INTRODUCTION

Multiple sclerosis (MS) is an autoimmune neurodegenerative disease of the central nervous system (CNS). This disease involves demyelination or the degradation of the myelin sheath, which insulates and protects axons to allow for efficient signaling of nerve impulses throughout the brain and spinal cord. Through stripping the myelin sheath from nerve cells, the neuron signal transmissions are either hindered or completely blocked. These processes are variable and unpredictable; thus, there is a wide spectrum of symptoms associated with this disease. Common symptoms include vision problems, numbness/tingling, problems with coordination or balance, and cognitive problems. MS is an etiologically complex disease where the genetic architecture is defined by genetic and likely clinical heterogeneity and is composed of the effects of single genes, polygenic inheritance, incomplete penetrance, and various environmental risk factors (1,2). Twin studies and other family-based studies have shown that there is a strong genetic component associated with MS (3-6).

Because of the autoimmune nature of MS, the Major Histocompatibility Complex (MHC) was an early candidate and was associated with MS in the 1970s (7,8). This association was later localized to the HLA-DRB1*1501-DQB*0602 haplotype. However, this haplotype only explains ~25-35% of the moderate genetic component attributed to MS (9-11). A number of genome-wide genetic linkage screens and numerous candidate gene studies have been completed testing for additional genetic effects outside the MHC region. However, until 2007, these various screens failed to consistently identify any new regions associated with MS. Interleukin-7 receptor alpha chain (IL7RA) was the first non-MHC locus to be identified. Following this discovery, twelve non-MHC genes have also been associated

The MHC signal is primarily attributed to the class II human leukocyte antigen (HLA) genes (more specifically, HLA-DRB1). However, the MHC region has several complicating characteristics including high gene content, extreme levels of polymorphism, and a dense pattern of linkage disequilibrium (LD). Because of these characteristics, it is difficult to discern the particular locale of the functional MHC association with MS (19). In addition, extensive conserved ancestral haplotypes in this region have been described; thus, creating difficulties in resolving if a single allele or the entire haplotype contributes to disease susceptibility. Despite these complications, data suggests that the MHC association with MS is much more complex than initially determined and that this region harbors additional susceptibility loci (19-21).

With the discovery of common variation, association studies took the place of linkage screens. Following the sequencing of the human genome and the identification of millions of single nucleotide polymorphisms (SNPs), genome-wide association studies (GWAS) became the norm. However, the advent of association studies did not produce quite the expected impact in investigations of complex diseases. This unfulfilled expectation is caused by variability in the parameters used in association studies, including effect size, the true disease allele frequency (DAF), the marker allele frequency (MAF), and LD between those alleles (22). With the dense pattern of LD in the MHC region, it is difficult to discern which regions may harbor independent (from HLA-DRB1*1501) effects and which regions hold residual effects (due to the LD) from HLA-DRB1*1501. It has been suggested that this...
region holds additional associated loci, but differentiating the independent effects has proved difficult.

A theoretical relationship exists between these parameters: for a given power, there must be an increase in the sample size to compensate for the reduced LD (between the marker and disease alleles) that is inversely proportional to \( r^2 \) (used as the measure of LD between the marker and disease alleles). Thus, despite how correlated two alleles are (i.e. \( D'=1 \)), if there is a big difference between MAF and DAF, then the \( r^2 \) will be smaller and the larger number of samples are needed to observe an association between the marker and the disease. Using this relationship and knowledge of the true disease allele, the odds ratio for the marker allele can be calculated using the other parameters (22).

Using a known disease allele and its observed effect size, it is possible to calculate odds ratios for markers in LD with this disease allele. This marker odds ratio (OR) calculation takes into account the MAF and the LD between the marker and the disease allele. The DRB1*1501 allele remains the strongest single effect for MS. Using the DRB1*1501 allele as the “disease allele” and the extreme LD in this region, we should observe a relationship between the marker allele odds ratios and the level of LD. However, if there are certain markers that have odds ratios much different than would be expected based on this residual effect from the major disease allele, this could indicate that these markers are associated with MS independently from the DRB1*1501 allele.

We hypothesize that independent secondary loci within the MHC region can be discerned by examining the loss of the effect size proportional to the loss in LD from the true disease allele. Thus, we examined the LD pattern across the MHC region. We identified
various markers across the MHC region and compared their observed ORs with the calculated ORs. The calculated OR is the measure of the residual effect from the association of HLA-DRB1*1501. When this measure differs from the observed OR, we hypothesize these are regions harboring independent susceptibility loci with MS.

METHODS

Samples

Samples for this analysis came from a large collaborative study where the same cases and controls were used for multiple projects. Through the International Multiple Sclerosis Genetics Consortium (IMSGC), we compiled 2,961 samples from the United States and United Kingdom (1,479 cases and 1,482 controls). The McDonald criteria were used to provide a positive diagnosis for MS (23). The controls were gender- and age-matched to the cases when possible. All four main types of MS were represented in our sample population and individuals had varying levels of disability. All individuals self-reported as non-Hispanic whites. A detailed description of these samples can be found elsewhere (24).

Genotyping

This collaborative study consisted of 48,767 SNPs for multiple projects genotyped on a single 60K Illumina BeadChip design. One of the projects on this BeadChip examined the MHC region specifically for further delineation of this complex region. Following a rigorous set of quality control processes, 2,343 single nucleotide polymorphisms (SNPs) were genotyped in the 28 Mb to 36 Mb region on chromosome 6 containing the MHC. Also included on the BeadChip was rs3135388, a SNP highly predictive of the HLA-DRB1*1501 allele. This SNP was used as the surrogate SNP for the large signal observed in the MHC (25).
For genotyping, cases and controls were randomized on each plate. The Illumina protocol entails genomic DNA amplification, fragmentation, hybridization to the BeadChip, extension of the bases on the bead, and finally imaging to read the chip. A detailed protocol is discussed elsewhere (26).

ANALYSIS

We calculated the allelic odds ratio (OR) for each SNP using two methods: observed OR and calculated OR. The observed OR is OR calculated using the odds of the each allele’s frequency in the cases vs. controls arbitrarily using the minor allele as the disease allele exposure. The calculated OR was generated using the theoretical relationship observed between LD and allele frequencies (MAF) and the effect size (equation 1). The proposed equation shows the residual OR of a particular marker based on the LD between this particular marker and the disease locus. Thus, this equation is displaying a residual effect as a ghost of the primary signal. LD (both $r^2$ and $D'$) were calculated using HaploView (27).

Eq 1: \[ OR_M = 1 + \frac{D \times (OR_T - 1)}{r} \times [s + (ps-D) \times (OR_T - 1)] \]

where \( OR_M \) is the marker allelic OR, \( OR_T \) is the trait or disease allelic OR (in our case rs3135388), \( D \) is the measure of LD between the marker and rs3135388, \( r \) is the allele frequency (minor allele) of the marker SNP in the controls, \( s \) is 1-\( r \) (the major allele of the marker SNP in the controls), and \( p \) is the disease allele frequency in the controls. This equation was derived and presented elsewhere (22).
RESULTS

The ORs for each SNP were separated into three categories: SNPs with no LD (based on $r^2$ values) with rs3135388 ($r^2=0$), with some LD ($r^2$ values between 0.001 and 0.2), and with high LD ($r^2 >0.2$) (Figures 1.1-1.3) to observe if SNPs with similar linkage to rs3135388 “behaved” similarly (i.e. had similar calculated and observed ORs). Because the calculated ORs can only be positive values, we used only absolute values of all our ORs for comparison. Based on the amount of LD between the SNPs and HLA, we observed the expected general trend for a positive correlation between the ORs (observed and calculated) of these SNPs. We used graphical representation of our data to visually observe outliers.

For those SNPs with no LD with rs3135388 (thus with a calculated log (OR)=0), we observed 3 outlier SNPs: rs9468892, rs3093550, and rs4647187 with observed ORs of 2.92 (log=0.46), 6.17 (log=0.79), and 8.26 (log=0.91) (Figure 1.1). We determined outliers as those SNPs that were visually separated from the clustering of SNPs in the groups we formed based on the amount of LD with our surrogate SNP (rs3135388).
For SNPs with some or high LD, we expect to see a strong correlation between the observed ORs and the calculated ORs. SNPs that do not cluster around the line y=x suggest independent associations with MS. Most of the SNPs with some LD \( (r^2 < 0.2) \) cluster around the line y=x indicating the expected correlation. There were 6 SNPs that did not cluster: rs2076535, rs2858880, rs1794265, rs3622952, rs3093556, and rs7769537 (Figure 1.2). For those SNPs with moderate LD \( (r^2 < 0.2) \), the log of the observed ORs and the log of the calculated ORs correlate very well (Figure 1.3).
Figure 1.2. Log of Observed OR (x-axis) vs. log of calculated OR (y-axis) for SNPs with some LD with rs3135388 (1,993 total SNPs).
We also looked at the absolute value of the difference between the calculated OR and the observed OR. Comparing this difference value with the distance from the HLA allele (rs3135388), one outlier was observed: rs7769537 with a difference of 2.89 (Figure 1.4). We observe a similar trend when comparing the r2 and D’ (which is defined as D/D_{max} when D>0 and D/D_{min} when D<0) values to the distance from HLA (Figure 1.5 and 1.6).
Figure 1.4. Difference between difference in calculated OR and observed OR vs distance from rs3135388.
Figure 1.5. Comparison of $r^2$ with the distance from the HLA allele (rs3135388).
DISCUSSION

Examining the correlation between the observed and calculated ORs, the SNPs in stronger LD with rs3135388 ($r^2>0.2$) had a stronger linear relationship. In other words, as the LD between the SNP and rs3135388 dissipated, the observed OR is consistent with the residual effect from the true disease allele. However, as the LD of the MHC region with rs3135388 disappears, this correlation reveals outliers.

Of those SNPs with no LD (Figure 1.1), 3 SNPs clearly did not follow the correlation pattern: rs9468892, rs3093550, and 4647187. According to the relationship with rs3135388, their calculated ORs were 1 (log=0). However, each had an observed OR over 2.0: 2.92
Of those SNPs with some LD (Figure 1.2), 6 SNPs did not cluster with the group: rs2076535, rs2858880, rs1794265, rs3622952, rs3093556, and rs7769537. All 6 of these SNPs had observed OR over 2.0: 2.17 (log=0.34), 2.33 (log=0.37), 2.50 (log=0.40), 2.50 (log=0.40), 2.74 (log=0.44), and 4.81 (log=0.68), respectively, which is larger than the calculated OR. These 6 SNPs had minimal LD with rs3135388 ($r^2=0.001-0.005$).

Because the majority of the SNPs in this region follow this correlation, one can reason that those SNPs that are visible outliers on the plots display independent effects from the main disease allele (in this case, rs3135388, which is a surrogate for HLA-DRB1*1501). This method of investigating the very convoluted MHC region has revealed 9 outliers representing all 3 MHC class regions. Closer investigation of these “outliers” reveals that observed odds ratios are over twice as large as the calculated odds ratio (Table 1.1).

All of these regions have potential biological relevance to MS, but none have been studied comprehensively with MS pathology. Two of the SNPs (rs3093550 and rs4647187) with no LD with the HLA-DRB1*1501 surrogate and one SNP with some LD (rs3093556) are in the lymphotoxin beta (LTB) gene, which is part of the tumor necrosis factor (TNF) superfamily. Inhibition of genes in the lymphotoxin pathway, including LTB, can be used in the therapy of various autoimmune disorders (28). One SNP with some LD with rs3135388 (rs3622952) was in the transporter 1, ATP-binding cassette sub-family B (TAP1) gene, which recently has been examined in metastatic lesions of head and neck squamous cell carcinoma and in chronic hepatitis B virus infected patients (29,30). This gene along with other MHC class I genes is vital for proper T cell recognition (31). Two of the SNPs with some LD with the HLA-DRB1*1501 surrogate (rs2858880 and rs1794265) are in HLA-DQA2. One of the SNPs
with no LD (rs9468892) is in the HLA complex group 27. Finally, the last 2 SNPs (rs2076535 and rs7769537) are located within open reading frames (C6orf10 and C6orf27, respectively) (Table 1.1).

Thus, using this relationship between allele frequencies and LD of alleles in dense LD genomic regions, we can identify which SNP is more likely to be the true disease loci and whether the two effects are independent or the result of residual LD in the region. The MHC region of chromosome 6 has been associated with MS since the mid 1970s; however, only one locus (HLA-DRB1*1501) has been identified and consistently replicated. Various other studies have suggested that additional loci outside of HLA-DR2 (i.e. HLA-C) are present in this region; however, differentiating these additional regions is not trivial (19-21). This method of comparing odds ratios can help better discern whether nearby significant loci are actually independent. We have identified six new genes or open reading frames that have suggestive association with MS.

For future studies, if one is unsure which loci is a true disease allele, then the calculation can be performed separately so both are considered the true disease marker. Using the relationship between the LD and the residual effect, one should be able to detect if there are other independent effects and where exactly they are positioned. This method could also be used when several markers display significance in a short spatial region to determine independence.
Table 1.1. Summary of Outlier SNPs

<table>
<thead>
<tr>
<th>SNP</th>
<th>Gene</th>
<th>MAF</th>
<th>$r^2$ (with rs3135388)</th>
<th>OR$_{\text{calculated}}$</th>
<th>log(OR$_{\text{calc}}$)</th>
<th>OR$_{\text{observed}}$</th>
<th>log(OR$_{\text{obs}}$)</th>
<th>OR difference (OR$<em>{\text{obs}}$-1)/(OR$</em>{\text{calc}}$-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs9468892</td>
<td>HLA complex group 27</td>
<td>0.004</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2.92</td>
<td>0.46</td>
<td>n/a</td>
</tr>
<tr>
<td>rs3093550</td>
<td>LTB</td>
<td>0.001</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>6.17</td>
<td>0.79</td>
<td>n/a</td>
</tr>
<tr>
<td>rs4647187</td>
<td>LTB</td>
<td>0.002</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>8.26</td>
<td>0.91</td>
<td>n/a</td>
</tr>
<tr>
<td>rs2076535</td>
<td>C6orf10</td>
<td>0.005</td>
<td>0.001</td>
<td>1.18</td>
<td>0.074</td>
<td>2.17</td>
<td>0.34</td>
<td>6.50</td>
</tr>
<tr>
<td>rs2858880</td>
<td>HLA-DQA2</td>
<td>0.005</td>
<td>0.001</td>
<td>1.18</td>
<td>0.074</td>
<td>2.33</td>
<td>0.37</td>
<td>7.39</td>
</tr>
<tr>
<td>rs3622952</td>
<td>TAP1</td>
<td>0.004</td>
<td>0.001</td>
<td>1.47</td>
<td>0.17</td>
<td>2.50</td>
<td>0.40</td>
<td>3.19</td>
</tr>
<tr>
<td>rs1794265</td>
<td>HLA-DQA2</td>
<td>0.016</td>
<td>0.005</td>
<td>1.23</td>
<td>0.089</td>
<td>2.50</td>
<td>0.40</td>
<td>6.52</td>
</tr>
<tr>
<td>rs3093556</td>
<td>LTB</td>
<td>0.002</td>
<td>0.001</td>
<td>1.74</td>
<td>0.24</td>
<td>2.74</td>
<td>0.44</td>
<td>2.35</td>
</tr>
<tr>
<td>rs7769537</td>
<td>C6orf27</td>
<td>0.003</td>
<td>0.001</td>
<td>1.92</td>
<td>0.28</td>
<td>4.81</td>
<td>0.68</td>
<td>4.14</td>
</tr>
</tbody>
</table>
REFERENCES


