The dichotomy between disease phenotype databases and the implications for understanding complex diseases involving the major histocompatibility complex

P. M. Clark*, M. Kunkel* & D. S. Monos*

Summary

Many genes related to innate and adaptive immunity reside within the major histocompatibility complex (MHC) and have been associated with a multitude of complex, immune-related disorders. Despite years of genetic study, this region has seen few causative determinants discovered for immune-mediated diseases. Reported associations have been curated in various databases including the Genetic Association Database, NCBI database of clinically relevant variants (ClinVar) and the Human Gene Mutation Database and together capture genetic associations and annotated pathogenic loci within the MHC and across the genome for a variety of complex, immune-mediated diseases. A review of these three distinct databases reveals disparate annotations between associated genes and pathogenic loci, alluding to the polygenic, multifactorial nature of immune-mediated diseases and the pleiotropic character of genes within the MHC. The technical limitations and inherent biases imposed by current approaches and technologies in studying the MHC create a strong case for the need to perform targeted deep sequencing of the MHC and other immunologically relevant loci in order to fully elucidate and study the causative elements of complex immune-mediated diseases.

Introduction

In the time since the first sequence-based assembly and gene map of the human major histocompatibility complex (MHC) was reported (The MHC Sequencing Consortium, 1999), this ~4 Mb region on the short arm of chromosome 6 has been associated with numerous diseases and traits, remaining one of the most studied regions of the genome. The MHC is among the most gene-dense regions of the genome, with over 180 protein-coding genes, and is defined as the region spanning from the gamma-aminobutyric acid receptor (GABBR) gene on the telomeric side of the chromosome to the kinesin family member C1 (KIFC1) gene towards the centromere (HG19 coordinates, chr6:29570005-33377699) (Horton et al., 2004; Stewart et al., 2004; Shina et al., 2009).

Although the sequence and gene map of the MHC have facilitated the association of numerous traits through single nucleotide polymorphism (SNP) genotyping microarrays and genomewide association studies (GWAS), fine mapping of pathogenic mutations to discrete genomic loci remains elusive due to the MHC’s inherent sequence complexity and structure. The MHC is known to contain long stretches of conserved and nonconserved genes, forming a mosaic structure with broad linkage disequilibrium (LD) (Kumanovics et al., 2003; Horton et al., 2004; Stenzel et al., 2004; Miretti et al., 2005). This complexity makes it difficult to resolve discrete causal variant loci using SNP genotyping microarrays, as these platforms inherently rely on the interrogation of tag SNPs that are in LD with SNPs residing within the same local haplotype block (Peiffer & Gunderson, 2009). As a result of the highly polymorphic nature and LD structure of the MHC, a tag SNP may not adequately characterize the haplotype diversity and associated variants. Tag SNPs also do not address the fact that both parental alleles are expressed in specific MHC genes (e.g. HLA) forming very long-range haplotypes associated with specific phenotypes (Petersdorf et al., 2007, 2013), suggesting gene coordination over very long distances. Furthermore, deconvolution of discrete causal variant loci from within the
same haplotype block, each with equal statistical significance, remains a significant challenge (particularly for complex traits), effectively hindering the fine mapping of causal, pathogenic mutations (Ioannidis et al., 2009). This is largely why numerous traits and diseases have been associated with genes and regions within the MHC by GWAS (de Bakker & Raychaudhuri, 2012; Welter et al., 2014) without the discovery of discrete causal variant loci.

This dichotomy between genes associated with specific phenotypes and identified pathogenic, disease susceptibility loci is apparent when comparing three seminal databases, including the Genetic Association Database (GAD) (Becker et al., 2004), NCBI database of clinically relevant variants (ClinVar) (Landrum et al., 2014) and the Human Gene Mutation Database (HGMD) (Stenson et al., 2012). The GAD is a submitter-curated repository of genetic association data and contains genes that are both positively and negatively associated with a variety of phenotypes (Becker et al., 2004). Although GAD was retired in early 2014, it remains a valuable resource of genetic association information retrieval, with over 160 000 annotated associations, curated from over 56 000 references. The HGMD is an expert-curated database of over 45 000 publications and contains over 150 000 germ line mutations in nuclear genes that are associated with inherited diseases (HGMD Professional version 2014.2) (Stenson et al., 2012). NCBI repository of medically relevant variants and associated phenotypes, ClinVar, contains over 260 000 user-curated associations (Landrum et al., 2014).

In reviewing these databases, we find that nowhere else in the genome is the disparity between associated loci, identified by traditional GWAS approaches and fine mapped disease susceptibility loci, more pronounced than within the MHC. This is despite the MHC’s persistent identification as the most associated region to complex immune disorders. We believe that this discordance arises from several factors including the sequence and structural complexity of the MHC, the limitations of traditional GWAS approaches and the pleiotropic nature of MHC genes involved in driving immune-mediated disease. To overcome these limitations and comprehensively evaluate the involvement of all variation within the MHC with complex diseases, we believe that future research should focus on more informative and comprehensive genetic approaches, including targeted deep sequencing of the entire MHC region.

**Evaluating database discordance**

We evaluated the three databases for concordance between regions of the genome associated with disease and regions where pathogenic variants have been identified. While our primary focus was on the MHC, we first considered all regions throughout the genome with high-magnitude peaks, exploring inherent caveats of each database and nuances captured by the data. To accomplish this, each data set was first condensed into a BED file containing unique combinations of genomic locations and the associated phenotype, and duplicate entries were removed. It is worth noting, however, that a single genomic location could be associated with multiple traits and a single trait could be associated with multiple loci. Trait-associated variant density plots were generated using the three filtered data sets (GAD, HGMD and ClinVar) and are shown for each chromosome (Supporting Information). The heights of the peaks are proportional to the number of trait-associated variants within a region, although a single genomic position may be counted multiple times as it may be associated with multiple phenotypes. As ClinVar is fundamentally a subset of HGMD, with 76% of pathogenic ClinVar variants also annotated by HGMD as pathogenic, we focus our subsequent review primarily on the GAD and HGMD data sets.

**Genomewide inspection of variants**

Generally, we observed three characteristic types of trait-associated variant density plots: loci in which there is a GAD peak and no annotated pathogenic HGMD loci; genes that harbour only HGMD loci and are not associated with GAD; and loci that are both associated with GAD and harbour pathogenic loci (Fig. 1a–c). Genome wide, there are 6031 of 8139 (74%) genes with annotated GAD associations that harbour no HGMD pathogenic variants and 1681 of 3789 (44%) genes with HGMD-annotated pathogenic mutations that do not have any associations within GAD. Thus, there are only 2108 genes that have both annotated GAD association and at least one pathogenic locus annotated by HGMD. From Fig. 1, it is evident that, generally, HGMD and ClinVar provide more granular information, annotating discrete pathogenic loci, whereas GAD simply reports associated loci, which may include an entire haplotype block, gene or discrete locus associated with a particular phenotype.

**HGMD-identified variants**

Beyond the granularity of the HGMD data set, our analysis also reveals that there are numerous high-magnitude HGMD peaks throughout the genome. Understanding the distinction between these peaks and databases is then critical to understanding the impact of trait-associated variant densities within the MHC. By scanning the genome, we evaluated high-magnitude HGMD peaks (>1000) on chr5, chr15 and chr16 and high-magnitude ClinVar peaks on chr13 and chr17 (Supporting Information) which harbour a high number of annotated HGMD or ClinVar pathogenic loci and very few if any GAD trait-associated loci. The HGMD peak on chr5 corresponds to pathogenic variants within the adenomatous polyposis coli (APC) gene. All of the pathogenic variants annotated within
this ~108-kbp locus confer susceptibility to APC and related cancers. The APC protein is an integral part of the beta-catenin signalling pathway and plays a primary role in tumour suppression by antagonizing the wingless-type (WNT) signalling pathway (Hanson & Miller, 2005; Sansom et al., 2007). The HGMD peak on chr15 corresponds to pathogenic variants within the fibrillin 1 (FBN1) gene, which encodes for the ubiquitously expressed fibrillin glycoprotein, a major constitutive element of extracellular microfibrils. These microfibrils provide force-bearing structural support in connective tissues throughout the body. The vast majority of pathogenic mutations in the ~237-kbp FBN1 locus confer susceptibility to Marfan syndrome and related conditions. The HGMD peak on chr16 corresponds to the pathogenic variants within the
genic diseases. Fer susceptibility to a variety of closely related mono-
eresponses, including cancer, a variety of vascular diseases, cleft palate, pre-eclampsia, Down syndrome, glaucoma, depression and schizophrenia. An additional high-magnitude GAD peak appears on chromosome 17, within the angiotensin I converting enzyme (ACE) gene, which is dipeptidyl carboxypeptidase and plays an important role in blood pressure regulation and electrolyte homeostasis. GAD has over 1000 phenotypes associated with the ACE locus, including many disease terms related to cardiovascular disease, various cancers, diabetes, Alzheimer’s disease and schizophrenia. Unlike the peaks seen in the HGMD and ClinVar, these GAD peaks (including those in the MHC) are associated with pleiotropic loci, which are involved in a variety of polygenic physiological processes and related disease pathways.

A closer look at variants associated with the MHC

The chromosomal variant density plots for each chromosome (Supporting Information) reveal a broad, high-magnitude GAD peak within the MHC (Fig. 2a). While the MHC has the highest density of associated GAD loci of any location in the human genome (Fig. 2b,c), there are disproportionately few HGMD-annotated pathogenic loci within the same region (Fig. 2d). Only 84 of the 181 (46.4%) associated GAD genes within the MHC harbour pathogenic loci (as annotated by HGMD) and yet every gene that harbours an annotated HGMD pathogenic locus within the MHC is also associated with at least one trait within GAD. All of the phenotypes associated with genes of the MHC and pathogenic loci were tabulated from the three databases to evaluate the frequency of phenotype-specific variants within the MHC (Table 1). While it is evident that there are a number of low-magnitude HGMD peaks within the MHC, these pathogenic variant loci are located within only 21 genes. For this analysis, only the unique set of paired genomic locations and annotated phenotypes were considered for each database. A word cloud generated from the tabulated data (Fig. 3) indicates phenotype commonality by each word’s size being proportional to the frequency of trait-associated loci or pathogenic loci. Given the important role of MHC genes in innate and adaptive immunity, it is not surprising that many complex autoimmune diseases including lupus, diabetes, rheumatoid arthritis, multiple sclerosis and IBD are more prominently indicated (have the highest number of associated or pathogenic loci within the MHC).

Why the MHC is lacking in causative variants

It is evident that the MHC has the highest density of GAD-annotated associations anywhere in the human genome (Supporting Information and Fig. 2). However, despite the high density of GAD annotations, there are relatively few HGMD-annotated pathogenic loci. We suspect that this may be primarily attributed to two inherent characteristics of the MHC. First, the high density of genes related to innate and adaptive immunity within the MHC and their likely shared involvement in aetiological mechanisms contributing to complex immune-mediated diseases has led to the identification of many pleotropic loci, the vast majority of which contain common variants with relatively small effect sizes. Secondly, the inherent design of SNP genotyping arrays (which rely on tag SNPs within LD blocks to identify associated loci) and, more recently, the implementation of whole-exome sequencing (WES, which limits variant analysis to coding regions only) limit the ability to determine
variability in regions that are not in LD with a tag SNP or regions outside of exons, respectively. These limitations reduce the effectiveness of genome surveys for fine mapping causal variants.

The distinguishing features of GAD and HGMD regions

Comparisons between databases identify a variety of regions with trait-associated variants in the GAD (Fig. 1a) that are absent causal variants in HGMD and ClinVar, even for very highly associated regions (MHC, MTHFR, ACE). Conversely, a number of regions with causally linked variants in HGMD and ClinVar have no corresponding association data curated by the GAD (Fig. 1b), even when multiple phenotypes have been determined (APC, FBN1, TSC2, LDLR, BRCA1 and BRCA2). There are also regions across the genome that show phenotype-associated variants in both the GAD and HGMD, indicating both associated and causal variants have been determined and are coincident (Fig. 1c). While it is not unusual for a GWAS-associated SNP to have unsupported evidence of causality, it appears odd that causally identified SNPs in the HGMD or ClinVar have no associated data in the GAD.

Primarily the distinction appears to be one of disease complexity and rarity. Monogenic disorders (i.e. cystic fibrosis) and rare disorders (often familial) are often excluded from GWAS as the causative gene is either already known, or the prevalence of disease is low in the general population. This type of data shows up as a strong phenotype-associated peak in HGMD with no (or very little) corresponding peak in the GAD. Many of the listed HGMD genes are associated with rare familial disease found in only a few families. In comparison, genes from the HGMD also found in the GWAS catalogue tend towards more housekeeping/common pathway or broad function genes that, not surprisingly, would be involved in a more diverse spectrum of possible disorders (collagen gene, cytochromes, etc.). But regions associated with complex disorders with no obvious causative elements display the opposite effect: a strong peak in GAD with no corresponding peak in HGMD. Beyond discrete GAD-identified genes spread across the genome, the MHC stands out in particular (Fig. 2a), not only for the quantity of associated phenotypes, but also for the sheer size of the region involved and the relative paucity of identified causal variants (Fig. 2b). Hundreds of thousands of base pairs encompassing the HLA-A, B, C DR and DQ genes show very high phenotype density along with non-HLA (class III) genes like complement, TNF, TAP and MICA/B (to name a few) (Fig. 2c). The small number of clinically

Figure 2. The major histocompatibility complex (MHC) on chromosome 6 (a) (sliding window average of 80-kbp window at 1000-bp intervals) was more closely examined by separately plotting the Gad and Human Gene Mutation Database (HGMD) data using a sliding window average of 1-kbp window at 100-bp intervals) (b–d). The high number of associations in the MHC region (b,c) (particularly for the MHC genes, b) is in stark contrast to the paltry number of causative variants found in (d) the HGMD database, clearly indicating the dichotomy between the identification of disease-associated vs. disease-causing variations.
The limitations of GWAS

Genomewide association studies by their nature study association between a polymorphic site within any given region and a (disease) phenotype. The density of the GWAS array being used, the individual variation within the sample under study, the LD of the genetic region and the population under study all impact GWAS results and its effectiveness. At its very best, GWAS can only indicate a region of interest, and it cannot necessarily discern the actual causative element. For monogenic disorders, GWAS is either unnecessary or can be performed in limited amounts to localize the gene of interest. But for complex disorders, GWAS has limited ability to interconnect the causative elements of disease. Technologies to support GWAS have evolved to maximize effectiveness (pinpoint the associated region) through greater probe (SNP) density (now millions of SNPs per array) allowing for more closely spaced SNPs across a given region. More recently, Illumina Immunochip has been developed to interrogate 195,806 SNPs and 718 small insertions–deletions, with 8459 variants within the MHC alone (Parkes et al., 2013). Despite the advances in the development of new and improved SNP arrays with increased probe densities and ever-increasing sample sizes for GWAS, there remain a number of technical challenges and limitations associated with using fixed arrays. Rosenfeld et al. (2012) point out that our recent improved understanding of the variability of the human genome through large-scale sequencing projects (identifying millions more polymorphisms) can lead to substantial confounding issues and even erroneous results when using fixed arrays due to differences between the reference genome (used to create array-based probes) and actual individual genomes. Regardless of the array being used, anywhere from 13% to 40% of the array probes on any Illumina or Affymetrix platform are negatively impacted by nonreference genome-annotated polymorphisms and structural variants, leading to confounding GWAS data. It was noted that 51% of array probes are affected by polymorphisms within 10 bp of the SNP site of interest. In addition, the fact that LD blocks are significantly reduced in size as the number of SNPs increases and even the concept of using tag SNPs to target disease-causing variation is suspected (Rosenfeld et al., 2012). This physical complexity underlies both the potentially invalid assumptions of existing bioinformatic programmes and the potentially nonconcordant genetic data produced from different groups (19% difference between the 1000 Genomes Project data and the Complete Genomics data sets), even when using the same DNA samples.

SNP genotyping data have been utilized to infer HLA genotypes, which have also been associated with a number of particular phenotypes (Lessard et al., 2013; Chang et al., 2015; Ollila et al., 2015). Although the latest HLA genotype imputation algorithm, HIBAG (Zheng et al., 2014), exhibits improved accuracy (typically >90%) and call rates as compared to previous HLA genotype imputation methods such as HLA*IMP (Dilthey et al., 2011, 2013) and BEAGLE (Browning & Browning, 2007), the accuracy of imputed HLA alleles remains variable between ethnicities and specific HLA loci, due to limited data sets to adequately train the algorithm, particularly for rare alleles (Levin et al., 2014). As a result, inferred HLA genotypes are not capable of adequately resolving the genetic contributions of low frequency rare variants within HLA loci and such results must be objectively interpreted. Furthermore, HLA imputation methods are only able to resolve
HLA genotypes at the two field resolution level and may not be used to infer any associations between synonymous variations or variations within noncoding regions of HLA alleles. However, despite these limitations, HLA genotype imputation has proven to be a valuable tool in associating specific HLA alleles with various traits from large-scale SNP genotyping data in the absence of targeted HLA sequencing data (Sanchez-Mazas & Meyer, 2014).

The limited utility of exome sequencing

In comparison, next-generation sequencing (NGS) has significantly improved our ability to evaluate disease-associated regions on a base by base level. And as sequencing costs decrease and data output increases, even small benchtop systems like the Illumina MiSeq make feasible and cost-effective the ability to evaluate all exonic polymorphic content within a given region. A number of exome-based sequencing kits have been developed to provide insight into disease processes by way of protein coding changes. But this exon-only approach greatly oversimplifies the genetics of complex disorders and ignores basic understandings of often noncoding gene/genomic control overexpression, translation, regulation and protein presentation.

de Bakker & Raychaudhuri (2012) point out that while imputation can support a better understanding of the independent effects of variation within the MHC (using the existing large GWAS data sets), only complete sequencing of the MHC will support determining to what extent noncoding variation plays a role in disease association and immune dysfunction. Research by Vandiedonck et al. (2011) and others clearly indicates that both gene expression and regulation (modulated by coding, noncoding and intergenic variation) play a large role in determining immune function. Recent work has demonstrated that variation in the 3′ untranslated region of HLA-C could significantly influence the response to HIV infection through moderation of HLA-C expression (Kulkarni et al., 2011). Combined with the knowledge that rare variants likely play a much greater role in both immune flexibility and function than initially thought (Kiezun et al., 2012; Klitz et al., 2012), these data lead ourselves and others to conclude that the MHC will only reveal its true complexity when all variation (coding, noncoding and intergenic) is evaluated. The concept of complete variation analysis should help to explain why many associated SNPs within the MHC have very small effect sizes. It is very likely that these associated SNPs are involved in a combinatorial manner with other coding and noncoding SNPs that together would demonstrate a much higher effect. Therefore, without evaluating noncoding regions of the genome, we are largely ignoring these additional factors that most certainly influence, if not outright cause, disease.

Targeted deep sequencing as the right approach to investigating the MHC and other complex genomic regions

With the advent of personalized medicine, facilitated by NGS platforms and particularly as newer targeted sequencing approaches are developed, the opportunity to understand and analyse genetic complexity within the MHC in the light of complex disease is at hand. We suggest that without a comprehensive, high depth sequencing approach that (i) allows for the detection of all variants, (ii) can lead to the de novo assembly of the MHC and (iii) can allow for long-range haplotype structure to be fully resolved, the MHC will continue to be resistant to determining causative variants in complex disorders. This assumes that causative variants (in the sense that they alone cause disease) are there at all. The more likely scenario is that complex diseases are caused by a combination of variants (both cis and trans) that individually are ‘necessary’ in order for the disease to develop, but by themselves are not ‘causative’. Therefore, looking for ‘causative’ variation may indeed be a search for combinations of both rare and common variation, both coding and noncoding, from within and outside the MHC.
A good example of this scenario is the rare disorder, bare lymphocyte syndrome, in which the actual cause of the disease is not defective MHC class II molecules, but rather defective transcriptional regulation from multiple transcription factors on distal chromosomes (Reith & Mach, 2001). Here, the requisite genes of the immune system are intact, but nonfunctional due to failed regulational control (lack of expression). Recent work has demonstrated epistatic effects through specific combinations of KIR3DL1 alleles and HLA-B loci that influenced both AIDS progression and HIV RNA abundance (Martin et al., 2007). Additional studies by Petersdorf et al. (2007, 2013) who examined individual SNPs in the context of long-range physical haplotype structure confirmed that single polymorphic positions, separated by mega-bases, can influence immunologic (transplantation) outcomes that are associated with specific haplotypes, not simply specific genotypes. This long-range gene interaction and LD across distal genes is an indicator of both the cooperative evolution of the MHC’s genes and its complex coordination with the immune system as a whole. This multigene–multihaplotype (both parental alleles are expressed in the HLA genes) interaction therefore likely predisposes immune diseases to involve multiple genes through variations that by themselves may not be clinically significant (i.e. associated with), but whose collective influence can lead to gene dysregulation and multigene miss-function (i.e. pathogenic outcome). Knight et al. point out that beyond the difficulty in determining disease causation by polymorphic and structural variation alone, gene expression differences in the HLA class II molecules may be equally responsible for the observed effect in autoimmune disorders (Handunnetthi et al., 2010).

The assessment by Parkes et al. in evaluating shared loci between autoimmune diseases clearly indicates the pleiotropic nature of these disorders. Ankylosing spondylitis, coeliac disease, inflammatory bowel disease, psoriasis, rheumatoid arthritis and type I diabetes all have significant overlap between disease-associated loci, even when comparing differing disease classes (sero positive vs. sero negative). And while the MHC genes often remain the most strongest associated with a specific disease, individual diseases are associated with individual HLA alleles, haplotypes or even specific populations (Mitsunaga et al., 2012). The very fact that multiple independent signals have been reported for diseases such as multiple sclerosis, T1D, rheumatoid arthritis and ankylosing spondylitis clearly indicates that other influences, possibly both MHC derived and non-MHC, are at play. This is particularly apparent in the number of discordant associations found in autoimmune diseases. As Parkes et al. point out, variants at IL27, IL10, STAT3, CD40 and FCGR2A while associated with increased risk for Crohn’s disease, ulcerative colitis, Behçet’s disease and ankylosing spondylitis are protective for T1D, rheumatoid arthritis, systemic lupus erythematosus and multiple sclerosis. These examples and others both within and outside the MHC clearly indicate that multiple functional pathways are involved that are impacted differently based upon the combination of gene interactions affected by each individual’s genetic make-up.

Conclusions

The past decade has seen remarkable progress in defining the genetic basis of disease and nondisease phenotypes. Array-based technologies like those from Affymetrix and Illumina have provided highly accurate methods for determining single nucleotide variation across the genome. These efforts have allowed the development of imputation methods that support the screening of large populations of patients and have led to critical insights into the genetics of normal and disease states. Technological innovations in NGS have made it possible to investigate the genome at the exonic level with great efficiency and accuracy and have led to remarkable progress in understanding the biological mechanisms of certain diseases. Together, these technologies have made it possible to study human disease in the context of evolution and biological function at a level that was simply not possible just a decade ago. These technologies have led to highly valuable public databases like the GAD, HGMD, ClinVar and others that have supported record progress in the determination of causal variants in monogenic diseases. But complex, multigenic disorders like those associated with the MHC (that are also influenced by environmental and epigenetic factors) have not shared in the success (causal determination) of monogenic disease research. And while significant efforts to try and correct this deficiency are underway, it is clear that current GWAS or even exomic sequencing technology alone will be insufficient to resolve the problem.

The inherent limitations of these technologies will never be able to fully answer the outstanding questions surrounding the study of complex diseases. SNP arrays by their very nature cannot take into account the millions of unique polymorphisms that are now clearly understood to exist in any given patient. The very fact that these non-reference-based polymorphisms exist will always undermine the accuracy and utility of GWAS data. Exome sequencing, although highly useful and informative, can never account for noncoding variation that directly influences or causes disease by way of effects in translational control, transcription, intracellular processing, presentation and cellular interaction. Combined with the highly polymorphic, mosaic nature of the MHC, these limitations, as indicated by the divergence between the GAD and HGMD data sets, are significant hurdles that must be overcome to truly understand how complex diseases are influenced by genetic variability and complex gene interactions.

We believe that recent developments in targeted sequencing approaches, de novo sequencing algorithms and long-range haplotype structure determination, are
absolutely needed to fully evaluate and understand the complexity of the MHC. Without a complete understanding of the variation, structural interaction and gene control within this complex region, we will never be able to fully determine the ‘necessary’ and possibly directly causative elements of disease. We believe the (perhaps) obvious distinction between the GAD and the HGMD in regard to associated vs. pathogenic loci is not simply a matter of lack of data (i.e. the need for more GWAS or even exome sequencing with ever larger and diverse populations), but that more complete data will make all the difference in defining the genetic basis of complex disorders involving the MHC and other complex genomic regions. By evaluating all genetic complexity, in the context of long-range interactions and haplotype structure, we also believe that the concept of complex disease causation will need to be reconsidered. It is quite probable that specific, individual variation (both common and unique) by itself is never causative of complex disease. Instead, it is only through the coordination of these disease facilitating variants in combination with environmental factors that complex disease will develop. If this hypothesis is correct, then current GWAS and exome sequencing projects will likely never find the majority of causes for complex disorders, no matter the diversity or size of the population under study. And while a more comprehensive targeted approach to sequencing the MHC should be considerably more informative, we must not neglect to include epigenetic and environmental factors that also likely play a role in diseases related to the MHC. Therefore, while no one method may address all the needs for studying complex diseases of the MHC, high-resolution targeted sequencing is the next appropriate step in the process.

References

de Bakker, P.I. & Raychaudhuri, S. (2012) Interrogating the major histocompatibility complex with high-throughput genomics. Human Molecular Genetics, 21, R29.
Becker, K.G., Barnes, K.C., Bright, T.J. & Wang, S.A. (2004) The genetic association database. Nature Genetics, 36, 431.
Browning, S.R. & Browning, B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American Journal of Human Genetics, 81, 1084.
Chang, A.L., Raber, I., Xu, J., Li, R., Spitale, R., Chen, J. et al. (2015) Assessment of the genetic basis of rosacea by genome-wide association study. The Journal of Investigative Dermatology, 135, 1548.
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, 968.
Dilthey, A., Leslie, S., Moutsianas, L., Shen, J., Cox, C., Nelson, M.R. & McVean, G. (2013) Multi-population classical HLA type imputation. PLoS Computational Biology, 9, e1002877.
Handunnetthi, L., Ramagopalan, S.V., Ebers, G.C. & Knight, J.C. (2010) Regulation of major histocompatibility complex II gene expression, genetic variation and disease. Genes and Immunity, 11, 99.

Hanson, C.A. & Miller, J.R. (2005) Non-traditional roles for the Adenomatous Polyposis Coli (APC) tumor suppressor protein. Gene, 361, 1.
Horton, R., Wilming, L., Rand, V., Loving, R.C., Bruford, E.A., Khodiyar, V.K. et al. (2004) Gene map of the extended human MHC. Nature Reviews Genetics, 5, 389.
Inoki, K., Ouyang, H., Zhu, T., Lindvall, C., Wang, Y., Zhang, X. et al. (2006) TSC2 integrates Wnt and energy signals via a coordinated phosphorylation by AMPK and GSK3 to regulate cell growth. Cell, 126, 955.
Ioannidis, J.P., Thomas, G. & Daly, M.J. (2009) Validating, augmenting and refining genome-wide association signals. Nature Reviews Genetics, 10, 318.
Kiezun, A., Garimella, K., Do, R., Stitzel, N.O., Neale, B.M., McLaren, P.J. et al. (2012) Exome sequencing and the genetic basis of complex traits. Nature Genetics, 44, 623.
Klitz, W., Hedrick, P. & Louis, E.J. (2012) New reservoirs of HLA alleles: pools of rare variants enhance immune defense. Trends in Genetics, 28, 480.
Kulkarni, S., Savan, R., Qi, Y., Gao, X., Yuki, Y., Bass, S.E. et al. (2011) Differential microRNA regulation of HLA-C expression and its association with HIV control. Nature, 472, 495.
Kumanovics, A., Takada, T. & Lindahl, K.F. (2003) Genomic organization of the mammalian MHC. Annual Review of Immunology, 21, 629.
Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M. & Maglott, D.R. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research, 42, D980.
Lessard, C.J., Li, H., Adrianto, I., Ice, J.A., Rasmussen, A., Grundahl, K.M. et al. (2013) Variants at multiple loci implicated in both innate and adaptive immune responses are associated with Sjogren’s syndrome. Nature Genetics, 45, 1284.
Levin, A.M., Adrianto, I., Datta, I., Iannuzzi, M.C., Trudeau, S., McKeigue, P., Montgomery, C.G. & Rybicki, B.A. (2014) Performance of HLA allele prediction methods in African Americans for class II genes HLA-DRB1, -DQB1, and -DPB1. BMC Genetics, 15, 72.
Martin, M.P., Qi, Y., Gao, X., Yamada, E., Martin, J.N., Pereyr, F. et al. (2007) Innate partnership of HLA-B and KIR3DL1 subtypes against HIV-1. Nature Genetics, 39, 733.
Miretti, M.M., Walsh, E.C., Ke, X., Delgado, M., Griffiths, M., Hunt, S. et al. (2005) A high-resolution linkage disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. American Journal of Human Genetics, 76, 634.
Mitsunaga, S., Suzuki, Y., Kuwana, M., Sato, S., Kaneko, Y., Homma, Y. et al. (2012) Associations between six classical HLA loci and rheumatoid arthritis: a comprehensive analysis. Tissue Antigens, 80, 16.
Ollila, H.M., Ravel, J.M., Han, F., Faraco, J., Lin, L., Zheng, X. et al. (2015) HLA-DRB1 and HLA class I confer risk of and protection from narcolepsy. American Journal of Human Genetics, 96, 136.
Ozcans, U., Ozcans, L., Yilmaz, E., Duvel, K., Sahin, M., Manning, B.D. & Hotamisligil, G.S. (2008) Loss of the tuberous sclerosis complex tumor suppressors triggers the unfolded protein response to regulate insulin signaling and apoptosis. Molecular Cell, 29, 541.
Parkes, M., Cortes, A., van Heel, D.A. & Brown, M.A. (2013) Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nature Reviews Genetics, 14, 661.
Peiffer, D.A. & Gunderson, K.L. (2009) Design of tag SNP whole genome genotyping arrays. *Methods in Molecular Biology*, 529, 51.

Petersdorf, E.W., Malkki, M., Gooley, T.A., Martin, P.J. & Guo, Z. (2007) MHC haplotype matching for unrelated hematopoietic cell transplantation. *PLoS Medicine*, 4, e8.

Petersdorf, E.W., Malkki, M., Horowitz, M.M., Spellman, S.R., Haagenson, M.D. & Wang, T. (2013) Mapping MHC haplotype effects in unrelated donor hematopoietic cell transplantation. *Blood*, 121, 1896.

Reith, W. & Mach, B. (2001) The bare lymphocyte syndrome and the regulation of MHC expression. *Annual Review of Immunology*, 19, 331.

Rosenfeld, J.A., Mason, C.E. & Smith, T.M. (2012) Limitations of the human reference genome for personalized genomics. *PLoS ONE*, 7, e40294.

Sanchez-Mazas, A. & Meyer, D. (2014) The relevance of HLA sequencing in population genetics studies. *Journal of Immunology Research*, 2014, 971818.

Sansom, O.J., Meniel, V.S., Muncan, V., Phesse, T.J., Wilkins, J.A., Reed, K.R., Vass, J.K., Athineos, D., Clevers, H. & Clarke, A.R. (2007) Myc deletion rescues Apc deficiency in the small intestine. *Nature*, 446, 676.

Shiina, T., Hosomichi, K., Inoko, H. & Kulski, J.K. (2009) The HLA genomic loci map: expression, interaction, diversity and disease. *Journal of Human Genetics*, 54, 15.

Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shaw, K. & Cooper, D.N. (2012) The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Curr Protoc Bioinformatics*, Chapter 1: Unit1.13.

Stenzel, A., Lu, T., Koch, W.A., Hampe, J., Guenther, S.M., De La Vega, F.M., Krawczak, M. & Schreiber, S. (2004) Patterns of linkage disequilibrium in the MHC region on human chromosome 6p. *Human Genetics*, 114, 377.

Stewart, C.A., Horton, R., Allcock, R.J., Ashurst, J.L., Atrazhev, A.M., Coggill, P. et al. (2004) Complete MHC haplotype sequencing for common disease gene mapping. *Genome Research*, 14, 1176.

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

Vandiedonck, C., Taylor, M.S., Lockstone, H.E., Plant, K., Taylor, J.M., Durrant, C., Broxholme, J., Fairfax, B.P. & Knight, J.C. (2011) Pervasive haplotypic variation in the spliceo-transcriptome of the human major histocompatibility complex. *Genome Research*, 21, 1042.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H. et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42, D1001.

Zheng, X., Shen, J., Cox, C., Wakefield, J.C., Ehm, M.G., Nelson, M.R. & Weir, B.S. (2014) HIBAG–HLA genotype imputation with attribute bagging. *The Pharmacogenomics Journal*, 14, 192.

Supporting Information

Additional supporting information may be found in the online version of this article:

Data S1 Trait associated variant density plots for each chromosome using the GAD, HGMD and ClinVar databases.