G. Consortium, 2014; Ripke et al., 2013; Sekar et al., 2016), Parkinson disease (Nalls et al., 2011), age-related macular degeneration (Cipriani et al., 2012), drug hypersensitivities (Profaizer & Eckels, 2012), and even with educational attainment (Rietveld et al., 2013) and wine preference (Pirastu et al., 2015). The potential reasons for such a large number of xMHC associations with a variety of traits are listed in **Box 2**.

Box 2. Potential reasons for disproportionate number of disease associations with xMHC region SNPs

- Extreme polymorphism (Durbin et al., 2010; Gaudieri et al., 2000)
- Extreme diversity of gene content (Roger Horton et al., 2004; Shiina et al., 2009)
- Pleiotropic (immune and non-immune) functions of HLA molecules (Hassan & Mourad, 2011; Truman, Garban, Choqueux, Charron, & Mooney, 1996)
- Selection acting on HLA loci and hitchhiking of deleterious alleles with them (Lenz et al., 2016; Mathieson et al., 2015)
- Presence of strongest trans-eQTLs (Fehrmann et al., 2011; Westra et al., 2013) and meQTLs (van Dongen et al., 2016) in the genome
- Effect of HLA alleles on the microbiome (Kubinak et al., 2015; Marietta, Rishi, & Taneja, 2015)

Another point relevant to any discussion of the xMHC in the pathogenesis of any disease, and to possible explanation of extra-ordinarily large numbers of disease associations with its variants is the trans-eQTL effects of xMHC SNPs (Fairfax et al., 2012; Fehrmann et al., 2011; Westra et al., 2013). It appears that the effect of xMHC SNPs on gene transcription extends well beyond the genes nearby, to genes on other chromosomes. With a recent large twin study showing that a substantial proportion of gene expression heritability is trans to the structural gene (Grundberg et al., 2012), the trans-eQTL effects of xMHC polymorphisms may be one of the mechanisms of their diverse disease associations (Fairfax et al., 2012; Fehrmann et al., 2011). Likewise, in the BIOS QTL Browser (van Dongen et al., 2016), the strongest meQTLs are xMHC variants overlapping with the strongest trans-eQTLs. Thus, xMHC is not only the most gene-dense and polymorphic region, but its polymorphisms also correlate with expression and methylation levels of distant genes. The high density of eQTLs and meQTL in xMHC may be the reason for the observation that xMHC genes have the highest number of genotype network across the genome (Dall'Olio et al., 2014).

What GWAS Could Have Shown

It is clear that GWAS have unravelled many unexpected associations throughout the genome including the xMHC. GWAS catalogue and other similar databases list thousands of associations from the xMHC, but their independence from one another and from HLA types already known to be associated with the same trait is not always examined. Different platforms use different sets of SNPs and the reported associations in the same trait may even be identical due to strong LD between the associated markers. There is currently no simple way of checking whether a SNP association corresponds to an already known HLA association although available HLA types together with genome-wide SNP genotypes from 1KG and HapMap samples may provide some clues (Erlich et al., 2011; Gourraud et al., 2014; Major, Rigo, Hague, Berces, & Juhos, 2013). Since imputation is now a common practice, the associations with imputed SNPs add another complication to the interpretation of xMHC associations. The best reports consist of examinations of LD between the reported marker (the lead SNP) and other known associations in the same region, imputations of HLA types and adjustments by them to check the independence of the SNP association, and a full imputation and association statistics. While some studies worked out the correlations at the time of publications, some earlier GWAS were not analysed comprehensively enough and a lot of associations reported as top hits could have been better scrutinised.

What GWAS has achieved has achieved is generally considered impressive, but more could have been done for the analysis of xMHC polymorphisms. At present, there are more than 16,000 HLA alleles (**Table 1**). Typing at this high-resolution level polymorphism is crucial for transplantation success, and as an aid in clinical diagnosis of certain disorders and drug toxicities. However, there are much simpler polymorphisms that have a huge impact on the physiological roles of the HLA proteins. Most well-known such polymorphisms concern HLA class I codons 114 and 116, HLA-Bw4/w6 and HLA-C1/C2 epitopes, -DRB1 codon 86, -DQB1 codon 57 and -DPB1 codon 56 (**Figure 2**).

These polymorphisms are generated by multiple nucleotide substitutions and cannot be identified by simple SNP typing. In addition, there are also phylogenetic groups (such as the DR53 family consisting of *HLA-DRB1*04*, **07* and **09*), cross-reactive groups (CREGs) of HLA class I alleles, and functional supertypes of both HLA class I and II alleles. Of these, HLA-Bw4/w6, HLA-C-C1/C2, -DRB1 codon 86, -DQB1 codon 57, and -DPB1 codon 56 are dimorphisms, which divide the alleles of the given locus into two mutually exclusive and collectively exhaustive groups. These polymorphisms show strong associations with diseases and their associations cannot be assessed by analysis of individual SNPs or individual HLA alleles. Special considerations are needed for their assessment and that has not been done in any genetic association study, including GWAS, to date.

There are already recognized associations with diseases or physiologic traits of some of these dimorphisms and other broad groupings, but recent studies have not recognized their value due to the shift to emphasis on individual SNP associations. With few exceptions such as *HLA-B*5701*, -*DRB1*1501* and the DR53 lineage, no single SNP is currently known to represent either a single HLA allele or any of the functional HLA groups. Besides those shown in Figure 2, there are additional sequence feature variant types (SFVT) which overlap with some of the groups shown in Figure 2. These SFVTs, when taken into account, can explain disease associations better than HLA types themselves (Karp et al., 2010; Thomson et al., 2010), but none of these functional groups are deliberately examined for their associations with disease in GWAS. Since there are no SNP proxies, which are likely to be constellations of SNPs rather than single SNPs, the only way to analyse associations with these functional groups is to impute HLA types and infer the SFVTs and other specificities for comparison between cases and controls.

The classical MHC functional groups can be inferred from HLA types. In current practice, however, neither HLA association studies nor GWAS -following the prediction of HLA types by recently developed algorithms(Karnes et al., 2017)- routinely examine associations of functional HLA groups. This may be due to aiming to keep the number of statistical comparisons to a minimum, or the lack of awareness. Dedicated genotyping assays are available for the better known dimorphisms: HLA-Bw4/Bw6 (Bari et al., 2011; Ugolotti et al., 2011; Yun et al., 2007) and HLA-C1/C2 (Bari et al., 2011; Schellekens et al., 2007; Ugolotti et al., 2011; Yun et al., 2007) as well as for *HLA-DPB1* (Cano & Fernandez-Vina, 2009) dimorphisms, which can be used in secondary studies following GWAS.

Currently, the HLA region is treated the same as any other region in the genome in GWAS data analysis, if not excluded from data analysis (see for example, Ref. (Deelen et al., 2014)). The HLA region has unique characteristics that need to be considered in data analysis. The most popular multiallelic HLA grouping currently in use for disease association studies is the DRB1 alleles bearing the “shared epitope” relevant in rheumatoid arthritis aetiology. *HLA-DRB1* alleles with amino acid sequences QKRAA, QRRAA, or RRRAA at positions 70-74 (shared epitope) are usually analysed as a single cluster in RA association studies (Barnetche, Constantin, Cantagrel, Cambon-Thomsen, & Gourraud, 2008; Bax, van Heemst, Huizinga, & Toes, 2011). This epitope exists on eight *HLA-DRB1* alleles (04:01, 04:04, 04:05, 04:08, 01:01, 01:02, 09 and 10:01). These alleles are usually grouped together in the analysis based on the HLA typing data. Likewise, *HLA-B/C* typing data are used to infer the HLA-Bw4 / Bw6 (Martin et al., 2002) and C1 / C2 epitope (Martin et al., 2010) status in certain disease association studies. These epitopes are not characterized by a single or a few SNPs but are possessed by heterogeneous groups of HLA alleles. As is currently done, GWAS data analysis does not detect associations with these epitopes.

Statistical Analysis of the xHLA Region GWAS Data

The xMHC region is currently analysed as anywhere else in the genome in GWAS. From the routine use of the additive model to the traditional thresholds for statistical significance, this approach is potentially counter-productive for detecting associations in this region. Besides, the unique features of xMHC previously discussed need to be taken into account for most effective analysis of the data from this region.

Confounding by genomic features of the xHLA region

When the unique features of the HLA region are not taken into account in GWAS analysis, a lot of data may be wasted. Dismissal of SNPs due to violation of HWE resulting from the presence of paralogs or CNV, and low genotype call rates because of structural variation (absence/presence polymorphisms) are a couple of examples of loss of valuable data. Most of these SNPs are eliminated during the SNP selection process for the GWAS chip, and if they make it to the chip, they face a similar outcome at the analysis phase. An example is the common deletion of the *C4A* gene. Currently, the number of *C4A* genes in the diploid genome is not determined prior to the analysis and all samples with zero (rare), one, two or more copies of *C4A* gene are analysed together. We believe the resulting problems with HWE are the main reasons for lack of data from the complement region of the MHC. SNPs in such regions can be genotyped by alternative methods to assess their contribution to disease risk. Thus, as far as the MHC region is concerned, no GWAS is truly genome-wide until highly functional regions of the MHC region are genotyped by complementary approaches.

Besides exclusion of SNPs due to genomic features of the region, the reliance on a single reference sequence based on just one HLA haplotype is also problematic. All HLA haplotypes are different in their length due to variable gene content, and the differences can be very large (R. Horton et al., 2008). This issue has recently been addressed and a new method based on a population reference graph for analysis of HLA region data is introduced (A. Dilthey, Cox, Iqbal, Nelson, & McVean, 2015). This method takes into account the sequencing data from eight common HLA haplotypes. While expected to be of primary use for mapping sequencing reads, it may also help with interpretation of the genotyping results where the current set of reference sequences is substantially incomplete (A. Dilthey et al., 2015).

Linkage disequilibrium

LD in the genome is important for the success of association studies and in the interpretation of results. High LD regions pose difficulty in identification of causal variants among the statistically similar SNP (ssSNP) set that has generated the association signal. LD varies in different parts of the genome and among populations, sometimes causes associations to disappear in a replication study or even to change their directions because of high correlation. High and long-range LD is interpreted as one of the hallmarks of the MHC (T. M. S. Consortium, 1999) with some haplotypes being better known for their long range LD than others.

LD extends over larger physical distances in xMHC than in the rest of the genome (31.1 kb versus 22.3 kb), but these blocks are shorter in genetic distances (0.012 cM versus 0.017 cM) (Vandiedonck & Knight, 2009). While on average the extent of LD may appear to be similar to elsewhere in the genome, a haplotype-specific LD variation has long been known (Ahmad et al., 2003; Cullen, Perfetto, Klitz, Nelson, & Carrington, 2002; Gregersen et al., 1988; Thomsen et al., 1994). As has already been known (Worwood et al., 1997), when assessed by the half-length of LD, the *HLA-B*0801* haplotype had an extra-ordinary degree of LD compared with the *HLA-B*1801* haplotype (3.5 vs 0.4Mbp) (Cullen et al., 1997). As a result of haplotypic variation in LD, the D’ values calculated on different HLA backgrounds show large variations in strength and extent (Blomhoff et al., 2006). As a consequence of haplotype frequency variation, the extent of global LD, the haplotype blocks constructed and the tags selected might be different in different studies of the xMHC region. Thus, the underlying HLA haplotypic architecture is an important parameter to take into account when constructing LD maps of the xMHC (Blomhoff et al., 2006)

Definition of statistical significance

Rather than using a traditional *P* value threshold, both odds ratios and *P* values may be taken into account for selection of SNPs. This hybrid approach has been shown to be superior to the ranking of SNPs by their *P* value in a simulation study (J. Wang & Shete, 2011). It has been shown that many of non-significant but “suggestive” SNPs may be associated with the disease (Lipman et al., 2011), but are missed due to the statistical threshold used. Replication of findings not exceeding the strict threshold in the first study should be considered as an equally valid approach in exploratory studies (Chanock et al., 2007). The associations of susceptibility alleles will rarely reach the required level of significance in GWAS if a Bonferroni correction is used, and the number of false negatives is likely to be large (Rice, Schork, & Rao, 2008). The drawbacks of Bonferroni type manipulations have been recognized, and solutions have been described (Lipman et al., 2011; Shi et al., 2011). The main approach for handling the multiple comparisons issue is becoming the false discovery rate (FDR) procedure, which provides adequate protection against type I error (Benjamini & Hochberg, 1995; Sabatti, Service, & Freimer, 2003). The FDR procedure is easy to apply and not as conservative as the Bonferroni correction. Thus, it does not increase the type II error rate while reducing type I error rate. The biological significance has received much less emphasis than statistical significance in GWAS. There are examples in the literature that statistical significance does not necessarily correlate with biological significance. For example, in GWAS for type 2 diabetes, *PPARG* rs1801282, one of the best replicated genetic effects with known functional correlation for this phenotype, has *P* values of 0.83, 0.019, 0.0013 in the individual studies and a value of 1.7×10^{-6} in the combined analysis of over 32,000 subjects (Williams et al., 2007). It is unlikely that this gene would be highlighted were it not for prior knowledge (Williams et al., 2007).

Model choice and multi-SNP approaches

GWAS data are generally analysed by using the additive genetic model assuming a uniform, linear increase in the odds ratio from wildtype genotype to heterozygous genotype and to variant homozygous genotype. This model is shown to be powerful enough to detect dominant effects, but may be underpowered to detect recessive or overdominant effects (Bush & Moore, 2012; Salanti et al., 2009). The extreme polymorphism of the HLA genes is due to balancing selection that encourages heterozygosity, and heterozygote advantage for HLA polymorphisms has been shown in infectious (Carrington et al., 1999) and autoimmune diseases (Nelson et al., 2004). Thus, currently overdominant model associations, which are common in the MHC region, are undetectable in GWAS data.

It has also been pointed out that exclusively recessive-fit or exclusively dominant-fit associations may be missed as a result of routine use of the additive model (Lettre, Lange, & Hirschhorn, 2007; Salanti et al., 2009; Sellers, 2004; Vukcevic, Hechter, Spencer, & Donnelly, 2011; Zheng et al., 2007). This issue is particularly problematic for the recessive model, especially when the minor allele frequency is not close to 50% (Freidlin, Zheng, Li, & Gastwirth, 2002; Lettre et al., 2007; Zheng et al., 2007). Examples of unravelling genetic associations when the best fitting association model is used have been presented in the literature (Puschmann et al., 2011; Salanti et al., 2009). Specifically, the existence of non-additive effects in the HLA region have been reported at least in autoimmune disorders (Goudey et al., 2017; Lenz et al., 2015). To overcome the potential of missing associations in non-additive models with the exclusive use of the additive model, one can either analyse the data under each model or if the inheritance model is not known, use a robust approach such as maximin efficiency robust test (MERT) or the maximum test (MAX) (Conneely & Boehnke, 2007; Freidlin et al., 2002; Gonzalez et al., 2008; Q. Li, Yu, Li, & Zheng, 2008). The testing of multiple genetic models for genome-wide genotype data can now be achieved online using GWAR even by inexperienced users (Dimou, Tsirigos, Elofsson, & Bagos, 2017).

The least absolute shrinkage and selection operator (LASSO), a shrinkage and variable selection method for linear regression penalizing the absolute size of coefficients, has been used for association analysis with a large number of SNPs simultaneously (Ayers & Cordell, 2010; Hoggart, Whittaker, De Iorio, & Balding, 2008; Shi et al., 2011). MOSGWA is a more recently developed alternative model selection approach which is based on a modification of the Bayesian Information Criterion (Dolejsi, Bodenstorfer, & Frommlet, 2014). MOSGWA detects a number of interesting SNPs for complex diseases, including those in the MHC region, which are not found by other methods. LASSO has been shown to reduce false-positive results while retaining statistical power (Shi et al., 2011) as well as to detect interactions in the MHC region otherwise undetectable (J. Wu, Devlin, Ringquist, Trucco, & Roeder, 2010). Although LASSO can simultaneously analyse all SNPs, it does not perform well to detect associations masked by the phenomenon called "unfaithfulness" in regions like the MHC characterized by correlations among markers (Yang et al., 2011). Correlation cancellation occurs in regions where so many markers are correlated and their individual contribution to the risk is weakened. In a genome-wide survey, associations masked by “unfaithfulness” involving SNPs with at least 1 Mb distance were identified, and all of them were located in the MHC (Yang et al., 2011). Such associations are unlikely to be detected by standard marginal tests or interaction tests, and the marginal effects of correlated SNPs do not express their significant joint effects faithfully due to the correlation cancellation. These hidden associations can be unmasked by the use of the software called "hidden pattern finder" (Yang et al., 2011). The unfaithfulness phenomenon has not been considered in the analysis of MHC region data in any GWAS, and may have resulted in missing existing associations in the MHC region. The recently proposed "multiple enhancer variant" hypothesis for common traits, which suggests that several variants in LD impacting multiple enhancers may collectively affect gene expression (Corradin et al., 2014), may well apply to the MHC region associations. If this is the case, correlation cancellation may result in missing such associations.

Bayesian approaches have many advantages over frequentist methods such as including prior information, easier and more intuitive interpretation of results and being more powerful in certain conditions (Balding, 2006). Bayesian approaches may be particularly attractive to model MHC region associations in GWAS as they are capable of combining different genetic risk models (Stephens & Balding, 2009) and modelling the relationships in an integrated "systems biology" manner, for example with hierarchical modelling to jointly evaluate numerous risk markers and covariates (Heron, O'Dushlaine, Segurado, Gallagher, & Gill, 2011; Stephens & Balding, 2009) as has been done in earlier HLA association studies (Thomas et al., 1992). The Bayesian GWAS framework uses external biological and functional genomics-based information to inform prior probabilities of SNP associations, and using priors based on independent functional knowledge could improve the statistical inference, but would be challenging because of heterogeneity and potential bias (Stranger, Stahl, & Raj, 2011).

Analysis of additional layers of variation in the classical MHC region

Exploration of HLA alleles, haplotypes, supertypes and lineages as susceptibility markers has not been given much importance in GWAS. It is practically impossible to run association studies for all HLA alleles defined by DNA sequencing at the highest resolution (n>16,000 as of March 2017), but algorithms have been developed to predict four-digit HLA alleles from HLA tag SNP data (Karnes et al., 2017), which has been used successfully (Neville et al., 2017), and also work in admixed populations (Nunes et al., 2016). This approach is useful, but there are many other levels of functional MHC specificities as discussed before. These polymorphisms can be incorporated in the analysis of GWAS data either by using proxy SNP constellations (when available), or by manipulating the data after HLA imputation.

The HLA alleles themselves show many important disease associations, but these cannot be unravelled by individual SNP analysis since HLA alleles are defined by multiple nucleotide substitutions. Most GWAS that have found top hits in the MHC have not used either HLA typing or HLA prediction to correlate their findings to known HLA alleles. When this examination is carried out, the MHC SNPs showing associations frequently correlate to an HLA allele or haplotype. We have, however, found examples that certain MHC SNPs that associate with disease risk correlate not with individual alleles, but evolutionary or functional groups of them (Kennedy, Singh, & Dorak, 2012). Most of the specificities shown in **Figure 2** could correspond to yet unknown multi-SNP haplotypes in GWAS data. Given that most are expected to represent HLA types showing evolutionary relationships, searching associations with them would reflect the spirit of the novel approach called evolutionary-based grouping of haplotypes in association analysis (Seltman, Roeder, & Devlin, 2003; Tzeng, 2005). The software package developed for cladistic-based analysis of genetic data (the Evolutionary-Based Haplotype Analysis Package, EHAP) has not been used in GWAS, but would have probably detected associations with HLA functional and/or evolutionarily-related clusters within the MHC. It is well known that such groupings exist, but even HLA association studies often fail to consider them. Given the popularity of GWAS and the number of top hits from the MHC, a more complete analysis may reveal associations stronger than existing ones.

In future GWAS analysis, combining conventional sequence variant analysis with the information on tissue-specific eQTL status, CNV, alternative splicing and epigenetic status is expected to be standard procedure, and should help most with the analysis of xMHC data. Besides, the use of population reference graphs to make use of all of the genomic data from this region (A. Dilthey et al., 2015), typing for all functional specificities not just for SNPs and HLA types, and the development of novel statistical methods taking into account the LD structure and other genomic features of the region should provide a more complete picture of the involvement of the extended MHC region in disease development.

Conclusion

Here, we contrasted the features of the xMHC region with the rest of the genome, and discussed how these differences may have affected the results from this region in existing GWAS, and how they may result in modifications of design, conduct, analysis and interpretation of future GWAS. The analysis of existing data using standard methods does not have the power to unmask all potential associations. The missing heritability concept for GWAS (Manolio et al., 2009) probably applies to the xMHC region more than other regions. This is due to insufficient coverage of the xMHC region in GWAS chips as a natural consequence of enrichment of this region by paralogous genes, extensive CNVs and structural variation. While we cannot thoroughly assess all existing SNPs in the xMHC region, overlooking other layers of functional specificities further contributes to the potential failure of GWAS to detect genetic predisposition conferred by the xMHC region variants. We conclude that despite already showing the highest number of disease associations, the true extent of the involvement of the xMHC region in disease genetics yet to be uncovered.

URLs and Resources for xMHC Region Research:

- HLA Nomenclature (Anthony Nolan Research Institute):
  http://hla.alleles.org/nomenclature/stats.html
  Regular updates on classic and non-classic HLA allele numbers, including pseudogenes.

- GRASP Database: https://grasp.nhlbi.nih.gov/Search.aspx
  The largest catalogue of GWAS results which can be searched by genomic location

- NCBI dbMHC: https://www.ncbi.nlm.nih.gov/projects/gv/mhc
  An NCBI database on MHC-related data

- NCBI MAP Annotation:
  https://www.ncbi.nlm.nih.gov/projects/mapview/map_search.cgi?taxid=9606
  The latest genome map where the up-to-date list of genes and transcripts can be found for any genomic location

- Human Genome Region MHC:
  https://www.ncbi.nlm.nih.gov/grc/human/regions/MHC?asm=GRCh38
  The Genome Reference Consortium site on the MHC region

- Top 100 Expressed Genes in Whole Blood in GTEx Database:
  http://www.gtexportal.org/home/eqtls/tissue?tissueName=Whole_Blood
  List of genes expressed at the highest level in GTEx project

- Database of Genetic Variants: http://dgv.tcag.ca/dgv/app/home
  A searchable catalogue of human genomic structural variation

- SNP2HLA: http://www.broadinstitute.org/mpg/snp2hla
  One of the software packages that impute classical HLA alleles and their amino acid sequences from SNP data

- HLA types of participants of:
  - 1KG: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0078410 (Table S1)
  - HapMap (class I): https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-12-42 (additional file 7)

- Immunogenetic bioinformatics sites:
  - IMGT Immunoinformatics website: http://www.imgt.org/about/immunoinformatics.php
  Links to databases, tools and resources on immunoglobulins, T cell receptors and major histocompatibility loci, including HLA gene sequences, polymorphisms and 3D structures..
  - IPD and IMGT/HLA database: http://www.ebi.ac.uk/ipd
  A centralised system for the study of polymorphism in genes of the immune system, including HLA and KIR genes.
  - ImmunoBase: https://www.immunobase.org

A web based resource focused on the genetics and genomics of immunologically related human diseases, including a genome browser for cumulative results and results from 20 autoimmune disorders.

- ImmPort: https://immport.niaid.nih.gov/home
A data warehouse to promote re-use of immunological data generated by NIH-NIAID funded investigators. Contains datasets of completed research projects, including HLA genetic associations.

References

Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P., & Inoko, H. (2002). Evidence of en bloc duplication in vertebrate genomes. *Nat Genet, 31*(1), 100-105. doi: 10.1038/ng855 ng855 [pii]

Achkar, J. P., Klei, L., de Bakker, P. I., Bellone, G., Rebert, N., Scott, R., . . . Duerr, R. H. (2012). Amino acid position 11 of HLA-DRbeta1 is a major determinant of chromosome 6p association with ulcerative colitis. *Genes Immun, 13*(3), 245-252. doi: gene201179 [pii] 10.1038/gene.2011.79

Ahmad, T., Neville, M., Marshall, S. E., Armuzzi, A., Mulcahy-Hawes, K., Crawshaw, J., . . . Welsh, K. I. (2003). Haplotype-specific linkage disequilibrium patterns define the genetic topography of the human MHC. *Hum Mol Genet, 12*(6), 647-656.

Aird, I., Bentall, H. H., & Roberts, J. A. (1953). A relationship between cancer of stomach and the ABO blood groups. *Br Med J, 1*(4814), 799-801.

Aly, T. A., Eller, E., Ide, A., Gowan, K., Babu, S. R., Erlich, H. A., . . . Fain, P. R. (2006). Multi-SNP analysis of MHC region: remarkable conservation of HLA-A1-B8-DR3 haplotype. *Diabetes, 55*(5), 1265-1269.

Amiel, J.L. (1967). Study of the leukocyte phenotypes in Hodgkin's disease. In E. S. Curtoni, P. L. Mattiuz & R. M. Tosi (Eds.), *Histocompatibility Testing 1967* (pp. 79-81). Copenhagen: Munksgaard.

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

Ayers, K. L., & Cordell, H. J. (2010). SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet Epidemiol, 34*(8), 879-891. doi: 10.1002/gepi.20543

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nat Rev Genet, 7*(10), 781-791. doi: nrg1916 [pii] 10.1038/nrg1916

Bari, R., Leung, M., Turner, V. E., Embrey, C., Rooney, B., Holladay, M., & Leung, W. (2011). Molecular determinant-based typing of KIR alleles and KIR ligands. *Clin Immunol, 138*(3), 274-281. doi: S1521-6616(10)00777-1 [pii] 10.1016/j.clim.2010.12.002

Barnetche, T., Constantin, A., Cantagrel, A., Cambon-Thomsen, A., & Gourraud, P. A. (2008). New classification of HLA-DRB1 alleles in rheumatoid arthritis susceptibility: a combined analysis of worldwide samples. *Arthritis Res Ther, 10*(1), R26. doi: ar2379 [pii] 10.1186/ar2379

Bax, M., van Heemst, J., Huizinga, T. W., & Toes, R. E. (2011). Genetics of rheumatoid arthritis: what have we learned? *Immunogenetics, 63*(8), 459-466. doi: 10.1007/s00251-011-0528-6

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57*(1), 289-300.

Bethke, L., Webb, E., Murray, A., Schoemaker, M., Johansen, C., Christensen, H. C., . . . Houlston, R. (2008). Comprehensive analysis of the role of DNA repair gene polymorphisms on risk of glioma. *Hum Mol Genet, 17*(6), 800-805. doi: ddm351 [pii] 10.1093/hmg/ddm351

Blanchong, C. A., Zhou, B., Rupert, K. L., Chung, E. K., Jones, K. N., Sotos, J. F., . . . Yung Yu, C. (2000). 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, 191*(12), 2183-2196.

Blomhoff, A., Olsson, M., Johansson, S., Akselsen, H. E., Pociot, F., Nerup, J., . . . Lie, B. A. (2006). 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, 7*(2), 130-140. doi: 6364272 [pii] 10.1038/sj.gene.6364272

Brandt, D. Y., Aguiar, V. R., Bitarello, B. D., Nunes, K., Goudet, J., & Meyer, D. (2015). Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 genomes project phase I data. *G3 (Bethesda), 5*(5), 931-941. doi: g3.114.015784 [pii] 10.1534/g3.114.015784

Brewerton, D. A., Hart, F. D., Nicholls, A., Caffrey, M., James, D. C., & Sturrock, R. D. (1973). Ankylosing spondylitis and HL-A 27. *Lancet, 1*(7809), 904-907. doi: S0140-6736(73)92026-6 [pii]

Broderick, P., Wang, Y., Vijayakrishnan, J., Matakidou, A., Spitz, M. R., Eisen, T., . . . Houlston, R. S. (2009). Deciphering the impact of common genetic variation on lung cancer risk: a genome-wide association study. *Cancer Res, 69*(16), 6633-6641. doi: 0008-5472.CAN-09-0680 [pii] 10.1158/0008-5472.CAN-09-0680

Busch, R., De Riva, A., Hadjinicolaou, A. V., Jiang, W., Hou, T., & Mellins, E. D. (2012). On the perils of poor editing: regulation of peptide loading by HLA-DQ and H2-A molecules associated with celiac disease and type 1 diabetes. *Expert Rev Mol Med, 14*, e15. doi: S1462399412000099 [pii] 10.1017/erm.2012.9

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS Comput Biol, 8*(12), e1002822. doi: 10.1371/journal.pcbi.1002822 PCOMPBIOL-D-12-01453 [pii]

Caillat-Zucman, S. (2009). Molecular mechanisms of HLA association with autoimmune diseases. *Tissue Antigens, 73*(1), 1-8. doi: TAN1167 [pii] 10.1111/j.1399-0039.2008.01167.x

Calderwood, S. K., & Ciocca, D. R. (2008). Heat shock proteins: stress proteins with Janus-like properties in cancer. *Int J Hyperthermia, 24*(1), 31-39. doi: 789891263 [pii] 10.1080/02656730701858305

Cano, P., & Fernandez-Vina, M. (2009). Two sequence dimorphisms of DPB1 define the immunodominant serologic epitopes of HLA-DP. *Hum Immunol, 70*(10), 836-843. doi: S0198-8859(09)00176-1 [pii] 10.1016/j.humimm.2009.07.011

Carrington, M., Nelson, G.W., Martin, M.P., Kissner, T., Vlahov, D., Goedert, J.J., . . . O'Brien, S.J. (1999). HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. *Science, 283*(5408), 1748-1752.

Chanock, S. J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D. J., Thomas, G., . . . Collins, F. S. (2007). Replicating genotype-phenotype associations. *Nature, 447*(7145), 655-660. doi: 447655a [pii] 10.1038/447655a

Chapman, S. J., & Hill, A. V. (2012). Human genetic susceptibility to infectious disease. *Nat Rev Genet, 13*(3), 175-188. doi: nrg3114 [pii] 10.1038/nrg3114

Chasman, D. I., Pare, G., Mora, S., Hopewell, J. C., Peloso, G., Clarke, R., . . . Ridker, P. M. (2009). Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis. *PLoS Genet, 5*(11), e1000730. doi: 10.1371/journal.pgen.1000730

Chaudhuri, S., Cariappa, A., Tang, M., Bell, D., Haber, D.A., Isselbacher, K.J., . . . Pillai, S. (2000). Genetic susceptibility to breast cancer: HLA DQB*03032 and HLA DRB1*11 may represent protective alleles. *Proc Natl Acad Sci U S A, 97*(21), 11451-11454.

Chen, D., Juko-Pecirep, I., Hammer, J., Ivansson, E., Enroth, S., Gustavsson, I., . . . Gyllensten, U. (2013). Genome-wide association study of susceptibility loci for cervical cancer. *J Natl Cancer Inst, 105*(9), 624-633. doi: djt051 [pii] 10.1093/jnci/djt051

Chubb, D., Weinhold, N., Broderick, P., Chen, B., Johnson, D. C., Forsti, A., . . . Goldschmidt, H. (2013). Common variation at 3q26.2, 6p21.33, 17p11.2 and 22q13.1 influences multiple myeloma risk. *Nat Genet*. doi: ng.2733 [pii] 10.1038/ng.2733

Cipriani, V., Leung, H. T., Plagnol, V., Bunce, C., Khan, J. C., Shahid, H., . . . Yates, J. R. (2012). Genome-wide association study of age-related macular degeneration identifies associated variants in the TNXB-FKBPL-NOTCH4 region of chromosome 6p21.3. *Hum Mol Genet, 21*(18), 4138-4150. doi: dds225 [pii] 10.1093/hmg/dds225

Collier, S., Sinnott, P. J., Dyer, P. A., Price, D. A., Harris, R., & Strachan, T. (1989). Pulsed field gel electrophoresis identifies a high degree of variability in the number of tandem 21-hydroxylase and complement C4 gene repeats in 21-hydroxylase deficiency haplotypes. *EMBO J, 8*(5), 1393-1402.

Conde, L., Halperin, E., Akers, N. K., Brown, K. M., Smedby, K. E., Rothman, N., . . . Skibola, C. F. (2010). Genome-wide association study of follicular lymphoma identifies a risk locus at 6p21.32. *Nat Genet, 42*(8), 661-664. doi: ng.626 [pii] 10.1038/ng.626

Conneely, K. N., & Boehnke, M. (2007). So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. *Am J Hum Genet, 81*(6), 1158-1168. doi: S0002929707637665 [pii] 10.1086/522036

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., . . . Hurles, M. E. (2010). Origins and functional impact of copy number variation in the human genome. *Nature, 464*(7289), 704-712. doi: nature08516 [pii] 10.1038/nature08516

Consortium, Schizophrenia Working Group of the Psychiatric Genomics. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature, 511*(7510), 421-427. doi: nature13595 [pii] 10.1038/nature13595

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

Contreras-Sesvold, C. L., Sambuughin, N., Blokhin, A., & Deuster, P. A. (2010). A protocol comparison for the analysis of heat shock protein A1B +A1538G SNP. *Cell Stress Chaperones, 15*(2), 205-209. doi: 10.1007/s12192-009-0134-9

Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., Myeroff, L., Willis, J., Cowper-Sal lari, R., . . . Scacheri, P. C. (2014). Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res, 24*(1), 1-13. doi: gr.164079.113 [pii] 10.1101/gr.164079.113

Cullen, M., Noble, J., Erlich, H., Thorpe, K., Beck, S., Klitz, W., . . . Carrington, M. (1997). Characterization of recombination in the HLA class II region. *Am J Hum Genet, 60*(2), 397-407.

Cullen, M., Perfetto, S. P., Klitz, W., Nelson, G., & Carrington, M. (2002). High-resolution patterns of meiotic recombination across the human major histocompatibility complex. *Am J Hum Genet, 71*(4), 759-776. doi: S0002-9297(07)60363-2 [pii] 10.1086/342973

Dall'Olio, G. M., Bertranpetit, J., Wagner, A., & Laayouni, H. (2014). Human genome variation and the concept of genotype networks. *PLoS One, 9*(6), e99424. doi: 10.1371/journal.pone.0099424 PONE-D-14-06484 [pii]

Dausset, J. (1981). The major histocompatibility complex in man. *Science, 213*, 1469-1474.

de Bakker, P. I., McVean, G., Sabeti, P. C., Miretti, M. M., Green, T., Marchini, J., . . . Rioux, J. D. (2006). A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat Genet, 38*(10), 1166-1172.

de Jong, M. M., Nolte, I. M., de Vries, E. G., Schaapveld, M., Kleibeuker, J. H., Oosterom, E., . . . van der Graaf, W. T. (2003). The HLA class III subregion is responsible for an increased breast cancer risk. *Hum Mol Genet, 12*(18), 2311-2319.

de Smith, A. J., Tsalenko, A., Sampas, N., Scheffer, A., Yamada, N. A., Tsang, P., . . . Blakemore, A. I. (2007). Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. *Hum Mol Genet, 16*(23), 2783-2794. doi: ddm208 [pii] 10.1093/hmg/ddm208

Deelen, J., Beekman, M., Uh, H. W., Broer, L., Ayers, K. L., Tan, Q., . . . Slagboom, P. E. (2014). Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. *Hum Mol Genet, 23*(16), 4420-4432. doi: ddu139 [pii] 10.1093/hmg/ddu139

Demant, P., Oomen, L. C., & Oudshoorn-Snoek, M. (1989). Genetics of tumor susceptibility in the mouse: MHC and non-MHC genes. *Adv Cancer Res, 53*, 117-179.

DeWolf, W.C., Lange, P.H., Einarson, M.E., & Yunis, E.J. (1979). HLA and testicular cancer. *Nature, 277*, 216-217.

Diepstra, A., Niens, M., Vellenga, E., van Imhoff, G. W., Nolte, I. M., Schaapveld, M., . . . Poppema, S. (2005). Association with HLA class I in Epstein-Barr-virus-positive and with HLA class III in Epstein-Barr-virus-negative Hodgkin's lymphoma. *Lancet, 365*(9478), 2216-2224.

Dilthey, A., Cox, C., Iqbal, Z., Nelson, M. R., & McVean, G. (2015). Improved genome inference in the MHC using a population reference graph. *Nat Genet, 47*(6), 682-688. doi: ng.3257 [pii] 10.1038/ng.3257

Dilthey, A., Leslie, S., Moutsianas, L., Shen, J., Cox, C., Nelson, M. R., & McVean, G. (2013). Multi-population classical HLA type imputation. *PLoS Comput Biol, 9*(2), e1002877. doi: 10.1371/journal.pcbi.1002877 PCOMPBIOL-D-12-01230 [pii]

Dilthey, A. T., Moutsianas, L., Leslie, S., & McVean, G. (2011). HLA*IMP--an integrated framework for imputing classical HLA alleles from SNP genotypes. *Bioinformatics, 27*(7), 968-972. doi: btr061 [pii] 10.1093/bioinformatics/btr061

Dimou, N. L., Tsirigos, K. D., Elofsson, A., & Bagos, P. G. (2017). GWAR: robust analysis and meta-analysis of genome-wide association studies. *Bioinformatics, 33*(10), 1521-1527. doi: btx008 [pii] 10.1093/bioinformatics/btx008

Dolejsi, E., Bodenstorfer, B., & Frommlet, F. (2014). Analyzing genome-wide association studies with an FDR controlling modification of the Bayesian Information Criterion. *PLoS One, 9*(7), e103322. doi: 10.1371/journal.pone.0103322 PONE-D-14-16059 [pii]

Dorak, M. T., Lawson, T., Machulla, H. K., Darke, C., Mills, K. I., & Burnett, A. K. (1999). Unravelling an HLA-DR association in childhood acute lymphoblastic leukemia. *Blood, 94*(2), 694-700.

Dorak, M.T. (2007). Genotyping with PCR: how to choose the right approach? *The Scientist, 21*(6), 70-72.

Durbin, R. M., Abecasis, G. R., Altshuler, D. L., Auton, A., Brooks, L. D., Gibbs, R. A., . . . McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467*(7319), 1061-1073. doi: nature09534 [pii] 10.1038/nature09534

Dux, A., & Demant, P. (1987). MHC-controlled susceptibility to C3H-MTV-induced mouse mammary tumors is predominantly systemic rather than local. *Int J Cancer, 40*(3), 372-377.

Endo, T., Imanishi, T., Gojobori, T., & Inoko, H. (1997). Evolutionary significance of intra-genome duplications on human chromosomes. *Gene, 205*(1-2), 19-27.

Erlich, R. L., Jia, X., Anderson, S., Banks, E., Gao, X., Carrington, M., . . . de Bakker, P. I. (2011). Next-generation sequencing for HLA typing of class I loci. *BMC Genomics, 12*, 42. doi: 1471-2164-12-42 [pii] 10.1186/1471-2164-12-42

Fairfax, B. P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., . . . Knight, J. C. (2012). Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat Genet, 44*(5), 502-510. doi: ng.2205 [pii] 10.1038/ng.2205

Fehrmann, R. S., Jansen, R. C., Veldink, J. H., Westra, H. J., Arends, D., Bonder, M. J., . . . Franke, L. (2011). Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet, 7*(8), e1002197. doi: 10.1371/journal.pgen.1002197 PGENETICS-D-11-00416 [pii]

Fitzgerald, L. M., Kumar, A., Boyle, E. A., Zhang, Y., McIntosh, L. M., Kolb, S., . . . Stanford, J. L. (2013). Germline missense variants in the BTNL2 gene are associated with prostate cancer susceptibility. *Cancer Epidemiol Biomarkers Prev, 22*(9), 1520-1528. doi: 1055-9965.EPI-13-0345 [pii] 10.1158/1055-9965.EPI-13-0345

Flajnik, M. F., & Kasahara, M. (2010). Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat Rev Genet, 11*(1), 47-59. doi: nrg2703 [pii] 10.1038/nrg2703

Freidlin, B., Zheng, G., Li, Z., & Gastwirth, J. L. (2002). Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum Hered, 53*(3), 146-152. doi: 64976 [pii]

Gaudieri, S., Dawkins, R. L., Habara, K., Kulski, J. K., & Gojobori, T. (2000). SNP profile within the human major histocompatibility complex reveals an extreme and interrupted level of nucleotide diversity. *Genome Res, 10*(10), 1579-1586.

Gonzalez, J. R., Carrasco, J. L., Dudbridge, F., Armengol, L., Estivill, X., & Moreno, V. (2008). Maximizing association statistics over genetic models. *Genet Epidemiol, 32*(3), 246-254. doi: 10.1002/gepi.20299

Gorer, P.A. (1937). The genetic and antigenic basis of tumour transplantation. *J Pathol Bacteriol, 44*(3), 691-697.

Goudey, B., Abraham, G., Kikianty, E., Wang, Q., Rawlinson, D., Shi, F., . . . Inouye, M. (2017). Interactions within the MHC contribute to the genetic architecture of celiac disease. *PLoS One, 12*(3), e0172826. doi: 10.1371/journal.pone.0172826 PONE-D-16-37547 [pii]

Gourraud, P. A., Khankhanian, P., Cereb, N., Yang, S. Y., Feolo, M., Maiers, M., . . . Oksenberg, J. (2014). HLA diversity in the 1000 genomes dataset. *PLoS One, 9*(7), e97282. doi: 10.1371/journal.pone.0097282 PONE-D-13-24434 [pii]

Graffelman, J., Jain, D., & Weir, B. (2017). A genome-wide study of Hardy-Weinberg equilibrium with next generation sequence data. *Hum Genet, 136*(6), 727-741. doi: 10.1007/s00439-017-1786-7 10.1007/s00439-017-1786-7 [pii]

Gregersen, P. K., Kao, H., Nunez-Roldan, A., Hurley, C. K., Karr, R. W., & Silver, J. (1988). Recombination sites in the HLA class II region are haplotype dependent. *J Immunol, 141*(4), 1365-1368.

Grundberg, E., Small, K. S., Hedman, A. K., Nica, A. C., Buil, A., Keildson, S., . . . Spector, T. D. (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet, 44*(10), 1084-1089. doi: ng.2394 [pii] 10.1038/ng.2394

Guo, H., Deng, Q., Wu, C., Hu, L., Wei, S., Xu, P., . . . Wu, T. (2011). Variations in HSPA1B at 6p21.3 are associated with lung cancer risk and prognosis in Chinese populations. *Cancer Res, 71*(24), 7576-7586. doi: 0008-5472.CAN-11-1409 [pii] 10.1158/0008-5472.CAN-11-1409

Haiman, C. A., Han, Y., Feng, Y., Xia, L., Hsu, C., Sheng, X., . . . Stram, D. O. (2013). Genome-wide testing of putative functional exonic variants in relationship with breast and prostate cancer risk in a multiethnic population. *PLoS Genet, 9*(3), e1003419. doi: 10.1371/journal.pgen.1003419 PGENETICS-D-12-02038 [pii]

Handunnetthi, L., Ramagopalan, S. V., Ebers, G. C., & Knight, J. C. (2010). Regulation of major histocompatibility complex class II gene expression, genetic variation and disease. *Genes Immun, 11*(2), 99-112. doi: gene200983 [pii] 10.1038/gene.2009.83

Hassan, G. S., & Mourad, W. (2011). An unexpected role for MHC class II. *Nat Immunol, 12*(5), 375-376. doi: ni.2023 [pii] 10.1038/ni.2023

Hauptmann, G., Tappeiner, G., & Schifferli, J. A. (1988). Inherited deficiency of the fourth component of human complement. *Immunodefic Rev, 1*(1), 3-22.

Heron, E. A., O'Dushlaine, C., Segurado, R., Gallagher, L., & Gill, M. (2011). Exploration of empirical Bayes hierarchical modeling for the analysis of genome-wide association study data. *Biostatistics, 12*(3), 445-461. doi: kxq072 [pii] 10.1093/biostatistics/kxq072

Hoggart, C. J., Whittaker, J. C., De Iorio, M., & Balding, D. J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet, 4*(7), e1000130.

Horton, R., Gibson, R., Coggill, P., Miretti, M., Allcock, R. J., Almeida, J., . . . Beck, S. (2008). Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. *Immunogenetics, 60*(1), 1-18. doi: 10.1007/s00251-007-0262-2

Horton, Roger, Wilming, Laurens, Rand, Vikki, Lovering, Ruth C., Bruford, Elspeth A., Khodiyar, Varsha K., . . . Beck, Stephan. (2004). Gene map of the extended human MHC. *Nat Rev Genet, 5*(12), 889-899.

Howell, W. M. (2014). HLA and disease: guilt by association. *Int J Immunogenet, 41*(1), 1-12. doi: 10.1111/iji.12088

Illing, P. T., Vivian, J. P., Purcell, A. W., Rossjohn, J., & McCluskey, J. (2013). Human leukocyte antigen-associated drug hypersensitivity. *Curr Opin Immunol, 25*(1), 81-89. doi: S0952-7915(12)00157-4 [pii] 10.1016/j.coi.2012.10.002

Ivanova, R., Henon, N., Lepage, V., Charron, D., Vicaut, E., & Schachter, F. (1998). HLA-DR alleles display sex-dependent effects on survival and discriminate between individual and familial longevity. *Hum Mol Genet, 7*(2), 187-194.

Jia, X., Han, B., Onengut-Gumuscu, S., Chen, W. M., Concannon, P. J., Rich, S. S., . . . de Bakker, P. I. (2013). Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One, 8*(6), e64683. doi: 10.1371/journal.pone.0064683 PONE-D-13-06894 [pii]

Karnes, J. H., Shaffer, C. M., Bastarache, L., Gaudieri, S., Glazer, A. M., Steiner, H. E., . . . Roden, D. M. (2017). Comparison of HLA allelic imputation programs. *PLoS One, 12*(2), e0172444. doi: 10.1371/journal.pone.0172444 PONE-D-16-48785 [pii]

Karp, D. R., Marthandan, N., Marsh, S. G., Ahn, C., Arnett, F. C., Deluca, D. S., . . . Scheuermann, R. H. (2010). Novel sequence feature variant type analysis of the HLA genetic association in systemic sclerosis. *Hum Mol Genet, 19*(4), 707-719. doi: ddp521 [pii] 10.1093/hmg/ddp521

Kasahara, M. (1999a). The chromosomal duplication model of the major histocompatibility complex. *Immunol Rev, 167*, 17-32.

Kasahara, M. (1999b). Genome dynamics of the major histocompatibility complex: insights from genome paralogy. *Immunogenetics, 50*(3-4), 134-145. doi: 90500134.251 [pii]

Kasahara, M., Hayashi, M., Tanaka, K., Inoko, H., Sugaya, K., Ikemura, T., & Ishibashi, T. (1996). Chromosomal localization of the proteasome Z subunit gene reveals an ancient chromosomal duplication involving the major histocompatibility complex. *Proc Natl Acad Sci U S A, 93*(17), 9096-9101.

Katsanis, N., Fitzgibbon, J., & Fisher, E. M. (1996). Paralogy mapping: identification of a region in the human MHC triplicated onto human chromosomes 1 and 9 allows the prediction and isolation of novel PBX and NOTCH loci. *Genomics, 35*(1), 101-108. doi: S0888-7543(96)90328-6 [pii] 10.1006/geno.1996.0328

Kennedy, A. E., Singh, S. K., & Dorak, M. T. (2012). Re: Genome-Wide Association Study of Classical Hodgkin Lymphoma and Epstein-Barr Virus Status-Defined Subgroups. *J Natl Cancer Inst, 104*(11), 884-885. doi: djs226 [pii] 10.1093/jnci/djs226

Klareskog, L., Catrina, A. I., & Paget, S. (2009). Rheumatoid arthritis. *Lancet, 373*(9664), 659-672. doi: S0140-6736(09)60008-8 [pii] 10.1016/S0140-6736(09)60008-8

Klitz, W., Aldrich, C. L., Fildes, N., Horning, S. J., & Begovich, A. B. (1994). Localization of predisposition to Hodgkin disease in the HLA class II region. *Am J Hum Genet, 54*(3), 497-505.

Kostyu, D.D. (1994). HLA: fertile territory for developmental genes? *Crit Rev Immunol, 14*(1), 29-59.

Kote-Jarai, Z., Olama, A. A., Giles, G. G., Severi, G., Schleutker, J., Weischer, M., . . . Eeles, R. A. (2011). Seven prostate cancer susceptibility loci identified by a multi-stage genome-wide association study. *Nat Genet, 43*(8), 785-791. doi: ng.882 [pii] 10.1038/ng.882

Kubinak, J. L., Stephens, W. Z., Soto, R., Petersen, C., Chiaro, T., Gogokhia, L., . . . Round, J. L. (2015). MHC variation sculpts individualized microbial communities that control susceptibility to enteric infection. *Nat Commun, 6*, 8642. doi: ncomms9642 [pii] 10.1038/ncomms9642

Kuchenbaecker, K. B., Ramus, S. J., Tyrer, J., Lee, A., Shen, H. C., Beesley, J., . . . Chenevix-Trench, G. (2015). Identification of six new susceptibility loci for invasive epithelial ovarian cancer. *Nat Genet, 47*(2), 164-171. doi: ng.3185 [pii] 10.1038/ng.3185

Kumar, V., Kato, N., Urabe, Y., Takahashi, A., Muroyama, R., Hosono, N., . . . Matsuda, K. (2011). Genome-wide association study identifies a susceptibility locus for HCV-induced hepatocellular carcinoma. *Nat Genet, 43*(5), 455-458. doi: ng.809 [pii] 10.1038/ng.809

Leal, S. M. (2005). Detection of genotyping errors and pseudo-SNPs via deviations from Hardy-Weinberg equilibrium. *Genet Epidemiol, 29*(3), 204-214.

Lenz, T. L., Deutsch, A. J., Han, B., Hu, X., Okada, Y., Eyre, S., . . . Raychaudhuri, S. (2015). Widespread non-additive and interaction effects within HLA loci modulate the risk of autoimmune diseases. *Nat Genet, 47*(9), 1085-1090. doi: ng.3379 [pii] 10.1038/ng.3379

Lenz, T. L., Spirin, V., Jordan, D. M., & Sunyaev, S. R. (2016). Excess of deleterious mutations around HLA genes reveals evolutionary cost of balancing selection. *Mol Biol Evol, 33*(10), 2555-2564. doi: msw127 [pii] 10.1093/molbev/msw127

Lerner, S.P., & Finch, C.E. (1991). The major histocompatibility complex and reproductive functions [Review]. *Endocr Rev, 12*(1), 78-90.

Lettre, G., Lange, C., & Hirschhorn, J. N. (2007). Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet Epidemiol, 31*(4), 358-362. doi: 10.1002/gepi.20217

Li, M., Li, C., & Guan, W. (2008). Evaluation of coverage variation of SNP chips for genome-wide association studies. *Eur J Hum Genet, 16*(5), 635-643. doi: 5202007 [pii] 10.1038/sj.ejhg.5202007

Li, Q., Yu, K., Li, Z., & Zheng, G. (2008). MAX-rank: a simple and robust genome-wide scan for case-control association studies. *Hum Genet, 123*(6), 617-623.

Lilly, F., Boyse, E.A., & Old, L.J. (1964). Genetic basis of susceptibility to viral leukaemogenesis. *Lancet, ii*, 1207-1209.

Lipman, P. J., Cho, M. H., Bakke, P., Gulsvik, A., Kong, X., Lomas, D. A., . . . Lange, C. (2011). On the follow-up of genome-wide association studies: an overall test for the most promising SNPs. *Genet Epidemiol, 35*(5), 303-309. doi: 10.1002/gepi.20578

Lu, S.J., Day, N.E., Degos, L., Lepage, V., Wang, P.C., Chan, S.H., . . . al., et. (1990). Linkage of a nasopharyngeal carcinoma susceptibility locus to the HLA region. *Nature, 346*(6283), 470-471.

Magnusson, P. K. E., Enroth, H., Eriksson, I., Held, M., Nyren, O., Engstrand, L., . . . Gyllensten, U. B. (2001). Gastric cancer and human leukocyte antigen: distinct DQ and DR alleles are associated with development of gastric cancer and infection by Helicobacter pylori. *Cancer Res, 61*(6), 2684-2689.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature, 456*, 18-21.

Major, E., Rigo, K., Hague, T., Berces, A., & Juhos, S. (2013). HLA typing from 1000 genomes whole genome and whole exome illumina data. *PLoS One, 8*(11), e78410. doi: 10.1371/journal.pone.0078410 PONE-D-13-23280 [pii]

Manolio, T. A. (2013). Bringing genome-wide association findings into clinical use. *Nat Rev Genet, 14*(8), 549-558. doi: nrg3523 [pii] 10.1038/nrg3523

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., . . . Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), 747-753. doi: nature08494 [pii] 10.1038/nature08494

Marietta, E., Rishi, A., & Taneja, V. (2015). Immunogenetic control of the intestinal microbiota. *Immunology, 145*(3), 313-322. doi: 10.1111/imm.12474

Martin, M. P., Borecki, I. B., Zhang, Z., Nguyen, L., Ma, D., Gao, X., . . . Rader, J. S. (2010). HLA-Cw group 1 ligands for KIR increase susceptibility to invasive cervical cancer. *Immunogenetics, 62*(11-12), 761-765. doi: 10.1007/s00251-010-0477-5

Martin, M. P., Gao, X., Lee, J. H., Nelson, G. W., Detels, R., Goedert, J. J., . . . Carrington, M. (2002). Epistatic interaction between KIR3DS1 and HLA-B delays the progression to AIDS. *Nat Genet, 31*(4), 429-434. doi: 10.1038/ng934 ng934 [pii]

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., . . . Reich, D. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature, 528*(7583), 499-503. doi: nature16152 [pii] 10.1038/nature16152

Matzaraki, V., Kumar, V., Wijmenga, C., & Zhernakova, A. (2017). The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol, 18*(1), 76. doi: 10.1186/s13059-017-1207-1 10.1186/s13059-017-1207-1 [pii]

McRae, A. F., Powell, J. E., Henders, A. K., Bowdler, L., Hemani, G., Shah, S., . . . Montgomery, G. W. (2014). Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome Biol, 15*(5), R73. doi: gb-2014-15-5-r73 [pii] 10.1186/gb-2014-15-5-r73

Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M. J., . . . Easton, D. F. (2015). Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat Genet, 47*(4), 373-380. doi: ng.3242 [pii] 10.1038/ng.3242

Miyadera, H., Ohashi, J., Lernmark, A., Kitamura, T., & Tokunaga, K. (2015). Cell-surface MHC density profiling reveals instability of autoimmunity-associated HLA. *J Clin Invest, 125*(1), 275-291. doi: 74961 [pii] 10.1172/JCI74961

Moutsianas, L., Enciso-Mora, V., Ma, Y. P., Leslie, S., Dilthey, A., Broderick, P., . . . Houlston, R. S. (2011). Multiple Hodgkin lymphoma-associated loci within the HLA region at chromosome 6p21.3. *Blood, 118*(3), 670-674. doi: blood-2011-03-339630 [pii] 10.1182/blood-2011-03-339630

Muhlbock, O., & Dux, A. (1974). Histocompatibility genes (the H-2 complex) and susceptibility to mammary tumor virus in mice. *J Natl Cancer Inst, 53*, 993-996.

Nalls, M. A., Plagnol, V., Hernandez, D. G., Sharma, M., Sheerin, U. M., Saad, M., . . . Wood, N. W. (2011). Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet, 377*(9766), 641-649. doi: S0140-6736(10)62345-8 [pii] 10.1016/S0140-6736(10)62345-8

Nelson, G. W., Martin, M. P., Gladman, D., Wade, J., Trowsdale, J., & Carrington, M. (2004). Cutting edge: heterozygote advantage in autoimmune disease: hierarchy of protection/susceptibility conferred by HLA and killer Ig-like receptor combinations in psoriatic arthritis. *J Immunol, 173*(7), 4273-4276. doi: 173/7/4273 [pii]

Neville, M. J., Lee, W., Humburg, P., Wong, D., Barnardo, M., Karpe, F., & Knight, J. C. (2017). High resolution HLA haplotyping by imputation for a British population bioresource. *Hum Immunol, 78*(3), 242-251. doi: S0198-8859(17)30015-0 [pii] 10.1016/j.humimm.2017.01.006

Nunes, K., Zheng, X., Torres, M., Moraes, M. E., Piovezan, B. Z., Pontes, G. N., . . . Meyer, D. (2016). HLA imputation in an admixed population: An assessment of the 1000 Genomes data as a training set. *Hum Immunol, 77*(3), 307-312. doi: S0198-8859(15)00557-1 [pii] 10.1016/j.humimm.2015.11.004

Oomen, L.C., Van der Valk, M.A., & Den Engelse, L. (1983). Tumour susceptibility in mice in relation to H-2 haplotype. *IARC Scientific Publications, 51*, 205-221.

Pirastu, N., Kooyman, M., Traglia, M., Robino, A., Willems, S. M., Pistis, G., . . . Gasparini, P. (2015). Genome-wide association analysis on five isolated populations identifies variants of the HLA-DOA gene associated with white wine liking. *Eur J Hum Genet, 23*(12), 1717-1722. doi: ejhg201534 [pii] 10.1038/ejhg.2015.34

Profaizer, T., & Eckels, D. (2012). HLA alleles and drug hypersensitivity reactions. *Int J Immunogenet, 39*(2), 99-105. doi: 10.1111/j.1744-313X.2011.01061.x

Puschmann, A., Verbeeck, C., Heckman, M. G., Soto-Ortolaza, A. I., Lynch, T., Jasinska-Myga, B., . . . Ross, O. A. (2011). Human leukocyte antigen variation and Parkinson's disease. *Parkinsonism Relat Disord, 17*(5), 376-378. doi: S1353-8020(11)00067-8 [pii] 10.1016/j.parkreldis.2011.03.008

Rapley, E. A., Turnbull, C., Al Olama, A. A., Dermitzakis, E. T., Linger, R., Huddart, R. A., . . . Stratton, M. R. (2009). A genome-wide association study of testicular germ cell tumor. *Nat Genet, 41*(7), 807-810. doi: ng.394 [pii] 10.1038/ng.394

Raychaudhuri, S., Sandor, C., Stahl, E. A., Freudenberg, J., Lee, H. S., Jia, X., . . . de Bakker, P. I. (2012). Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet, 44*(3), 291-296. doi: ng.1076 [pii] 10.1038/ng.1076

Rice, T. K., Schork, N. J., & Rao, D. C. (2008). Methods for handling multiple testing. *Adv Genet, 60*, 293-308. doi: S0065-2660(07)00412-9 [pii] 10.1016/S0065-2660(07)00412-9

Rietveld, C. A., Medland, S. E., Derringer, J., Yang, J., Esko, T., Martin, N. W., . . . Koellinger, P. D. (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science, 340*(6139), 1467-1471. doi: science.1235488 [pii] 10.1126/science.1235488

Rioux, J. D., Goyette, P., Vyse, T. J., Hammarstrom, L., Fernando, M. M., Green, T., . . . Hauser, S. L. (2009). Mapping of multiple susceptibility variants within the MHC region for 7 immune-mediated diseases. *Proc Natl Acad Sci U S A, 106*(44), 18680-18685. doi: 0909307106 [pii] 10.1073/pnas.0909307106

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kahler, A. K., Akterin, S., . . . Sullivan, P. F. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet, 45*(10), 1150-1159. doi: ng.2742 [pii] 10.1038/ng.2742

Robbins, F., Hurley, C. K., Tang, T., Yao, H., Lin, Y. S., Wade, J., . . . Hartzman, R. J. (1997). Diversity associated with the second expressed HLA-DRB locus in the human population. *Immunogenetics, 46*(2), 104-110.

Ropcke, G., Moen, C. J., Hart, A. A., & Demant, P. (1990). Effects of the MHC on hormonal induction of mammary tumors and function of hypophyseal isografts in the mouse. *Immunogenetics, 31*(5-6), 347-355.

Sabatti, C., Service, S., & Freimer, N. (2003). False discovery rate in linkage and association genome screens for complex disorders. *Genetics, 164*(2), 829-833.

Salanti, G., Southam, L., Altshuler, D., Ardlie, K., Barroso, I., Boehnke, M., . . . Ioannidis, J. P. (2009). Underlying genetic models of inheritance in established type 2 diabetes associations. *Am J Epidemiol, 170*(5), 537-545. doi: kwp145 [pii] 10.1093/aje/kwp145

Schellekens, J., Rozemuller, E. H., Borst, H. P., Otten, H. G., van den Tweel, J. G., & Tilanus, M. G. (2007). NK-KIR ligand identification: a quick Q-PCR approach for HLA-C epitope typing. *Tissue Antigens, 69*(4), 334-337. doi: TAN809 [pii] 10.1111/j.1399-0039.2007.00809.x

Schlattl, A., Anders, S., Waszak, S. M., Huber, W., & Korbel, J. O. (2011). Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions. *Genome Res, 21*(12), 2004-2013. doi: gr.122614.111 [pii] 10.1101/gr.122614.111

Sekar, A., Bialas, A. R., de Rivera, H., Davis, A., Hammond, T. R., Kamitaki, N., . . . McCarroll, S. A. (2016). Schizophrenia risk from complex variation of complement component 4. *Nature, 530*(7589), 177-183. doi: nature16549 [pii] 10.1038/nature16549

Sellers, T. A. (2004). Genetic ancestry and molecular epidemiology. *Cancer Epidemiol Biomarkers Prev, 13*(4), 499-500.

Seltman, H., Roeder, K., & Devlin, B. (2003). Evolutionary-based association analysis using haplotype data. *Genet Epidemiol, 25*(1), 48-58.

Shi, G., Boerwinkle, E., Morrison, A. C., Gu, C. C., Chakravarti, A., & Rao, D. C. (2011). Mining gold dust under the genome wide significance level: a two-stage approach to analysis of GWAS. *Genet Epidemiol, 35*(2), 111-118. doi: 10.1002/gepi.20556

Shiina, T., Ando, A., Suto, Y., Kasai, F., Shigenari, A., Takishima, N., . . . Inoko, H. (2001). Genomic anatomy of a premier major histocompatibility complex paralogous region on chromosome 1q21-q22. *Genome Res, 11*(5), 789-802. doi: 10.1101/gr.175801

Shiina, T., Hosomichi, K., Inoko, H., & Kulski, J. K. (2009). The HLA genomic loci map: expression, interaction, diversity and disease. *J Hum Genet, 54*(1), 15-39. doi: jhg20085 [pii] 10.1038/jhg.2008.5

Simon, S., Truedsson, L., Marcus-Bagley, D., Awdeh, Z., Eisenbarth, G. S., Brink, S. J., . . . Alper, C. A. (1997). Relationship between protein complotypes and DNA variant haplotypes: complotype-RFLP constellations (CRC). *Hum Immunol, 57*(1), 27-36.

Snoek, M., Albertella, M. R., van Kooij, M., Wixon, J., van Vugt, H., de Groot, K., & Campbell, R. D. (2000). G7c, a novel gene in the mouse and human major histocompatibility complex class III region, possibly controlling lung tumor susceptibility. *Immunogenetics, 51*(4-5), 383-386.

Sollid, L. M., Pos, W., & Wucherpfennig, K. W. (2014). Molecular mechanisms for contribution of MHC molecules to autoimmune diseases. *Curr Opin Immunol, 31*, 24-30. doi: S0952-7915(14)00106-X [pii] 10.1016/j.coi.2014.08.005

Stastny, P. (1978). Association of the B-cell alloantigen DRw4 with rheumatoid arthritis. *N Engl J Med, 298*(16), 869-871. doi: 10.1056/NEJM197804202981602

Stephens, M., & Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nat Rev Genet, 10*(10), 681-690. doi: nrg2615 [pii] 10.1038/nrg2615

Stranger, B. E., Stahl, E. A., & Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics, 187*(2), 367-383. doi: genetics.110.120907 [pii] 10.1534/genetics.110.120907

Su, Z., Gay, L. J., Strange, A., Palles, C., Band, G., Whiteman, D. C., . . . Jankowski, J. A. (2012). Common variants at the MHC locus and at chromosome 16q24.1 predispose to Barrett's esophagus. *Nat Genet, 44*(10), 1131-1136. doi: ng.2408 [pii] 10.1038/ng.2408

Thomas, D., Langholz, B., Clayton, D., Pitkaniemi, J., Tuomilehto-Wolf, E., & Tuomilehto, J. (1992). Empirical Bayes methods for testing associations with large numbers of candidate genes in the presence of environmental risk factors, with applications to HLA associations in IDDM. *Ann Med, 24*(5), 387-392.

Thomsen, M., Neugebauer, M., Arnaud, J., Borot, N., Sevin, A., Baur, M., & Cambon-Thomsen, A. (1994). Recombination fractions in the HLA system based on the data set 'provinces Francaises': indications of haplotype-specific recombination rates. *Eur J Immunogenet, 21*(1), 33-43.

Thomson, G., Marthandan, N., Hollenbach, J. A., Mack, S. J., Erlich, H. A., Single, R. M., . . . Helmberg, W. (2010). Sequence feature variant type (SFVT) analysis of the HLA genetic association in juvenile idiopathic arthritis. *Pac Symp Biocomput, 15*, 359-370. doi: 9789814295291_0038 [pii]

Thorleifsson, G., Walters, G. B., Gudbjartsson, D. F., Steinthorsdottir, V., Sulem, P., Helgadottir, A., . . . Stefansson, K. (2009). Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nat Genet, 41*(1), 18-24. doi: ng.274 [pii] 10.1038/ng.274

Timofeeva, M. N., Hung, R. J., Rafnar, T., Christiani, D. C., Field, J. K., Bickeboller, H., . . . Landi, M. T. (2012). Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. *Hum Mol Genet, 21*(22), 4980-4995. doi: dds334 [pii] 10.1093/hmg/dds334

Traherne, J. A. (2008). Human MHC architecture and evolution: implications for disease association studies. *Int J Immunogenet, 35*(3), 179-192. doi: EJI765 [pii] 10.1111/j.1744-313X.2008.00765.x

Trowsdale, J., & Knight, J. C. (2013). Major histocompatibility complex genomics and human disease. *Annu Rev Genomics Hum Genet, 14*, 301-323. doi: 10.1146/annurev-genom-091212-153455

Truman, J. P., Garban, F., Choqueux, C., Charron, D., & Mooney, N. (1996). HLA class II signaling mediates cellular activation and programmed cell death. *Exp Hematol, 24*(12), 1409-1415.

Tse, K. P., Su, W. H., Chang, K. P., Tsang, N. M., Yu, C. J., Tang, P., . . . Shugart, Y. Y. (2009). Genome-wide association study reveals multiple nasopharyngeal carcinoma-associated loci within the HLA region at chromosome 6p21.3. *Am J Hum Genet, 85*(2), 194-203. doi: S0002-9297(09)00298-5 [pii] 10.1016/j.ajhg.2009.07.007

Tzeng, J. Y. (2005). Evolutionary-based grouping of haplotypes in association analysis. *Genet Epidemiol, 28*(3), 220-231.

Ucisik-Akkaya, E., Davis, C. F., Gorodezky, C., Alaez, C., & Dorak, M. T. (2010). HLA complex-linked heat shock protein genes and childhood acute lymphoblastic leukemia. *Cell Stress Chaperones, 15*(5), 475-485.

Ugolotti, E., Vanni, I., Raso, A., Benzi, F., Malnati, M., & Biassoni, R. (2011). Human leukocyte antigen-B (-Bw6/-Bw4 I(80), T(80)) and human leukocyte antigen-C (-C1/-C2) subgrouping using pyrosequence analysis. *Hum Immunol, 72*(10), 859-868. doi: S0198-8859(11)00114-5 [pii] 10.1016/j.humimm.2011.05.007

Urayama, K. Y., Jarrett, R. F., Hjalgrim, H., Diepstra, A., Kamatani, Y., Chabrier, A., . . . McKay, J. D. (2012). Genome-wide association study of classical Hodgkin lymphoma and Epstein-Barr Virus status-defined subgroups. *J Natl Cancer Inst, 104*(3), 240-153. doi: djr516 [pii] 10.1093/jnci/djr516 van Dongen, J., Nivard, M. G., Willemsen, G., Hottenga, J. J., Helmer, Q., Dolan, C. V., . . . Boomsma, D. I. (2016). Genetic and environmental influences interact with age and sex in shapng the human methylome. *Nat Commun, 7*, 11115. doi: ncomms11115 [pii] 10.1038/ncomms11115

Vandiedonck, C., & Knight, J. C. (2009). The human major histocompatibility complex as a paradigm in genomics research. *Brief Funct Genomic Proteomic, 8*(5), 379-394. doi: elp010 [pii] 10.1093/bfgp/elp010

Vandiedonck, C., Taylor, M. S., Lockstone, H. E., Plant, K., Taylor, J. M., Durrant, C., . . . Knight, J. C. (2011). Pervasive haplotypic variation in the spliceo-transcriptome of the human major histocompatibility complex. *Genome Res, 21*(7), 1042-1054. doi: gr.116681.110 [pii] 10.1101/gr.116681.110

Vukcevic, D., Hechter, E., Spencer, C., & Donnelly, P. (2011). Disease model distortion in association studies. *Genet Epidemiol, 35*(4), 278-290. doi: 10.1002/gepi.20576

Wang, J., & Shete, S. (2011). A powerful hybrid approach to select top single-nucleotide polymorphisms for genome-wide association study. *BMC Genet, 12*, 3. doi: 1471-2156-12-3 [pii] 10.1186/1471-2156-12-3

Wang, Y., Broderick, P., Webb, E., Wu, X., Vijayakrishnan, J., Matakidou, A., . . . Houlston, R. S. (2008). Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet, 40*(12), 1407-1409. doi: ng.273 [pii] 10.1038/ng.273

Westra, H. J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., . . . Franke, L. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet, 45*(10), 1238-1243. doi: ng.2756 [pii] 10.1038/ng.2756

Whitehead, A. S., Woods, D. E., Fleischnick, E., Chin, J. E., Yunis, E. J., Katz, A. J., . . . Colten, H. R. (1984). DNA polymorphism of the C4 genes. A new marker for analysis of the major histocompatibility complex. *N Engl J Med, 310*(2), 88-91.

Williams, S. M., Canter, J. A., Crawford, D. C., Moore, J. H., Ritchie, M. D., & Haines, J. L. (2007). Problems with genome-wide association studies. *Science, 316*(5833), 1840-1842.

Woolcott, C. G., Shvetsov, Y. B., Stanczyk, F. Z., Wilkens, L. R., White, K. K., Caberto, C., . . . Goodman, M. T. (2010). Plasma sex hormone concentrations and breast cancer risk in an ethnically diverse population of postmenopausal women: the Multiethnic Cohort Study. *Endocr Relat Cancer, 17*(1), 125-134. doi: ERC-09-0211 [pii] 10.1677/ERC-09-0211

Worwood, M., Raha Chowdhury, R., Robson, K. J., Pointon, J., Shearman, J. D., & Darke, C. (1997). The HLA A1-B8 haplotype extends 6 Mb beyond HLA-A: associations between HLA-A, B, F and 15 microsatellite markers. *Tissue Antigens, 50*(5), 521-526.

Wright, F. A., Sullivan, P. F., Brooks, A. I., Zou, F., Sun, W., Xia, K., . . . Boomsma, D. I. (2014). Heritability and genomics of gene expression in peripheral blood. *Nat Genet, 46*(5), 430-437. doi: ng.2951 [pii] 10.1038/ng.2951

Wu, C., Wang, Z., Song, X., Feng, X. S., Abnet, C. C., He, J., . . . Chanock, S. J. (2014). Joint analysis of three genome-wide association studies of esophageal squamous cell carcinoma in Chinese populations. *Nat Genet, 46*(9), 1001-1006. doi: ng.3064 [pii] 10.1038/ng.3064

Wu, J., Devlin, B., Ringquist, S., Trucco, M., & Roeder, K. (2010). Screen and clean: a tool for identifying interactions in genome-wide association studies. *Genet Epidemiol, 34*(3), 275-285. doi: 10.1002/gepi.20459

Wu, Y. L., Savelli, S. L., Yang, Y., Zhou, B., Rovin, B. H., Birmingham, D. J., . . . Yu, C. Y. (2007). 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, 179*(5), 3012-3025. doi: 179/5/3012 [pii]

Wu, Y. L., Yang, Y., Chung, E. K., Zhou, B., Kitzmiller, K. J., Savelli, S. L., . . . Yu, C. Y. (2008). Phenotypes, genotypes and disease susceptibility associated with gene copy number variations: complement C4 CNVs in European American healthy subjects and those with systemic lupus erythematosus. *Cytogenet Genome Res, 123*(1-4), 131-141. doi: 000184700 [pii] 10.1159/000184700

Xie, M., Li, J., & Jiang, T. (2010). Accurate HLA type inference using a weighted similarity graph. *BMC Bioinformatics, 11 Suppl 11*, S10. doi: 1471-2105-11-S11-S10 [pii] 10.1186/1471-2105-11-S11-S10

Xie, T., Rowen, L., Aguado, B., Ahearn, M. E., Madan, A., Qin, S., . . . Hood, L. (2003). Analysis of the gene-dense major histocompatibility complex class III region and its comparison to mouse. *Genome Res, 13*(12), 2621-2636.

Yang, C., Wan, X., Yang, Q., Xue, H., Tang, N. L., & Yu, W. (2011). A hidden two-locus disease association pattern in genome-wide association studies. *BMC Bioinformatics, 12*, 156. doi: 1471-2105-12-156 [pii] 10.1186/1471-2105-12-156

Yun, G., Tolar, J., Yerich, A. K., Marsh, S. G., Robinson, J., Noreen, H., . . . Miller, J. S. (2007). A novel method for KIR-ligand typing by pyrosequencing to predict NK cell alloreactivity. *Clin Immunol, 123*(3), 272-280. doi: S1521-6616(07)00031-9 [pii] 10.1016/j.clim.2007.01.011

Zhang, X. C., Li, S. S., Wang, H., Hansen, J. A., & Zhao, L. P. (2011). Empirical evaluations of analytical issues arising from predicting HLA alleles using multiple SNPs. *BMC Genet, 12*, 39. doi: 1471-2156-12-39 [pii] 10.1186/1471-2156-12-39

Zhang, X., Tworoger, S. S., Eliassen, A. H., & Hankinson, S. E. (2013). Postmenopausal plasma sex hormone levels and breast cancer risk over 20 years of follow-up. *Breast Cancer Res Treat, 137*(3), 883-892. doi: 10.1007/s10549-012-2391-z

Zhang, Y. B., Li, X., Zhang, F., Wang, D. M., & Yu, J. (2012). A preliminary study of copy number variation in Tibetans. *PLoS One, 7*(7), e41768. doi: 10.1371/journal.pone.0041768 PONE-D-12-02453 [pii]

Zheng, G., Joo, J., Lin, J. P., Stylianou, M., Waclawiw, M. A., & Geller, N. L. (2007). Robust ranks of true associations in genome-wide case-control association studies. *BMC Proc, 1 Suppl 1*, S165.

Table 1. Descriptive information on genome coordinates, gene content and polymorphisms of the classical MHC region

a. Genome coordinates on chromosome 6 [a]

	Telomeric end	Centromeric end
Classical MHC region	29672373 (*ZFP57*)	33148800 (*HCG24*)
Classical class I region	29672373 (*ZFP57*)	31511124 (*MICB*)
Class III region	31511125 (*PPIAP9*)	32224067 (*NOTCH4*)
Classical class II region	32224068 (*C6orf10*)	33148800 (*HCG24*)

b. Gene content

Total number of genes (all categories)	271
Protein-coding genes	151 [b]
Non-coding RNA	39 [c]
Pseudogene	81

c. Polymorphism

i. Classical HLA gene polymorphisms [d]

Total number of HLA alleles	16,755
Total number of HLA class I alleles	12,351
HLA-A	3,913
HLA-B	4,765
HLA-C	3,510
Total number of HLA class II alleles	4,404
HLA-DRA	7
HLA-DRB1	2,311
HLA-DQA1	78
HLA-DQB1	1,079
HLA-DPA1	45
HLA-DPB1	828

ii. Sequence polymorphisms [e]

Total number of SNPs classical MHC region	253,309
Class I region	125,747
Class III region	51,221
Class II region	76,341

[a]: From NCBI Map Annotation Release 108.6 in March 2017. Genes in brackets are the most centromeric and most telomeric ones in each region.

[b]: including 5 open reading frame and 20 yet uncharacterised genes.

[c]: Including 8 antisense-RNA, 9 microRNA, 1 long non-coding RNA, 8 antisense and 6 small nuclear/nucleolar RNA genes.

[d]: From HLA Nomenclature website (Anthony Nolan Research Institute), March 2017 update.

[e]: From Ensembl (GRCh38.p7; March 2017) using the coordinates given above (SNPs and indels excluding flagged variants).

Table 2. Gene set enrichment analysis results of the complete xMHC gene list on PANTHER tool for gene list analysis [a,b]

Gene ontology biological process	Fold enrichment	*P* value
Antigen processing and presentation	> 5	1.18E-16
Antigen processing and presentation of peptide antigen	> 5	2.33E-16
Nucleosome assembly	> 5	1.35E-15
Antigen processing and presentation of exogenous peptide antigen	> 5	5.02E-15
Antigen processing and presentation of exogenous antigen	> 5	1.23E-14
Chromatin assembly	> 5	1.52E-14
Interferon-gamma-mediated signaling pathway	> 5	2.33E-14
Protein-DNA complex assembly	> 5	9.66E-14
Nucleosome organization	> 5	9.66E-14
Chromatin assembly or disassembly	> 5	3.02E-13
DNA packaging	> 5	2.14E-12
Protein-DNA complex subunit organization	> 5	3.10E-12
Response to interferon-gamma	> 5	7.22E-12
Cellular response to interferon-gamma	> 5	8.11E-12
Immune response	3.49	1.83E-11
DNA conformation change	> 5	1.30E-10
Cellular macromolecular complex assembly	4.8	2.59E-09
Antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	> 5	7.47E-09
Antigen processing and presentation of peptide antigen via MHC class I	> 5	1.49E-08
Defense response	2.95	1.01E-07
Regulation of immune system process	2.99	1.07E-07
Innate immune response	3.48	3.92E-07
Antigen processing and presentation of exogenous peptide antigen via MHC class II	> 5	6.67E-07
Antigen processing and presentation of peptide antigen via MHC class II	> 5	8.62E-07
immune system process	2.44	9.56E-07
Antigen processing and presentation of endogenous peptide	> 5	1.05E-

antigen		06
Positive regulation of cell-cell adhesion	> 5	1.97E-06
Regulation of cell-cell adhesion	> 5	2.86E-06
Antigen processing and presentation of endogenous antigen	> 5	3.03E-06
Protein complex assembly	3.3	5.47E-06
Protein complex biogenesis	3.3	5.47E-06
Regulation of T cell activation	> 5	5.96E-06
Positive regulation of T cell activation	> 5	8.38E-06
Positive regulation of immune system process	3.39	9.26E-06
Regulation of leukocyte cell-cell adhesion	> 5	9.40E-06
Antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-independent	> 5	1.02E-05
Positive regulation of homotypic cell-cell adhesion	> 5	1.12E-05
Antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent	> 5	1.17E-05
Positive regulation of leukocyte cell-cell adhesion	> 5	1.20E-05
Regulation of homotypic cell-cell adhesion	> 5	1.46E-05
Regulation of lymphocyte activation	> 5	1.72E-05
Antigen processing and presentation of exogenous peptide antigen via MHC class I	> 5	2.00E-05
Regulation of immune response	3.36	2.08E-05
Macromolecular complex assembly	2.91	2.84E-05
Antigen processing and presentation of endogenous peptide antigen via MHC class I	> 5	3.34E-05
MHC protein complex assembly	> 5	8.35E-05
Cytokine-mediated signaling pathway	4.55	9.61E-05

[a] PANTHER tool is accessible at http://www.pantherdb.org

[b] The list is truncated at the arbitrary statistical threshold of $P < 1 \times 10^{-4}$

Table 3. Representative HLA and disease associations, and their corresponding GWAS associations [a]

Disease	**HLA association**	**GWAS association** (SNP ID; chromosome 6 position [b])	**GWAS *P* value**	**GWAS reference** (Pubmed ID)
Psoriasis	*HLA-C*06:02* (*PSORS1*)	rs4406273; 31298313 ([c])	4.5E-723 [d]	23143594
Myasthenia gravis	*HLA-C*07:01*	rs7750641; 31161533 ([c])	1.7E-114	23055271
HIV-1 control	*HLA-B, HLA-C*	rs9264942; 31306603	2.8E-35	21051598
Ankylosing spondylitis	*HLA-B*27*	rs7743761; 31368323	5.0E-304	20062062
Malaria	*HLA-B*53*	No association in xMHC	-	-
Abacavir drug hypersensitivity	*HLA-B*57:01*	No GWAS	-	-
Dengue shock syndrome	-	rs3132468; 31507709	4.4E-11	22001756
Sarcoidosis	-	rs2076530; 32396039	3.0E-11	22936702
Idiopathic membranous nephropathy	*HLA-DRB1*03*	rs2187668; 32638107 ([c])	8.0E-93	21323541
Type 1 diabetes	*DRB1*04-DQA1*03:01-DQB1*03:02; DRB1*03-DQA1*05:01-DQB1*02:01*	rs9273363; 32658495	1.0E-307	17554300
Rheumatoid arthritis (cyclic citrullinated peptide positive)	*HLA-DRB1*04:01, HLA-DQA1*03:01*	rs660895; 32609603 ([c])	1.0E-300	23143596
Systemic lupus erythematosus	*HLA-DRB1*03:01*	rs1270942; 31951083 ([c])	2.0E-165	26502338
Multiple sclerosis	*HLA-DRB1*05:01*	rs3135388; 32445274 ([c])	3.8E-225	19525953
Systemic sclerosis (Anti-topoisomerase-I antibody positive)	*DRB1*11:04-DQA1*05:01-DQB1*03:01*	rs3129763; 32623148	9.2E-187	21779181
Systemic sclerosis (Anti-centromere antibody positive)	*DRB1*11:04*	rs9275390; 32701379	1.1E-130	21779181
Pemphigus vulgaris	*HLA-DQB1*03:01*	rs9275184; 32686937	7.7E-21	22437316
Leprosy	*(HLA-DRB1, DQA1)*	rs9271100; 32608701	8.0E-95	25642632
Narcolepsy	*HLA-DQB1*06:02*	rs9271117; 32609018	6.0E-14	24204295
Ulcerative colitis	*HLA-DRB1*11:01*	rs6927022; 32644620	4.7E-133	23128233
Graves' disease	*HLA-DRB1*03:01, HLA-DQA1*05:01*	rs1521; 31382927	2.0E-65	21841780
Celiac disease	*HLA-DQA1*05:01, HLA-DQB1*02:01*	rs2187668; 32638107 ([c])	5.8E-209	20190752
Selective IgA deficiency	*HLA-DQB1*02:01*	rs116041786; 32634619	3.0E-92	27723758

[a] The HLA and disease associations are based on (Trowsdale & Knight, 2013) with some additions. GWAS data was extracted from GRASP v2.0.0.0 (https://grasp.nhlbi.nih.gov) and EBI GWAS Catalog (http://www.ebi.ac.uk/gwas).

[b] Chromosome 6 positions are hg19 coordinates.

[c] These GWAS associations correspond to the known HLA allelic associations.

[d] The psoriasis association is the statistically most significant association in any GWAS.

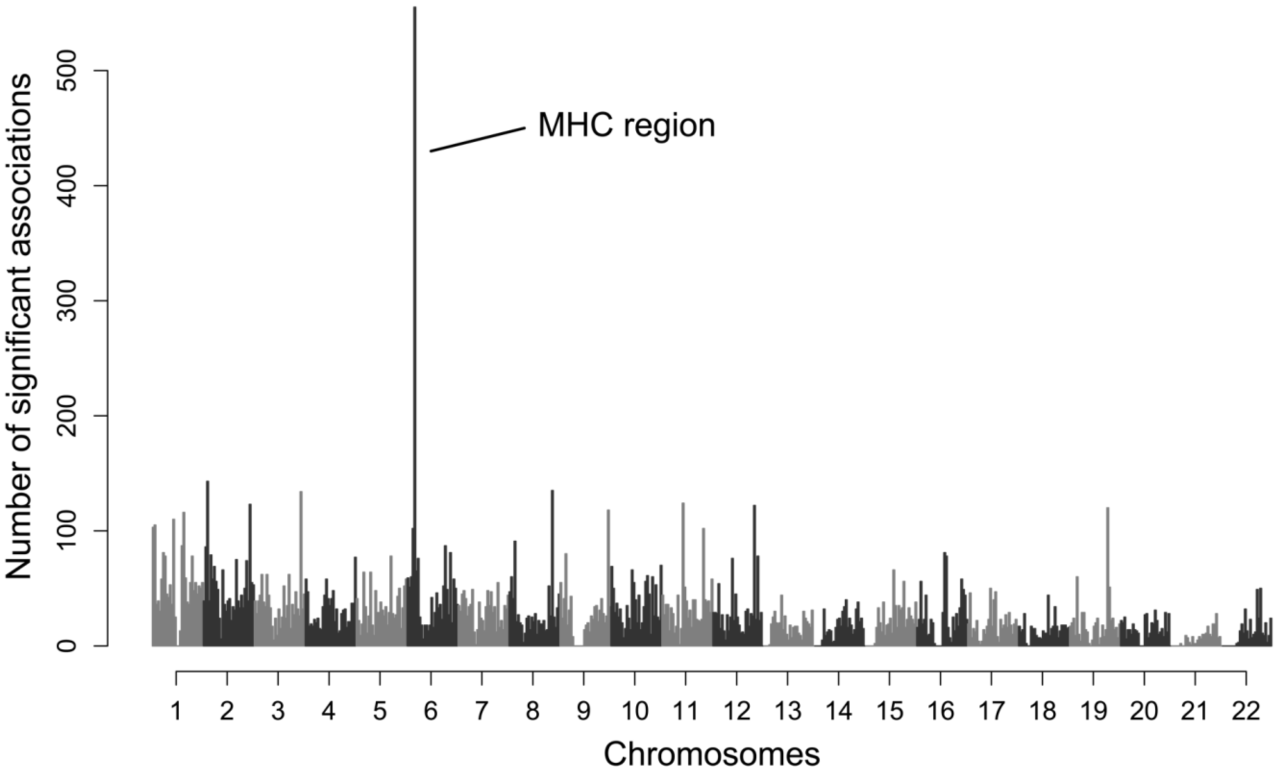

Figure 1. **Number of significant GWAS associations along the genome**. The chromosomal location of significant trait associations from GWAS (*N* = 18,682) are shown for all autosomes. Data from NHGRI GWAS catalog. Reproduced from “Lenz TL, Spirin V, Jordan DM, Sunyaev SR. Excess of Deleterious Mutations around HLA Genes Reveals Evolutionary Cost of Balancing Selection. Mol Biol Evol 2016;33(10):2555-64. doi: 10.1093/molbev/msw127” by permission of Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

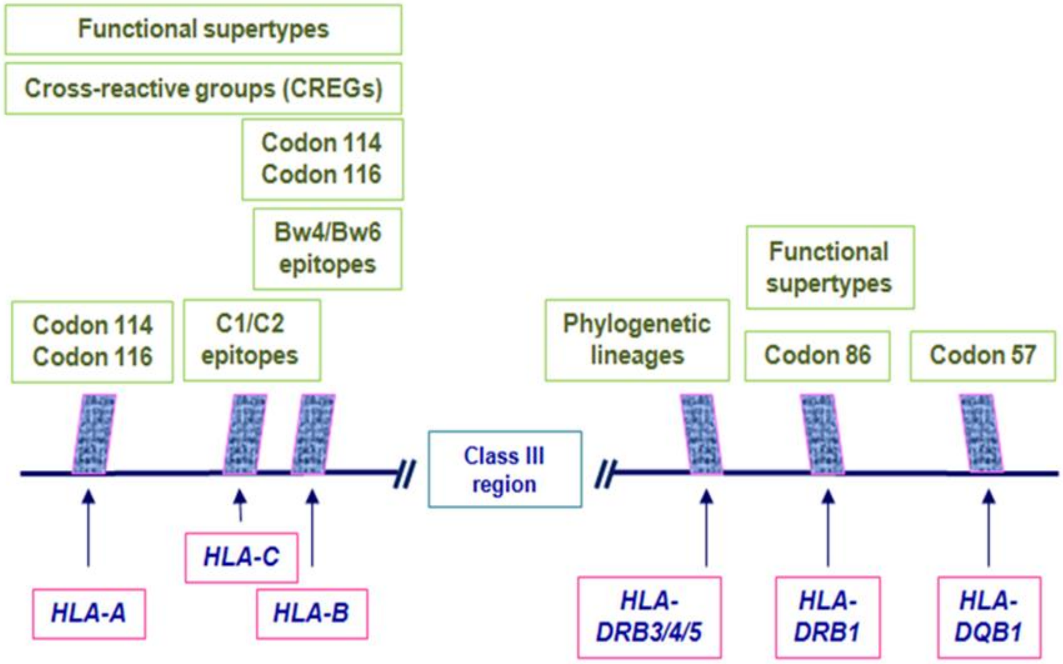

Figure 2. Well-known groupings of HLA alleles based on genetic, functional or evolutionary features.