Interrogating the major histocompatibility complex with high-throughput genomics

Paul I.W. de Bakker1,2,3,5,* and Soumya Raychaudhuri3,4,5,6,7

1Department of Medical Genetics, and 2Department of Epidemiology, University Medical Center Utrecht, Utrecht, The Netherlands, 3Division of Genetics, and 4Division of Rheumatology, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA, USA, 5Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA, 6Partners HealthCare Center for Personalized Genetic Medicine, Boston, MA, USA, and 7Faculty of Medical and Human Sciences, University of Manchester, Manchester, UK

Received August 19, 2012; Revised and Accepted September 6, 2012

The major histocompatibility complex (MHC) region on the short arm of chromosome 6 harbors the largest number of replicated associations across the human genome for a wide range of diseases, but the functional basis for these associations is still poorly understood. One fundamental challenge in fine-mapping associations to functional alleles is the enormous sequence diversity and broad linkage disequilibrium of the MHC, both of which hamper the cost-effective interrogation in large patient samples and the identification of causal variants. In this review, we argue that there is now a valuable opportunity to leverage existing genome-wide association study (GWAS) datasets for in-depth investigation to identify independent effects in the MHC. Application of imputation to GWAS data facilitates comprehensive interrogation of the classical human leukocyte antigen (HLA) loci. These datasets are, in many cases, sufficiently large to give investigators the ability to disentangle effects at different loci. We also explain how querying variation at individual amino acid positions for association can be powerful and expand traditional analyses that focus only on the classical HLA types.

INTRODUCTION

Across the entire genome, the major histocompatibility complex (MHC) region stands out by virtue of harboring the largest number of bona fide genetic associations for a range of conditions. These not only include inflammatory, autoimmune and infectious diseases but also different forms of cancer, drug-induced hypersensitivity and, more recently, neuropsychiatric disease. The earliest associations within the MHC were identified in the 1970s, using serological reagents and cell-based assays that measure different human leukocyte antigens (HLAs) (1,2). These initial studies revealed large effects that could be detected with relatively small sample sizes. Decades later, despite the impressive cumulative list of MHC associations, there has been a surprising lack of progress in pinpointing causal variants underlying these disease associations.

The MHC is often regarded as one of the most complex regions of the genome because of its enormous sequence diversity, broad linkage disequilibrium (LD) and high gene density (3–5). This complexity has severely hampered efforts to study the region in detail. As a result of LD, multiple nearby variants may have equivalent statistical evidence of association, and thus it can be difficult to assign causality to a specific variant (6). Inability to pin down a causal variant limits the ability of investigators to elucidate the functional mechanisms of validated associations and to understand disease pathogenesis. Even larger sample sizes than were required for their initial discovery will likely be needed to reach sufficient power to tease apart the role of correlated variants. The question remains to what extent imputation with dense reference panels (like 1000 Genomes) or genotyping with dense SNP arrays (like Immunochip) might help to narrow associations with functional variation at known loci (7,8). The MHC region provides a ripe opportunity to investigate these specific issues.

In this review, we highlight the potential for fine-mapping associations within the MHC. For these studies to be done

*To whom correspondence should be addressed at: University Medical Center Utrecht, Stratenum 0.310, PO Box 85500, 3508 GA Utrecht, The Netherlands. Tel: +31 887550406; Fax: +31 887555410; Email: pdebakker@umcutrecht.nl

© The Author 2012. Published by Oxford University Press. All rights reserved.
For Permissions, please email: journals.permissions@oup.com
effectively large numbers of samples need to be typed for HLA alleles, which are potentially driving many of the observed associations. Imputation-based approaches, specifically developed for the HLA genes within the MHC, make it possible to type classical alleles in silico for large sample sets with available data from genome-wide association studies (GWASs) (9–13). We make the case that such methods will likely add substantial value to the existing GWAS datasets. Our aim is to demystify some of the perceived complexities in this subfield and to propose a systematic way to interrogate genetic variation in the MHC.

**MHC IN A NUTSHELL**

The MHC comprises a contiguous 4 Mb region on the short arm of chromosome 6. The extended MHC (xMHC), as its name suggests, spans an even larger 7.6 Mb region and comprises more than 8000 unique alleles (14). The xMHC region include tumor necrosis factor, and members of the complement and antigen-processing pathways. Other MHC genes with a less obvious link to immunity include the histones, tRNAs and olfactory receptors. The combination of protein-changing and non-coding variation in these genes provides important insights into the evolution of the immune system, and can challenge the interpretation of observed genetic associations. The so-called 8.1 ancestral haplotype is a case in point; this haplotype spans essentially the entire MHC from class I to II and has been implicated in multiple immune disorders (14). Of course, there is also the possibility that non-HLA genes play a role in complex diseases; for example, C2/CFB in age-related macular degeneration (15).

**A COMPLEX NOMENCLATURE**

Extreme sequence variation is one of the characteristic features of HLA genes. This is evidenced by sequence data from the International Immunogenetics database (IMGT/HLA v 3.9); more than 8000 unique alleles have been documented for the classical HLA genes (16). With >2600 reported haplotypes, HLA-B is probably the most polymorphic gene of the genome. Within the HLA coding regions, some codons contain multiple SNPs (sometimes with more than two alleles), creating a range of possible amino acids at certain sites in HLA proteins. Over the years, an elaborate nomenclature system has been developed to catalog allelic variation observed in each HLA gene (17), firmly rooted in the original antibody-based serotyping used to study HLA molecular isotypes. In this nomenclature, the two-digit type (e.g. HLA-DRB1*01) refers to groups of alleles defined based on serological reagents (‘serotyping’) that were widely used before the advent of polymerase chain reaction (PCR)-based reagents and DNA sequencing. Since antibodies bind to the exposed extra-cellular aspects of an HLA molecule, differences in two-digit alleles reflect mostly variation in the exposed parts of HLA protein structures. The four-digit type (e.g. HLA-DRB1*01:01) refers to a unique amino acid sequence of the gene product; that is, synonymous changes are ignored. Synonymous changes within exons and non-coding changes outside exons are defined by six- and eight-digit types, respectively. All these haplotype definitions are codified by the standard HLA dictionary and use naming conventions that may seem rather cryptic to a non-expert. Figure 2 shows HLA-DRB1*01 and *02 alleles in a multiple sequence alignment, highlighting relations between two-, four- and six-digit types.

It is important to emphasize that the nomenclature does not naturally capture all SNPs, indels or polymorphic amino acid positions. Instead, it defines gene-wide haplotypes, where each unique haplotype gets its own unique number. Historically, the nomenclature has evolved from grouping HLA alleles according to serotype (at two-digit resolution) to a more complete classification on the basis of protein-changing and non-coding variation in DNA sequence (four-digit and beyond). The hierarchical structure of the HLA nomenclature reveals the implicit assumption about the relative functional impact of non-synonymous, synonymous and non-coding variation.

**GENOTYPING HLA ALLELES**

There is a long history in the immunogenetics community of sharing protocols, reagents and expertise, facilitated by the International HLA and Immunogenetics Workshop series (the 16th meeting took place in May/June 2012 in Liverpool, UK). The current standard for HLA typing involves the design of sequence-specific primers to isolate and amplify the desired HLA gene(s) of interest followed by Sanger sequencing. Most studies focus on capturing coding variations within the classical HLA genes (that is, to four-digit resolution), typically achieved by querying the exons coding for the peptide-binding grooves of the HLA proteins where most of the inter-individual diversity is concentrated. Indeed, a commonly used protocol for typing class I genes involves sequencing exons 2 and 3, whereas typing of class II genes is often based on sequencing only exon 2. More comprehensive ( unbiased) protocols exist, but are more costly due to the additional PCR steps and greater sequencing burden to cover the remaining exons.

One limitation of Sanger sequencing is that it is often difficult to resolve closely related alleles (referred to as genotyping ambiguity), because Sanger sequencing assays a diploid mixture with no straightforward way to establish the haplotype phase between nearby sites. Novel protocols based on
next-generation sequencing may offer an attractive alternative (18–20).

We note that it is not feasible to use high-throughput microarray-based SNP typing (which has propelled GWAS in the last few years) to query HLA genes directly, because the sequence diversity and polymorphism density makes it virtually impossible to design a robust set of oligonucleotide probes. As a result, most SNPs within the MHC on genotyping microarrays are located away from the polymorphic hotspots in the classical HLA genes. Their corresponding oligonucleotide probes for these SNPs can robustly differentiate between the known alleles and do not suffer from interference from neighboring SNPs within the probe sequence. In terms of SNP density, the earliest microarrays contain a few hundred SNPs across the MHC, whereas the more recent Immunochip includes thousands of SNPs.

### IMPUTATION OF CLASSICAL HAPLOTYPES

Instead of genotyping, it is possible to impute classical HLA alleles by inferring them from surrounding SNP markers across the MHC (4). In the simplest case, an HLA allele might be ‘tagged’ by an SNP nearby that is in perfect LD. Alternatively, an HLA allele might be tagged by a set (haplotype) of multiple markers. More complex haplotype patterns can be exploited using sophisticated imputation algorithms, which are now well established in the context of GWAS and meta-analysis (7).

Admittedly, SNP-based imputation is unlikely to replace classical typing for clinical applications, such as organ matching for transplantation, where accuracy in a single individual is critical. Imputation might however be an attractive alternative for large-scale association studies, considering that SNPs are not only easy and cheap to genotype but are also widely available due to the recent investment into GWAS. If errors are randomly distributed, and not systematic in nature, imputation might result in a loss in power without an increased false-positive rate. This makes imputation ideal for the re-analysis of the existing GWAS datasets.

Since the MHC is characterized by extensive LD, it is highly amenable to effective imputation, despite its high overall haplotype diversity. This LD is the consequence of a relatively low recombination rate (almost 3-fold lower compared with the genome-wide average of 1.2 cM/Mb) (4). The first high-resolution genetic map of the MHC was constructed by genotyping >6000 SNPs and by collecting classical four-digit HLA types in four HapMap populations (CEU for Northern and Western European, YRI for African, and CHB and JPT for East-Asian populations) (4). This public resource provided for the first time a detailed picture of the LD structure between SNPs and classical HLA alleles, and demonstrated that specific SNPs (often at considerable distance) were informative for classical HLA alleles. This led to the important realization that classical HLA alleles could be tagged and imputed through LD to surrounding SNPs.

Figure 1. Overview of the MHC region with validated GWAS associations from the NHGRI GWAS catalog (http://www.genome.gov/gwastudies/) plotted as a function of position along chromosome 6 (hg18). Strongest SNP associations are annotated by trait or disease. Classical HLA-A, -B, -C, -DP, -DQ and -DR genes are highlighted (red vertical ticks) among the many hundreds of genes across the region.
Prior to application of HLA imputation within the MHC, several studies have leveraged tight LD between SNPs and HLA alleles to interpret MHC associations. For example, rs2395029 emerged as the top hit for two unrelated GWAS of HIV-1 virus load set point (21) and drug-induced liver injury due to fluclouxacinil (22). This SNP lies about 100 kb upstream of HLA-B, but it was already known to be in near-perfect LD with HLA-B*57:01 in Europeans (4). This same SNP was also successfully evaluated as a screening tool for HLA-B*57:01-associated abacavir hypersensitivity (23). In another example, six tag SNPs were identified (24) and independently validated in another population to predict classical HLA-DQ alleles associated with celiac disease (DQA1*05/DQB1*02 (DQ2.5), DQA1*0201/DQB1*0202 (DQ2.2), DQA1*0505/DQB1*0301 (DQ7) and DQA1*03/DQB1*0302 (DQ8)) (25).

Investigators have since devised more sophisticated imputation methods ((9,11–13)) and with large reference panels these tools have since been applied to association studies. Most of these methods leverage the observation that chromosomes with high identity by descent within the MHC region are likely to share the same HLA alleles. Then for a given individual with SNP genotypes within the MHC, statistical methods can identify MHC segments with high identity-by-descent, and use that information to infer HLA alleles. We expect that the major determinant of imputation quality will be the sample size and representativeness of the reference panel. How ancestry differences between the reference panel and GWAS dataset affect imputation quality must be evaluated empirically, but the MHC may be more sensitive to this given the high degree of population differentiation of HLA alleles across populations.

Multiple sclerosis

Multiple sclerosis (MS) is an inflammatory disease of the central nervous system characterized by neurodegeneration and progressive disability. It has been well known that MHC variation has the largest genetic influence on disease risk (26), but attempts to fine-map the MHC association signal beyond the primary DRB1*15:01 association has proved challenging with conflicting results (27,28). Recently, a large-scale GWAS meta-analysis of 10 000 MS cases and 17 000 controls, drawn from multiple European populations, had adequate power for the discovery of 29 novel MS loci outside MHC (29). Investigators imputed classical HLA alleles with the HLA-IMP tool (12). As expected, the primary signal mapped to HLA-DRB1*15:01 (P < 10^−320). Subsequent conditional analysis identified additional classical allele associations with consistent effects across countries/cohorts for HLA-A*02:01, HLA-DRB1*03:01 (which was indistinguishable from HLA-DQB1*02:01), HLA-DRB1*13:03 and possibly for HLA-DPB1*03:01. These results suggest that the MHC association cannot be mapped to a single allele at a single locus—but rather constitutes independent effects at multiple loci. Further fine-mapping analysis proposed a possible model where three mutations on specific branches of the estimated genealogical tree of HLA-DRB1 were shared by the associated classical alleles at this locus. This model implicated an additional (risk) association for DRB1*08:01 by virtue of sharing the same putative mutation as DRB1*13:03; this hypothesis will need to be tested.

TESTING FUNCTIONAL VARIATION

We argue that HLA genes should be investigated much in the same way that we test variations outside of the MHC. That is, in addition to haplotype-based testing (tantamount to testing classical alleles for association), we assert that testing individual polymorphic sites within a gene might be informative, and may produce a more parsimonious model to explain an observed association. Since classical HLA alleles are defined by unique nucleotide sequences, variations at individual coding and non-coding sites at the nucleotide level can easily be inferred by simple lookup. Because the classical alleles also define unique amino acid sequences, it is equally straightforward to infer variations at polymorphic amino acid positions in the corresponding HLA proteins. Thus, it is possible to impute not only classical HLA alleles but also individual nucleotide variants and amino acid variants.
Our working hypothesis is that some amino acid positions in a protein are structurally or functionally more important than others, and can therefore be tested (through imputation) for association. This is particularly appropriate if disease susceptibility is modulated by specific amino acid sites (or multiple sites); in that case, a testing approach based solely on classical HLA alleles may not be as powerful. If a model based on specific amino acids can explain the data better than a model based on classical haplotypes, this should provide an additional biological insight as the association signal is much more localized compared with a four-digit allele, which by definition spans the entire gene/protein.

We note that the goal of explaining genetic associations in terms of protein structure and function is not a novel idea. Previously, amino acid associations have been described for HLA-DQB1 position 57 in type 1 diabetes (30), HLA-DRB1 position 74 in Graves’ disease (31,32), HLA-DRB1 position 11 in sarcoidosis (33), HLA-DRB1 positions 67, 71 and 74 in autoimmune polyglandular syndrome (34), HLA-DRB1 position 13 in MS (35), and HLA-DRB1 positions 37 and 86 in primary sclerosing cholangitis (36). Comparative analyses of a growing number of three-dimensional structures from the Protein Data Bank have focused mostly on canonical ‘pockets’ within the peptide-binding groove to rationalize classical HLA allele associations (37–39). An illustrative example comes from the crystal structure of HLA-DQ bound to an immunogenic epitope from dietary gluten (the known antigen in celiac disease), which provides structural clues as to how DQ2 molecules may preferentially bind proline-rich gluten peptides (40). For most autoimmune diseases, however, we still do not know the (auto)antigens involved. Fine-mapping functional variants may produce useful information to narrow the search for possible peptides presented by HLA molecules associated with disease risk.

By coupling HLA imputation in large GWAS datasets to association testing at amino acid resolution, we anticipate that the potential gains can be greater than that has been achieved so far. The combination of large sample sizes and the expected effect sizes in the HLA is likely to yield much better power for stepwise conditional analysis allowing more precise ranking of models by their goodness-of-fit. In addition, GWAS data also provide for excellent control against population stratification, a key confounding variable for HLA association studies (41). HLA alleles are strongly differentiated across global populations and therefore highly sensitive to type 1 error when slight differences in the ancestry of cases and controls are not adequately controlled for.

**Host control of HIV replication**

In the International HIV Controllers Study, we performed a GWAS in about 1000 HIV controllers and 2700 HIV progressors. HIV controllers constitute a relatively small subgroup of HIV-positive individuals able to suppress viral replication down to levels that often go below detection in the absence of treatment. We observed 313 SNPs at genome-wide significance, concentrated in the class I region. One of those SNPs (rs2395029, mentioned above) had already been described as an effective tag SNP for classical HLA allele B*57:01. To assess whether there were other independent associations, we imputed classical alleles and corresponding amino acid polymorphisms at the HLA-A, -B and -C loci, allowing us to test individual classical alleles as well as polymorphic amino acid positions. Three HLA-B amino acid sites (97, 67 and 70) clearly stood out and explained the MHC association significantly more effectively than the well-known classical alleles such as B*57(01) and B*27(05). These sites are located in the peptide-binding groove, highlighting the importance of the peptide–MHC interaction in the control of HIV. Position 97 ($P = 10^{-45}$) has as many as six common alleles and explained most of the observed MHC association (Figure 3). Three amino acids at this position (valine, asparagine and tryptophan) are carried exclusively by protective haplotypes (B*57:01, B*27:05 and B*14, respectively). Threonine is carried by protective haplotype B*52 (and others), while serine and arginine are carried by a variety of risk haplotypes including B*07 and B*35. We subsequently performed a similar amino acid-based analysis of HIV control in African Americans and were able to replicate position 97 as the most important polymorphism in HLA-B (42). In African Americans, position 97 explains the association of not only the B*57:01 allele (the most prevalent B*57 allele in Northern–Western European populations) but also the B*57:03 allele (the most prevalent B*57 allele in African Bantu populations). Interestingly, we also detected an effect at position 245 located far from the peptide-binding groove. Threonine at this position is uniquely carried by HLA-B*81:01, an allele that is essentially absent from European populations. It has been suggested by others that this position may influence the binding of the CD8 co-receptor to the HLA-B molecule. These findings underscore the value of fine-mapping in multiple ethnicities.

**Rheumatoid arthritis**

Separately, our group has applied a similar imputation and fine-mapping approach to understand the basis of the MHC association with rheumatoid arthritis (43). For more than 20 years, the widely accepted ‘shared epitope’ model postulates that MHC risk to RA is conferred by a short stretch of amino acids (positions 70–74) in one of the two alpha-helices that enclose the peptide-binding groove of the HLA–DR molecule (44). Taking advantage of available GWAS data in 5000 cases and 15 000 controls from six different cohorts, we imputed classical alleles and polymorphic amino acids and tested them for association. Surprisingly, the strongest association signal did not map to any of the shared epitope positions, but mapped to position 11 ($P < 10^{-581}$) located on the beta-sheet floor of the peptide-binding groove. Once position 11 was corrected for, we observed independent, but more modest effects, at the shared epitope positions 71 and 74. Adjusting for these DRB1 effects, additional amino acid associations were found in HLA-B and HLA-DPB1. Interestingly, the HLA-B effect is due to B*08, which is often described as part of the long 8.1 ancestral haplotype across the class I and II regions. The presence of both class I and II involvement suggests a role for both CD8+ cytotoxic T cells and CD4 helper T cells. The biological connection to other autoimmune diseases that share these same allelic associations is an interesting observation.
The PGF cell line is currently the only cell line relying on a single reference sequence (the sequence from consanguineous cell lines, selected on the basis of high frequency in European populations and their putative role in disease association). Recent studies however suggest a prominent role for the regulatory variation in the MHC, including eQTLs (45–48). This signals the urgent need for a more comprehensive approach to study all variation across the MHC, not limited to the coding regions of classical HLA genes. This will likely happen in two consecutive steps: first through imputation using an MHC-wide reference panel and second by complete sequencing.

There have been a few parallel attempts to characterize long-range haplotype sequences across the MHC. For example, the MHC Haplotype Project determined the full-length sequences of eight common MHC haplotypes derived from consanguineous cell lines, selected on the basis of high frequency in European populations and their putative role in type 1 diabetes and MS (49–51). These DNA sequences provide a valuable framework for polymorphism discovery and gene annotation. The extensive variation observed between these MHC haplotypes illustrates the limitations of relying on a single reference sequence (the sequence from the PGF cell line is currently the de facto human genome reference sequence for the MHC).

Whole-genome sequencing is still too expensive on a large scale, but can generate comprehensive coverage across the MHC. We are hopeful that soon this will become routine, but better approaches are needed for aligning divergent sequence reads to the reference sequence (or all eight reference MHC sequences). De novo assembly of short reads (without aligning to a reference) looks promising, but these approaches are still in their infancy (52). Some progress has been made with flow cytometric sorting of chromosome 6 for more cost-effective sequencing (as it accounts for ~6% of the whole genome), but this approach would also demand better alignment or assembly strategies. Once these analytical issues are resolved and large-scale sequencing becomes more affordable, it will be possible to obtain complete sequencing information. Whole-exome sequencing datasets have demonstrated a relative excess of rare, functional variants, reflecting the effects of negative selection, with the implication that sequencing in larger samples will continue to discover novel variants (53). The abundance of rare alleles may be even more pronounced in the HLA according to a recent analysis (54), underscoring the potential value of MHC sequencing when this can be done efficiently.

**CONCLUSION**

With HLA imputation, it will be possible to extract much more information from existing GWAS datasets with an initial MHC association signal. Large-scale reference panels and effective imputation methods are currently available to type HLA genes in silico and fine-map associations within classical HLA genes. Unfortunately, adequately sized reference panels are not available still for many ethnic groups. Another major limitation is that these reference panels lack MHC-wide ascertainment of variation. We are hopeful that the onslaught of whole-genome sequencing efforts will allow the construction of large reference panels for multiple global populations. With those emerging tools and resources, it will become possible to study the relative importance of coding and non-coding variation, and of rare and common variation, and the role of non-HLA genes. We hope that detailed insights into functional variations in the MHC will generate opportunities to develop novel therapies for a wide range of diseases.

**Conflict of Interest statement.** None declared.

**FUNDING**

We acknowledge financial support from the National Institutes of Health (1 R01 AR062886-01, 5K08AR055688), the Arthritis Foundation, and the Collaboration for AIDS Vaccine Discovery of the Bill and Melinda Gates Foundation. P.I.W.d.B. is a recipient of a VIDI Award from the Netherlands Organization for Scientific Research (NWO).

**REFERENCES**


