Interrogating the complex role of chromosome 16p13.13 in multiple sclerosis susceptibility: independent genetic signals in the CIITA–CLEC16A–SOCS1 gene complex

Rebecca L. Zuvich1,∗, William S. Bush1, Jacob L. McCauley2, Ashley H. Beecham2, Philip L. De Jager3,4, the International Multiple Sclerosis Genetics Consortium, Adrian J. Ivinson5, Alastair Compston6, David A. Hafler4,7, Stephen L. Hauser8,9, Stephen J. Sawcer6, Margaret A. Pericak-Vance2, Lisa F. Barcellos10, Douglas P. Mortlock1 and Jonathan L. Haines1

1Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville, TN 37232-0700, USA, 2John P. Hussman Institute for Human Genomics, University of Miami, Miller School of Medicine, Miami, FL 33136, USA, 3Program in NeuroPsychiatric Genomics, Center for Neurologic Diseases, Department of Neurology, Brigham & Women's Hospital and Harvard Medical School, Boston, MA 02115, USA, 4Program in Medical & Population Genetics, Broad Institute of Harvard University and Massachusetts Institute of Technology, Cambridge, MA 02139, USA, 5Harvard NeuroDiscovery Center, Harvard Medical School, Boston, MA 02155, USA, 6Department of Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, BOX 165, Hills Road, Cambridge CB2 0QQ, UK, 7Department of Neurology, Yale School of Medicine, New Haven, CT 06520-8018, USA, 8Department of Neurology and 9Institute for Human Genetics, School of Medicine, University of California, San Francisco, CA 94143-0435, USA and 10Division of Epidemiology, School of Public Health, University of California, Berkeley, CA 94720-7356, USA

Received March 10, 2011; Revised May 17, 2011; Accepted May 31, 2011

Multiple sclerosis (MS) is a neurodegenerative, autoimmune disease of the central nervous system, and numerous studies have shown that MS has a strong genetic component. Independent studies to identify MS-associated genes have often indicated multiple signals in physically close genomic regions, although by their proximity it is not always clear if these data indicate redundant or truly independent genetic signals. Recently, three MS study samples were genotyped in parallel using an Illumina Custom BeadChip. These revealed multiple significantly associated single-nucleotide polymorphisms within a 600 kb stretch on chromosome 16p13. Here we present a detailed analysis of variants in this region that clarifies the independent nature of these signals. The linkage disequilibrium patterns in the region and logistic regression analysis of the associations suggest that this region likely harbors three independent MS disease loci. Further, we examined cis-expression QTLs, histone modifications and CCCTC-binding factor (CTCF) binding data in the region. We also tested for correlated expression of the genes from the region using whole-genome expression array data from lymphoblastoid cell lines. Three of the genes show expression correlations across loci. Furthermore, in the GM12878 lymphoblastoid cell line, these three genes are in a continuous region devoid of H3K27 methylation, suggesting an open chromatin configuration. This region likely only contributes minimal risk to MS; however, investigation of this region will undoubtedly provide insight into the functional mechanisms of these genes. These data highlight the importance of taking a closer look at the expression and function of chromosome 16p13 in the pathogenesis of MS.

∗To whom correspondence should be addressed at: 2215 Garland Avenue, Light Hall 519, Nashville, TN 37232. Tel: +1 6153226437; Fax: +1 6153221453; Email: rebecca.l.zuvich@vanderbilt.edu

© The Author 2011. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
INTRODUCTION

Multiple sclerosis (MS) is an autoimmune neurodegenerative disease of the central nervous system. MS has a well-established genetic component, but extended families with MS are uncommon and multiple genetic studies have yielded virtually no consistent results. This, coupled with the fact that MS has a complex clinical presentation, suggests a heterogeneous genetic etiology. Due to the autoimmune nature of the disease, there was early success in the 1970s with the discovery that variants in the major histocompatibility complex (MHC) were associated with MS, specifically the HLA-DRB1\(^*\)1501 allele. The identification of several non-MHC MS disease susceptibility loci has been achieved through recent genome-wide association studies (GWAS) (1–3). To date, 16 non-MHC genes have emerged as MS disease susceptibility loci, including \(\text{IL7RA}\), \(\text{IL2RA}\), \(\text{CLEC16A}\), \(\text{CD226}\), \(\text{CD58}\), \(\text{TYK2}\), \(\text{CD6}\), \(\text{IRF8}\), \(\text{TNFRSF1A}\), \(\text{MMEL1}\), \(\text{KIF21B}\), \(\text{TMEM39A}\), \(\text{IL7}\), \(\text{SOCS1}\), \(\text{CIITA}\) and \(\text{CD40}\) (4–15).

An important remaining question is how many variants explain the genetic component of MS. Although this may prove to be an impossible question to answer precisely, recent studies suggest that there are several hundred common alleles across the genome with very modest effect sizes that collectively influence MS risk (16) and many additional rare alleles may further explain the variable nature of this complex disease (17–19). Additionally, environmental variables and interactions (both gene–gene and gene–environment) also likely contribute to the ‘missing heritability’ of most complex diseases (20). The discovery of a genomic region on chromosome 16 where multiple independently associated single-nucleotide polymorphisms (SNPs) are clustered provoked us to further investigate what impact this region may have on MS pathogenesis. A \(~600\) kb stretch of chromosome 16 contains several genes that were the focus of three recent publications: a candidate gene study of \(\text{CIITA}\) (8), an extensive follow-up of the GWAS performed by the International Multiple Sclerosis Genetics Consortium (IMSGC) that included SNPs from \(\text{CLEC16A}, \text{PRM1}, \text{PRM2} \) and \(\text{PRM3}\) (14) and a candidate pathway approach examining genes involved in the \(\text{IL7RA}-\text{IL7}\) pathway including \(\text{SOCS1}\) (15). While this chromosome 16 region directly contributes a small amount to risk, investigation of this region will undoubtedly provide insight into the functional mechanisms of these genes in general and could help clarify the underlying pathology for MS. Here, we collectively examine the region spanning these loci. Through our interrogation of this region, we detect several areas of ‘independent genetic signal’ [that is, not sharing linkage disequilibrium (LD)], we find some of these risk variants are differentially sensitive to HLA-DRB1*1501 genotype (8), and that these variants could potentially regulate the transcription of multiple genes in the region due to shared \textit{cis}-regulatory mechanisms.

RESULTS

The Armitage trend tests on the overall and HLA stratified data sets for all SNPs in this region are presented in Supplementary Material, Table S1. Upon inspection of the LD structure of this region, all the SNPs in \(\text{CIITA}\) are in the same LD block (based on \(D^\prime\) values; average \(D^\prime\) value = 0.48; average \(r^2\) value = 0.07). Likewise, the SNPs in \(\text{CLEC16A}\) are in a single LD block (average \(D^\prime\) value = 0.91; average \(r^2\) value = 0.46); the remaining SNPs are in one large LD block (average \(D^\prime\) value = 0.60; average \(r^2\) value = 0.14; encompassing \(\text{CLEC16A–SOCS1}\) intergenic region, \(\text{SOCS1}, \text{TNP2}, \text{PRM1–3}, \text{C16orf75}, \text{LOC388210} \) and \(\text{LOC400499}\)). Also, this LD pattern was consistent when stratifying the samples by \(\text{HLA}\) (\(\text{HLA}^+\) versus \(\text{HLA}^-\)) and by affection status (cases versus controls) (data not shown). These data suggest that each of these regions is independent (i.e. little to no residual LD across blocks with an average \(D^\prime\) value of 0.19 and average \(r^2\) value of 0.01). Given the low \(r^2\) values between SNPs of the three regions, it is highly unlikely that these signals are redundant due to LD (Figs 1 and 2).
The most significant \( P \)-values from the Armitage trend test in each region include a SNP from \( CIITA \) (rs4774), a SNP from \( CLEC16A \) [rs7184083, which had the lowest \( P \)-value in this gene in our data set and is in LD (\( r^2 = 0.31; D' = 1.00 \)) with the previously identified variant, rs12708716], and two SNPs from the \( SOCS1 \) region (one in \( CLEC16A–SOCS1 \) intergenic region, rs1640923; and one in \( PRM1 \), rs12922733) (Table 1). These specific SNPs are highlighted because they are the most significant SNP from each of the three regions, and each has since been replicated in independent studies (12–14). We performed logistic regression analysis for each pair-wise combination of these four SNPs to examine possible interactions. In those SNP pairs with significant interactions, we then modeled each SNP adjusting for the interaction term using logistic regression (Table 2).

In the overall data set, the SNP in \( CLEC16A \) (rs7184083) had the most significant main effect of the four SNPs (Table 1). There are also significant interactions between SNPs in \( CIITA \) and \( CLEC16A \) and between SNPs in \( CIITA \) and \( PRM1 \) (Table 2). However, with even a modest correction for multiple testing, the significance of these interactions disappears. All four SNPs had significant \( P \)-values for their main effects after accounting for the interaction effect (data not shown). There was also a significant haplotype effect observed between the major alleles of both rs1640923 and rs12922733 in the third region (data not shown).

From available whole-genome expression data derived from Hapmap CEU lymphoblastoid cell lines, we examined the potential correlations in expression levels between 11 unique transcripts. These transcripts cover nine genes: \( CIITA \), \( DEXI \), \( CLEC16A \) (two alternatively spliced transcripts), \( SOCS1 \), \( PRM1 \), \( PRM2 \), \( PRM3 \), \( C16orf75 \) (two alternatively spliced transcripts) and \( Q6ZTK2_HUMAN \) (encoding a putative uncharacterized protein, \( LOC400499 \)). Highly significant correlations were seen between expression probes related to the adjacent genes \( DEXI \), \( CLEC16A \) and \( SOCS1 \) (Table 3).

To corroborate these results, we examined histone and CCCTC-binding factor (CTCF) ChIP data from the UCSC.
From the region, and these three SNPs in cell line K562. Therefore, an ‘active’ domain spans SOCS1. Interestingly, the ‘active’ domain also extends to include in all cells except the lymphoblastoid cell line GM12878. CLEC16A and SOCS1 in GM12878 cells and the myelogenous leukemia cell line K562. Therefore, an ‘active’ domain spans DEXI, CLEC16A and SOCS1 in at least some lymphoblastoid/myeloid cell lines. CTCF is strongly localized to the boundary of this domain and the SOCS1 gene itself, as well as several other sites within and around the constitutively ‘active’ domain. The PRM gene cluster is ‘inactive’ in all cell lines within ChIP data, consistent with the known tests-specific functions of these genes in spermatogenesis, while C16orf75 is globally ‘active’.

We also annotated significantly associated SNPs for expression QTLs (eQTLs) based on an analysis by Veyrieras et al. (21). Of all 39 significant Armitage tests ($P < 0.05$) in the region, 21 of them are eQTLs for C16orf75, most with highly significant associations to expression changes in the gene/reading frame ($P = 0.01$ to $1.5 \times 10^{-11}$). The remaining 18 significant Armitage test SNPs are eQTLs for other genes in the region (CLEC16A, SOCS1, PRM1 and PRM2) (Fig. 3). Interestingly, three SNPs are eQTLs for lipopolysaccharide-induced TNF factor (LITAF), a gene 200 kb downstream from the region, and these three SNPs are more significant in the HLA− strata.

**DISCUSSION**

One hundred and forty-nine SNPs were genotyped and analyzed across the 16p13 region for MS susceptibility. The CEU LD data for this region of chromosome 16 suggests that there are three distinct LD blocks (Fig. 3). After identifying the previously replicated SNP with the most significant association to MS (strongest $P$-value) from each of the three sub-regions, we further investigated the region for interaction effects. All of the nominally significant pairwise interaction models (Table 2) had significant main effects and the significance of the interactions disappears with correction for multiple testing. In addition, we also observed significant correlation of expression probes between DEXI, CLEC16A and SOCS1 (Table 3), which could indicate a common regulatory element. Although the HLA locus contributes more to MS risk than any other locus, our work again confirms that the chromosome 16p13 region independently contributes to MS risk. While the change in risk conferred by this region is small compared to that of HLA, further study of the 16p13 genes is likely to give insight to the less well understood, non-HLA-mediated cellular events that augment MS pathology.

The patterns of LD and the tests of independence in the logistic regression model (by adjusting for an interaction term) both suggest that each of these SNPs (and hence their tagged genes) has an independent effect(s) on MS susceptibility. Furthermore, these data suggest that there are interactive effects within this 600 kb region that also contribute to MS susceptibility.

We were surprised to discover that several SNPs were eQTLs for PRM1 (and one for PRM2) in several different non-testes cell lines, despite the likely testes-specific expression and function of these in vivo and obvious ‘repressive’ chromatin marks on the PRM genes in cell lines. We speculate that the region between CLEC16A and SOCS1 contains a regulatory domain that influences expression of multiple genes in this region and that genetic variants in this region alter MS risk through expression of more than one gene. The long-range nature of this cis-regulation might influence low levels of ‘leaky’ PRM1 transcription in cultured cell lines, explaining the eQTL effect on PRM1. Alternatively, some of the eQTL signals could reflect indirect (trans) effects on gene expression. Other SNPs in this region are shown to be associated with other autoimmune disorders (22,23).

Because of the differential associations with HLA strata, we have identified four sub-regions of independent association in this genomic region. The Subregion I variant(s) may be functioning through CITA exclusively, as this area has little LD and no evidence for eQTL effects (Figs 3 and 4). Likewise, Subregion II is contained with the CLEC16A transcription unit and in an LD block associated with this gene. In contrast, Subregion IIIa shares little LD with the structural genes in the region, but contains SNPs with eQTL effects on several genes in the region. The role of Subregion IIIb is unclear, although it is doubtful that it acts via altering PRM gene cluster expression as expression is likely to be tightly restricted to testes in vivo. More likely, Subregion IIIb variants affect cis-regulation of or are simply in LD with one or more SNPs that act through SOCS1 or C16orf75. Therefore, further searches for undiscovered variants in these genes may be warranted.

Because this is a small chromosomal region, it is reasonable to hypothesize that some regulatory mechanisms for separate
genes have co-evolved. Sets of genes that are highly conserved across species could support a model for some degree of common regulatory control and thus, expression of several or all genes in this region.

In addition to comparing expression, searching for known regulatory regions could also be influential in the search for the mechanism of disease pathology. If any other SNPs with significant P-values are in LD with a known regulatory region, it would suggest possible co-regulation of these genes. These effects could be located in regulatory elements that affect the transcription of the whole region, similar to the locus control regions that affect globin gene cluster expression (24).

Figure 3. Association results with gene map, cis-eQTLs and epigenetic data. (A) Manhattan plot of association data (negative log P-values) relative to gene map, (B) physical location of genes, (C) cis-eQTLs for SNPs with significant P-values, arranged by target gene. (D) Phased HapMap CEU LD data. (E) Location of CTCF binding and histone3 lysine 27 methylation ENCODE peaks for four selected cell lines (GM15878, K562, NHEK, HMEC, which are immortalized lymphoblastoid, leukemic, keratinocytic, mammary epithelial) (March 2006/NCBI 36/ hg18 from http://genome.ucsc.edu/) (33–35). The red box highlights an interesting region that has significant P-values in the HLA—stratum, which overlaps with a cis-eQTL and CTCF binding site.

Figure 4. Diagram of genetic signals and genes in the chromosome 16 MS complex. The four regions that harbor distinct genetic signals (black bars) are labeled with roman numerals. DEXI, CLEC16A and SOCS1 are colored green to reflect their correlated expression in cell lines. Region III may harbor eQTLs that act in cis to affect transcription of multiple genes in the region (arrows), some of which are implicated functionally in MS pathology (red arrows). Alternatively, the eQTL effects may not be cis-effects but could be indirect. Region IV has a dashed line to indicate LD with C16orf7.
PRM1 and PRM2, the genes with the most significant SNPs in the LD block containing SOCS1, do not appear to have a biological function relevant to known MS etiology (25). However, these SNPs, or SNPs in LD with them, could be involved in regulating transcription of SOCS1, which is a more logical candidate gene for MS pathogenesis. This is especially interesting given the previously shown role of SOCS1 in regulating inflammation. SOCS1 deficiency in T cells leads to hyperactivity in Th1 cells and a reduction in Th17 response, which has been linked to the pathogenesis of autoimmune disease (26). In addition, SOCS1 deficiency causes the breakdown of self MHC-T cell receptor recognition during thymic differentiation. This process leads to a release of autoreactive T cells (27,28). This mechanism could explain the breakdown in the normal control of T cell differentiation leading to T-cell hyper-reactivity, which may ultimately contribute to the attack on the myelin sheath.

Some differences in gene expression across this region argue against co-regulation. For example, CIITA expression in lymphoblastoid cells does not correlate with that of other genes in the region. On the other hand, the expression-correlated genes DEXI, CLEC16A and SOCS1 lie in a contiguous ‘active’ region of H3K27 methylation, and the eQTL data may support cis-effects that act across gene boundaries (e.g. LITAF) (Supplementary Material, Table S1).

The confluence of significant results on this arm of chromosome 16 has raised important questions. We have demonstrated that there are multiple independent signals in this genomic region. In various cell lines that are publicly available, these genes are not expressed together despite the conserved syneny of the 600 kb region. If regulatory elements located here (e.g. the SOCS1 region) overlap with MS-associated SNPs, they might influence expression of the nearby more biologically plausible gene, SOCS1. As there is no direct evidence for gene co-regulation, more studies are needed to confirm a model for SNPs that regulate multiple genes in cis. Further studies of expression and genotype correlations in additional lymphoblastoid cell lines and/or T cells from MS patients and controls should help refine these possibilities. These data suggest distinct functional mechanisms for these independent genetic effects in the etiology of MS.

MATERIALS AND METHODS
Genotyping techniques
SNPs of interest in the region were typed using the Illumina iSelect Custom BeadChip platform. A total of 48,767 SNPs were successfully arrayed on the custom beadchip and subsequently genotyped as part of a multi-center collaborative research project through the IMMSGC. A total of 149 SNPs in the region of chromosome 16 (bp: 10,870,067–11,462,368 for Build 36 Version 127) were analyzed for this work using a previously described final data set (14,15). SNPs in the chromosome 16 region were chosen for various projects: candidate gene study (CIITA), candidate pathway study (SOCS1 region) and as a replication study for top GWAS hits (CLEC16A) (12–14). All available SNPs (based on Illumina’s success algorithm) were selected for genotyping in the regions of interest.

Quality control
The data set consisted of 2961 individuals (1479 cases and 1482 controls) collected from the USA and the UK. The McDonald criteria were used to provide a positive MS diagnosis (29). The controls were collected from the same US sites as the cases and from the 1958 Birth Cohort in the UK. Controls were selected to match the ratio of female to male cases. Samples with clinically isolated syndrome (CIS) were excluded from the UK subset. Only 7% of the US subset had CIS at the time of their examination; however, these individuals almost always develop MS. The samples were all non-Hispanic individuals of European ancestry. The table of demographics can be found elsewhere (Table 4 of IMMSGC 2009) (14).

We tested for SNP and sample genotyping efficiency, allele frequencies, gender errors and Hardy–Weinberg equilibrium (HWE) using the Whole-genome Association Study Pipeline (WASP) (30). We removed individuals with low genotyping efficiency (<95%), with gender discrepancies (based on X-chromosome heterozygosity scores), with apparent relatedness (using Identity by State—IBS statistics) and individuals who had <0.90 probability of being from European descent. SNPs were eliminated with low genotyping efficiency (<95%), with no minor allele (monomorphic) and with deviation from HWE in the overall population (<0.0001). This process was repeated to ensure that the data were thoroughly cleaned. After stringent quality control criteria, we retained 2722 individuals (1343 cases and 1379 controls; 1735 females and 987 males) and 46,874 SNPs.

Analysis
We examined 149 SNPs using Armitage trend tests. From our previous study (15), we stratified SNPs from the genes in the IL7RA-IL7 pathway by the HLA-DRB1*1501 risk allele using the surrogate SNP rs3135388: by the presence of the risk allele (genotypes AA and AG) HLA+ or HLA− stratum) and the absence of the risk allele (genotype GG) (HLA− stratum) (11). Given that our previous study demonstrated SNPs within SOCS1 were differentially significant in either the HLA+ or HLA− strata, we chose to perform this stratified analysis on all SNPs within this region of chromosome 16 (including those within CLEC16A and CIITA).

To determine whether the signals (SNPs with the lowest P-value based on the Armitage trend test from each of the previous studies) were independent, we examined the LD patterns among these 149 SNPs. Using Haploview, we identified LD blocks and examined r2 and D′ values within this chromosomal region (31). To test for possible interactions, multi-factor dimensionality reduction was performed for the most significant SNPs from each gene region (32). In addition, we built a logistic regression model to examine interactions among all possible pair-wise combinations for the most significant SNPs (based on P-values) from each gene region. If there was a significant interaction between two SNPs using logistic regression, then we tested each SNP for an association with MS after adjusting for an interaction term.
Annotation

We examined the synteny of this region with the hypothesis that evolutionary pressures may be driving the genomic configuration of the multiple MS loci in this region. We used the UCSC Genome Browser (http://genome.ucsc.edu/) to determine when these genes converged on the same chromosome in the evolutionary scale. In addition, we examined UCSC Genome Browser ChiP data across multiple cell lines, including GM12878 (a CEPH lymphoblast cell line), HUVSMC, (umbilical vein endothelial), K562 (myelogenous leukemia) and NHEK (normal human epidermal keratinocytes) for histone marks that indicate active and inactive chromatin. We then used the Ensembl Genome Browser (http://www.ensembl.org/index.html), which has information on known regulatory regions, to determine whether any of the significant SNPs lie within known regions of histone marks for either active or inactive chromatin. Whole-genome gene expression data for lymphoblastoid cell lines (PMID: 17289997) were used for eQTL analysis and to examine correlated gene expression. These data were collected on CEU Hapmap samples using an Illumina WG-6v1 whole-genome expression array.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG online.

ACKNOWLEDGEMENTS

We acknowledge the work done by the Biorepository and the Center for Genome Technology within the John P. Hussman Institute for Human Genomics (University of Miami). We thank the Computational Genomics Core within the Center for Human Genetics Research (Vanderbilt University); and most importantly, we thank our study participants and families.

Conflict of Interest statement. There have been no involvement that might raise the question of bias in the work reported or in the conclusions, implications, or opinions stated.

FUNDING

The International MS Genetics Consortium is supported by grants (NIH R01NS32830), societies, foundations and a number of individual donors. This work was supported by the Medical Research Council (G0700061), the National Institute of Health (R01NS049477) and the Cambridge NIHR Biomedical Research Center. P.L.D.J. is a Harry Weaver Neuroscience Scholar Award of the National MS Society (NMSS); he is also a William C. Fowler Scholar in MS Research. D.A.H. is a Jacob Javits Scholar of the NIH. We acknowledge use of genotype data from the British 1958 Birth Cohort DNA collection, funded by the Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02.

REFERENCES


