Genome-wide Association Studies: Findings at the Major Histocompatibility Complex Locus in Psychosis

Aiden Corvin and Derek W. Morris

The major histocompatibility complex (MHC) is one of the most intensively investigated, genetically diverse regions of the genome. In its extended form, it encodes more than 400 genes critical to immunity but is also involved in many other functions. In 2009, three simultaneously published genome-wide association studies (GWAS) reported the first compelling evidence for involvement of the MHC in schizophrenia susceptibility. In this review, we describe the structure and function of the MHC, discuss some of the challenges for genetic analysis of the region, and provide an update on findings from GWAS studies before describing potential approaches to interpreting the role of the locus in schizophrenia etiology. The GWAS literature supports involvement of the MHC locus in schizophrenia susceptibility. Current evidence suggests that the MHC plays a more significant role in schizophrenia susceptibility than in other psychiatric disorders. Because of the substantial diversity at the locus, there are differences in the implicated risk variants between ancestral groups, as there are for many other disorders. This is somewhat different than the pattern emerging at other loci. The association findings presently capture large genomic regions, with at least some evidence to suggest that multiple signals may be involved. Based on notable successes in other disorders, we suggest approaches to refining association signals at the locus. Finally, we discuss that these genetic data may be used to understand how the MHC contributes to the complex etiology of schizophrenia.

Key Words: Genome-wide association studies (GWAS), human leukocyte antigen (HLA), imputation, major histocompatibility complex (MHC), psychosis, schizophrenia

Discovered in the mouse in 1936, the major histocompatibility complex (MHC) is one of the most investigated, genetically diverse regions of the genome (1). The first MHC gene products were discovered on the surface of leukocytes, so the locus also became known as the human leukocyte antigen (HLA) region. Major histocompatibility complex involvement in schizophrenia etiology has been the subject of speculation, and the first genetic investigation was reported in 1974 (2). This and subsequent studies were inconclusive, until three genome-wide association studies (GWAS), published simultaneously in 2009, reignited researchers' interest (3–5). We describe the structure and function of the MHC, discuss some of the challenges for genetic analysis of the region, and detail the findings from GWAS studies before considering approaches to interpreting the etiological role of the locus.

Structure and Function of the MHC

The human MHC, representing ~.1% of the genome, is a 4 million base pair (megabase [Mb]) region located on the short arm of chromosome 6 (6p21). The locus contains ~250 genes and encodes classical and transplantation HLA genes but also many other immune and nonimmune genes. An extended MHC of 7.6 Mb comprising more than 400 annotated genes and pseudogenes has been defined more recently (6). The MHC brings unique challenges. To respond to a large number of highly variable antigens, many genes have evolved rapidly to be extremely polymorphic, but related genes of similar function have also developed (polygeny). The MHC is complex, with extreme sequence diversity, substantial linkage disequilibrium (LD) and high gene density (7). An elaborate nomenclature system cataloguing allelic variation in each HLA region has arisen. Since 1968, the naming and quality control of HLA genes/alleles have been the responsibility of the World Health Organization Committee for Factors of the HLA system (http://www.ebi.ac.uk/ipd/imgt/hla/) and this now includes information on 9310 different alleles (April, 2013).

Major histocompatibility complex molecules were first investigated for their ability to identify compatibility (histocompatibility) for allogeneic tissue grafting but they have a much wider immunological role. Major histocompatibility complex genes have many other functions and their role in brain development is the subject of growing scrutiny (8,9). Three classes of MHC molecules have been identified, including a core of 21 highly polymorphic HLA genes. The MHC class I region is located at the telomeric end of the MHC and encodes the HLA class I molecules, HLA-A, HLA-B, and HLA-C. These are present on all nucleated cells and present endogenous antigens to CD8+ T cells. The class II region lies centromeric within the region and encodes HLA class II genes HLA-DRA, HLA-DRB1 (and its homologs HLA-DRB3, HLA-DRD4, HLA-DRD5), HLA-DQA1,HLA-DQB1, HLA-DPA1, and HLA-DPB1. Class II molecules are generally only expressed in cells involved in immune responses, where these molecules present antigens to CD4+ helper T cells. Located between these regions, the class III genes represent other non-HLA immune proteins, including components of the complement cascade, cytokines, heat shock proteins, transcription factors, other signaling molecules, transfer RNAs and olfactory receptors (10).

Using this framework, the nomenclature of the HLA region defines gene-wide haplotypes, where each haplotype has its own unique number. This evolved from simple two-digit numbers (e.g., HLA-DQA1*01) arising from the original antibody-based serotyping used to study isoforms of HLA molecules. Antibodies bind to extracellular elements of the HLA molecule, so differences based on serotyping represented variation in the exposed parts of the HLA protein. A four-digit coding was developed to capture nonsynonymous, unique gene products. More recently,

From the Department of Psychiatry and Neuropsychiatric Genetics Research Group, Institute of Molecular Medicine, Trinity College Dublin, Dublin, Ireland.

Address correspondence to Aiden Corvin, M.D., Ph.D., Trinity College Dublin, Department of Psychiatry and Neuropsychiatric Genetics Research Group, Institute of Molecular Medicine, Dublin 2, Ireland; E-mail: acorvin@tcd.ie.

Received May 27, 2013; revised and accepted Sep 18, 2013.

0006-3223/$36.00
http://dx.doi.org/10.1016/j.biopsych.2013.09.018

synonymous changes within exons were defined by a six-digit code and noncoding changes outside exons by an eight-digit code. Importantly, for interpreting genome-wide association studies, the nomenclature does not systematically capture information on all single nucleotide polymorphisms (SNPs), indels, or copy number changes at the locus.

Given the role of the MHC in immune function and its level of diversity, its involvement in disease susceptibility is unsurprising, and early disease association was confirmed for certain autoimmune disorders (e.g., type 1 diabetes, ankylosing spondylitis). In the GWAS era, hundreds of disease associations have been defined, including most autoimmune disorders, inflammatory disorders, infectious diseases, cancer, neurological disorders, and hypersensitivity responses to medications [e.g., abacavir (11), carbamazepine (12); reviewed in (13)]. Identifying these associations has been useful in guiding treatment but also in uncovering unexpected relationships between disease entities (14).

Genetic Studies of the MHC

Two sources of evidence prompted speculation about the role of the MHC locus in schizophrenia. First, there is evidence for involvement of infection in schizophrenia risk, mostly related to prenatal or early life exposures (15). Second, increased rates of autoimmune and inflammatory disorders are reported in schizophrenia (16). Since the initial 1974 study (2), hundreds of association studies testing all three classes of MHC molecules have been reported [reviewed in (17,18)]. With hindsight, many of the positive results reported may have represented type I errors in studies of small sample size, performing tests on many alleles, with variable genotyping quality and often with poorly matched control populations—problems that plagued most fields of complex disorder genetics in the pre-GWAS era. In their review, Wright *et al.* (18) suggested some evidence for protective alleles at DQB1*0602 and/or DRB1*04.

Genome-wide association studies represented a paradigm shift from piecemeal studies of individual loci to comprehensive genome-wide assessment of common risk variation. This was made possible by improved technology allowing assays of SNP markers numbering in the millions, directly or by imputation (19). Testing a million markers ratchets up the risk of false positives so a conservative threshold of $p < 5 \times 10^{-8}$ is taken as evidence of genome-wide association. Coupling technology with unparalleled collaborative efforts and large sample sizes, common risk variants have been confirmed for many diseases [e.g., diabetes, inflammatory bowel disease, psoriasis, cerebrovascular disease (20–23)] and traits [e.g., lipid levels, height (24,25)] (see the National Human Genome Research Institute GWAS catalogue at http://www.genome.gov/gwastudies).

Schizophrenia GWAS

Three GWAS studies, simultaneously published by *Nature* in 2009, provided more definitive evidence implicating the MHC in schizophrenia (3–5). Although other common schizophrenia risk variants of small effect (odds ratios <1.2) and rare structural variants of larger effect (odds ratios between 2 and 20) have also been identified (26), the MHC finding remains the most significant and consistent across subsequent schizophrenia GWAS.

In their discovery sample of 3322 cases and 3587 control subjects, the International Schizophrenia Consortium (ISC) (3) identified association with >450 SNPs spanning the MHC locus, including a genome-wide significant SNP rs3130297 ($p = 4.79 \times 10^{-8}$; Table 1) at 32.3 Mb in the MHC class I region 7 kilobase from the previously implicated *NOTCH4* gene. Combining data with the Molecular Genetics of Schizophrenia (2687 US cases and 2653 control subjects of Caucasian ancestry) (4) and European consortium on schizophrenia genetics (SGENE) (2005 cases and 12,837 control subjects of European ancestry) (5) consortia studies identified seven genome-wide significant SNP signals. The strongest was from SNP rs13194053 at 27.3 Mb, which is 5 Mb away from the original signal (Table 1). In their study, the SGENE consortium had access to additional samples, giving a total of 12,945 cases and 34,591 control subjects and yielding further support for five genome-wide significant SNPs (5). This identified four SNPs at 27.2-28.4 Mb and a fifth at 32.3 Mb, again close to *NOTCH4*.

These studies provided a rude introduction to the challenges of association mapping the MHC: strong signals arising across a large region, multiple signals but no clear example of independently associated SNPs in the same sample, and difficulty localizing signals to attribute them to underlying genes. These issues are addressed below. Critics argued that the variation in the degree of association across cohorts, and innate variability at the MHC locus across populations, pointed to cryptic substructure due to population stratification rather than true association as a likely explanation for these results (27). This concern has been comprehensively addressed by subsequent studies.

We reviewed the schizophrenia GWAS literature based on the National Human Genome Research Institute GWAS catalogue; PubMed using the search terms schizophrenia, psychosis, and genome-wide association studies; and from international conference abstracts. For most diseases, common risk alleles have a modest effect on risk, so individual studies, unless they include thousands of subjects, are underpowered. Schizophrenia is no exception, and we place particular emphasis on these larger studies (28). Further independent studies have been reasonably consistent in generated genome-wide significant evidence for risk SNPs at this locus in Irish (rs204999) (29), Swedish (rs886424) (30), Japanese (rs2071287) (31), and Chinese datasets (rs1635) (32) (Table 1). Other individual studies reported additional associations with specific risk SNPs, albeit not to the genome-wide corrected threshold but providing further replication evidence (33–35).

A meta-analysis of GWAS data, by the Schizophrenia workgroup of the Psychiatric GWAS (now Genomics) Consortium (PGC-SZ), including 17 separate studies ($n = 9394$ cases and 12,462 control subjects), identified 129 genome-wide significant SNPs mapping to the extended MHC (6p21.32-p22.1) (36). This analysis largely overlapped with the data reported in the original three studies. Taking 22 of these SNPs forward, 5 were genome-wide significant, following a combined analysis with a further 8422 cases and 21,397 control subjects. Each of these five SNPs has, in turn, been replicated in an independent UK sample of 2640 cases and 2878 control subjects from the CLOZUK cohort (37). The addition of these samples to the PGC-SZ analysis represents the largest reported association sample (20,476 cases and 36,737 control subjects) and the most significant signal comes from rs2021722 ($p = 1.05 \times 10^{-14}$). Strange *et al.* (29) reported an even stronger association at rs2523722 ($p = 1.47 \times 10^{-16}$) in a slightly smaller overlapping sample. These two SNPs are in high LD and map at 30.3 Mb in the MHC class I region (Figure S1 in Supplement 1), representing further evidence for replication in