Multiple sclerosis (MS) is a common autoimmune disease that targets myelin in the central nervous system (CNS). Multiple genome-wide association studies (GWAS) over the past 10 years have uncovered more than 200 loci that independently contribute to disease pathogenesis. As with many other complex diseases, risk of developing MS is driven by multiple common variants whose biological effects are not immediately clear. Here, we present a historical perspective on the progress made in MS genetics and discuss current work geared towards creating a more complete model that accurately represents the genetic landscape of MS susceptibility. Such a model necessarily includes a better understanding of the individual contributions of each common variant to the cellular phenotypes, and interactions with other genes and with the environment. Future genetic studies in MS will likely focus on the role of rare variants and endophenotypes.

MS Is a Genetic Disease
MS is a chronic disease of the CNS that affects more than two million individuals worldwide [1,2]. It is typically diagnosed between the third and forth decade of life, occurs more frequently in women than men, and is considered the most common cause of nontraumatic neurological disability in young adults. Although the trigger(s) of MS remains unknown, its pathogenesis is best explained by a multifactorial model that incorporates interactions between genetic, epigenetic, and infectious, nutritional, climatic, or other environmental influences, including Epstein–Barr virus (EBV) infection, sunlight exposure, and smoking [3–7]. This array of factors results in the loss of immune homeostasis and self-tolerance (see Glossary) manifested in brain and spinal cord infiltration by activated peripheral mononuclear cells, and the development of unregulated pathologic inflammatory responses against structural components of the CNS. Myelin loss, gliosis, and the resulting axonal pathology culminate in progressive, often severe, neurological dysfunction. Emerging data show that the neuroinflammation and neurodegeneration that occur in MS are overlapping and have a complex dependence [8–11] but the autoimmune model of pathogenesis has set early the tone for immunotherapy as the primary clinical management strategy, first by global immune suppression using aggressive anti-inflammatory drugs, and, more recently, by selectively targeting specific elements of the immune response [12–15]. The recognition that axonal damage unresponsive to immunotherapy is the main driver of disability in MS highlights the need to emphasize the genetics of gray and white matter cell death and repair as pressing research frontiers in this disease.

Heritable contributions to MS risk are undisputable. Pivotal epidemiological data consistent with a prominent role of genetic factors in MS include the high disease prevalence in distinct ancestral groups (particularly those of northern European origin) compared with others...
from the USA and UK using an unorthodox family-based (patients with MS and both parents) completed by the International Multiple Sclerosis Genetics Consortium (IMSGC) on subjects found in the sample size, and allele frequency. Initial power calculations for MS estimated that, to detect the power of genetic association studies mostly depends on three variables: effect size, heritability were overestimated.

The recognition during the early 1970s of the influence of specific human leukocyte antigen (HLA) variants within the major histocompatibility complex (MHC) gene complex (chromosome 6p21) in MS susceptibility represented the first empirical demonstration linking disease risk with common genetic variation, but it was the polygenic model of MS heritability that provided the theoretical justification for assembling multicenter large DNA data sets to pursue GWAS (Box 1).

Ten Years of GWAS in MS

The first concerted efforts to identify susceptibility genome-wide were done through linkage analysis in multicase families. These studies failed to identify any consistent statistically significant signal [36–38]. Yet, a suggestive logarithm of the odds ratio (OR) (LOD) score found in the HLA region was later confirmed in larger studies, providing a statistical road map to take full advantage of the emerging microarray technology [39,40]. The first GWAS in MS was completed by the International Multiple Sclerosis Genetics Consortium (IMSGC) on subjects from the USA and UK using an unorthodox family-based (patients with MS and both parents) study design for discovery coupled with case-control replication [41]. Several GWAS and large-scale targeted studies followed in the ensuing 10 years [42–48], culminating in the collation of a genetic data set representing 47 351 MS subjects and 68 284 controls that showed unequivocal statistical evidence for the association of 200 autosomal susceptibility variants outside the MHC, one chromosome X variant, and 32 independent associations within the extended MHC [49]. Altogether, different measures of the heritability explained by all associations are still in the 20–30% range, of which a substantial proportion can be assigned to the MHC region. While the explanation for this missing heritability may be rooted in gene×gene and gene×environment (G×E) interactions and epigenetic factors, it is also conceivable that the early measures of heritability were overestimated.

The power of genetic association studies mostly depends on three variables: effect size, sample size, and allele frequency. Initial power calculations for MS estimated that, to detect an association with OR >1.2 investigating common variants (those with minor allele frequency
Self-tolerance: immunological tolerance is the failure to mount an immune response to an antigen. Self-tolerance refers to the (normal) failure to respond to one’s own antigens, thus avoiding autoimmunity.

Splicing acceptor site: in the splicing of RNA, the site at the 3' end of an intron.

>5%), 10 000 samples would be required to ensure that a $P$ value of $10^{-7}$ is a hundred times more likely to be true than false (by contrast, the posterior odds of the same $P$ value with only 2000 patients would be in favor of a false discovery) [35]. While in later years, a $P$ value threshold of $5 \times 10^{-8}$ for GWAS emerged as a consensus [50], a systematic analysis of published studies suggested a further relaxation of this $P$ value would result in the discovery of additional genuine associations [51].

When all MS associations are analyzed by year of discovery (a proxy for study size), they show an ever-decreasing median odds ratio (OR; a measure of effect size) (Figure 1). For example, the association with HLA-DRB1*15:01 has an OR $\approx 3.5$ and was first reported in 1972 with just a few hundred cases, while the first discovered non-HLA loci (in 2007) have OR $\approx 1.15$ and nearly 1000 cases were needed. The most recent studies (2011, 2013, and 2017) analyzed tens of thousands of cases (10 000, 15 000, and 40 000, respectively), which enabled the reporting of statistically replicated signals down to OR $= 1.05$. This asymptotic trend toward OR $\approx 1$ also means that the cost:benefit incentive for performing additional GWAS searching for additional common susceptibility signals in Europeans may no longer be justified or realistic. Associations in that low effect size-range mean that those variants are almost as frequent in cases as they are in controls, thus raising the argument of whether they are biologically relevant. Quantitative risk scores have been developed to assess the significance of the ‘genetic load’ of individual

---

**Figure 1.** Decreasing Odds Ratio (OR) as Study n Increases. Box plots represent the median and quartiles of the OR of studies performed at the indicated years. With exception of 2010, each new study revealed associations with smaller OR.
patients and whether all associations count equally towards risk [52,53]. However, at least in these early versions, these algorithms are not able to accurately predict risk of individual subjects and, thus, are not yet an option to be used in the clinic for precision medicine purposes.

However, this does not mean the GWAS era is over in MS. Due to inherently different linkage disequilibrium (LD) structures, studies of other populations (African-American/European, Hispanics, or Asians) is a meritorious proposal for fine-mapping and identification of population-specific effects, notwithstanding the difficulties in assembling the required large DNA sample sets [54,55]. Likewise, the study of disease heterogeneity (clinical subtypes or endo-phenotypes among cases) by GWAS is a valid approach with high potential for discoveries of disease modifiers, particularly with the large cohorts that have been already assembled. Here, accurate and quantitative description of the phenotypes, beyond whether a subject has the disease or not, is crucial to success. Finally, the search for rare(r) variants by exome sequencing or similar approaches is an active field of investigation.

Nearly all the identified associations map to noncoding regions of the genome, either in intragenic or intergenic regions. Naturally, those mapping to intragenic regions were explored first (IL7RA, TNFSF10, etc.) as their relationship to function is thought to relate to splicing, mRNA stability, or promoter activity of the gene in question. However, most associations lie in genomic regions distant from any known gene. In those cases, the most likely explanation is that they modify regulatory sites and indirectly influence gene activity. Examples of regulatory regions include open chromatin, histone modifications, enhancers, and repressors. Efforts to integrate statistical evidence of association with their potentially regulatory effect are underway.

In the most recent MS GWAS, an analysis of open chromatin regions was performed using DNase I hypersensitivity data collated from 56 tissues from the NIH Roadmap Epigenomics Project (REPi) [56]. For each tissue, the proportion of SNPs in the list that were located on DHS regions to the proportion of all SNPs located on DHS regions active in that tissue were compared. Notably, statistically significant enrichment was observed only in nine types of cell from the immune system, including CD3, CD4, CD8 (T cells), CD14 (monocytes), and CD19 (B cells) [49]. This finding highlights the immune nature of genetic MS susceptibility.

Sharing with Other Autoimmune Diseases

One salient characteristic of the MS susceptibility map is the large proportion of the associations (either the exact same variant, or within the same gene or locus) that are shared with other autoimmune diseases, such as type 1 diabetes mellitus (T1DM), rheumatoid arthritis (RA),

Box 1. The MHC Is the Most Robust MS Genetic Susceptibility Locus

GWAS have confirmed that the main MS susceptibility association signal genome-wide maps to the HLA-DRB1 locus in the MHC. This region is located on the short arm of chromosome 6 at p21.3, spans almost 4000 kb of DNA, and contains approximately 165 closely linked genes, some 50% of which have pivotal roles in the immune system [82]. These include the HLA genes, which have been extensively studied as the principal determinants of allogeneic transplantation outcomes [83]. The polymorphism of many of the HLA genes is extraordinary, with more than 14000 alleles identified to date. More than 100 infectious, autoimmune, and inflammatory disease phenotypes, including MS, as well as drug reactions and cancers, are associated with HLA gene variations [84,85]. The association of the HLA locus with MS risk was first described several decades ago [86] and has since been observed across all populations studied [87]. HLA-DRB1 allelic heterogeneity vis-a-vis MS risk and copy number effects have also been described [88]. Individuals who are HLA-DRB1*15:01 homozygotes carry a high-risk genotype with ORs exceeding 7.0, compared with a range between 3.5 and 5.0 for heterozygotes HLA-DRB1*15:01/X. In addition to risk, HLA-DRB1*15:01 has been associated with phenotypic markers of disease severity [89]. A decade-long effort with genomic screens provided statistical evidence for over 30 independent allelic and genetic associations within the extended MHC region. Hierarchical allelic lineages and epistatic effects affecting risk are described but poorly understood [90,91]. It is sobering to recognize that, despite the fact that examination of HLA variation at this level has been performed in MS cohorts comprising many thousands of individuals, a complete understanding of the role of the MHC in disease pathogenesis remains elusive.
systemic lupus erythematosus (SLE), and Crohn disease (CD), among others. Interestingly, some of these shared associations confer risk for one disease but protection for another. For example, allelic heterogeneity was reported in the IL2RA locus, where rs11594656-A was shown to be associated with susceptibility to one disease (MS) and protection against another (T1DM). Furthermore, rs41295061-A only conferred susceptibility to T1DM but not MS. Finally, rs2104286-G was only associated with T1DM. While somewhat surprising at first, genetic and allelic heterogeneity across autoimmune diseases is now recognized to be a distinctive feature of this group of disorders.

Figure 2 shows associated genes for most autoimmune diseases with GWAS data (obtained from the GWAS Catalog). It clearly shows a core of shared loci (blue triangles) and several disease-specific ones represented by different colors (MS in yellow). This remarkable sharing is the highest among the complex diseases, including cancers, neurological, and metabolic (data not shown). This particular architecture confirms that MS is at its core an autoimmune disease and suggests that susceptibility to autoimmunity is an inherited risk that later in life may be compounded and shaped by additional genetic (and epigenetic) determinants and environmental exposures to ultimately define the predominant effector mechanism (cellular versus humoral) and target organ of the autodestructive process.

From Mapping to Function

One of the most difficult steps in any GWAS is to infer the biologically relevant consequences of each statistical association. Given that most associations map to noncoding regions, in the absence of obvious functional signals, earlier studies simply reported as candidate causative the closest gene(s) to the most significant independent effect. However, with continuous improvement in our understanding of the molecular machinery of gene regulation, it has now become clear that this crude approach needs further refinement. One way to confirm the functional effect of genetic associations is to conduct in vitro or in vivo experiments in which the presence of the variant is studied in the context of expression of the putative regulated gene target(s). The first such demonstration in MS was the discovery that rs6897932, a polymorphism located in chromosome 5p13.2 and previously associated with disease susceptibility (OR = 1.18) [41], influenced the function of the gene encoding interleukin 7 receptor (IL7RA).

Specifically, inheritance of the risk allele (C) disrupts a splicing acceptor site and results in transcriptional skipping of exon 6 of the gene, thus altering the relative amounts of soluble and membrane-bound isoforms of the gene [58]. More recently, several single nucleotide polymorphisms (SNPs) within the gene encoding RNA helicase DEAD box polypeptide 39B (DDX39B) were shown to be associated with MS risk in a meta-analysis [59]. In particular, rs2523506-A, located within the 5′-untranslated region (UTR) of this gene, was shown to reduce mRNA translation of DDX39B, which is in turn a potent activator of IL7R exon 6 transcription. This example provides functional evidence that two associated loci can work together to confer susceptibility to MS. Given the large number of independent associations, and the networked molecular pathways that underlie cellular behavior, this example may only be the tip of the iceberg in unraveling the complex genetic architecture of MS.

Another well-characterized example is the intronic SNP rs1800693 in the gene encoding tumor necrosis factor receptor super family 1A (TNFRSF1A). In this case, the risk allele also results in skipping of exon 6 and the production of a novel soluble form of the TNF receptor. This soluble protein is able to inhibit TNF signaling inside the cells, mirroring somehow the exacerbating effects of TNF-blocking drugs on MS course [60].

One of the most complex regions in the MS susceptibility map is 1p22.1, which has been the subject of several reports with seemingly contradictory results; that is, different genes in the locus were linked to the association signal [61–64]. In this locus, GWAS meta-analysis identified
multiple statistically independent genome-wide effects, three of which were found under the same association peak. In 2011, the disease-associated SNP rs11804321, which is located in the last intron of the gene encoding ecotropic viral integration site 5 (EVI5), was shown to overlap with an insulator element that modulates the expression of the neighboring GFI1 via the transcriptional repressor CCCTC-binding factor (CTCF) [64]. More recently, it was reported that the nonsynonymous, strongly associated exonic SNP rs11808092 in the EVI5 gene itself induces changes in superficial hydrophobicity patterns of the coiled-coil domain of EVI5 protein, which, in turn, affects the EVI5 interactome. In particular, this work showed that

Figure 2. An Autoimmunity Gene Network. All known associations with common autoimmune diseases were obtained from the Genome-Wide Association Study (GWAS) Catalog [92] and mapped to the closest or most likely affected gene, as reported by the authors of each publication. Then, data were displayed graphically as a network using the software Cytoscape [93]. Nodes represent either diseases (circles) or genes (triangles), while edges indicate genetic associations. Genes associated with more than one disease are shown in yellow and lie at the center of the network. Genes associated with only one disease are displayed with the same color as their corresponding disease and organized in an outward fashion.
the EVI5 protein bearing the risk allele selectively interacts with sphingosine 1-phosphate lyase (SGPL1), an enzyme important for the creation of the S1P gradient, which is relevant to the adaptive immune response and the therapeutic management of MS [62]. Thus, it is conceivable that the association peak in this region truly represents multiple independent functional effects.

A specific DNA polymorphism in TNFSF113B, insertion-deletion GCTGT→A (BAFF-var), was recently reported to be associated with MS risk in Sardinians [65]. Whereas the frequency in Sardinians is approaching 30%, as a result of environmental pressures, BAFF-var frequency in other ancestral groups, including Europeans, ranges from <1% to 6%. TNFSF113B encodes the cytokine BAFF, which belongs to the TNF ligand superfamily. BAFF is expressed in B lineage cells and acts as a potent B cell differentiator and activator. Noteworthy, BAFF-var results in a shorter transcript that escapes miRNA inhibition, resulting in increased production of soluble BAFF. In addition to disease risk, BAFF-var was shown to be significantly associated with several endophenotypic traits, such as B cell and monocyte counts and total IgG, IgA, and IGM levels. The authors hypothesized that BAFF-var carriers would have a weaker response to B cell-depleting therapies due to an accelerated resurgences of memory B cells.

Additional studies with individual risk variants have been performed but current understanding of the function of each bona fide susceptibility variant in the molecular pathogenesis of MS is superficial. Moreover, since DNA variants are likely to act in concert to confer risk, studies on their interaction and on how biological pathways are affected by them are needed.

**Gene x Environment Interactions in MS**

A reasonable interpretation of the incidence increase and distinctive geoprevalence distribution of MS implicates precipitating environmental or lifestyle triggers and/or the disappearance of protective environmental factors. The biological interactions between these environmental factors and genetic variance could also explain the inherent complexity of MS as a genetic trait. In recent years, several nongenetic risk factors for MS have been identified. These include vitamin D deficiency, exposure to EBV after early childhood and manifestations of infectious mononucleosis, and cigarette smoking [66,67]. Although the causality of these associations remains controversial, there is justifiable interest in formally assessing the additive interactions between exposure and the emerging MS-susceptibility DNA map [68]. These studies have been limited to HLA genes so far, with positive results reported for active and passive smoking, EBV infection and infectious mononucleosis, and childhood and/or adolescent obesity (reviewed in [5]). For example, in a pooled analysis of six data sets, smokers carrying HLA-DRB1*15 and lacking HLA-A*02 had a 13-fold increased disease risk compared with never-smokers without these genetic risk factors (OR 12.7, 95% CI 10.8–14.9) [69]. Novel modeling approaches will afford the analysis of gene x environment interactions genome-wide [70], potentially addressing an important knowledge gap in our understanding of MS heritability.

**From Individual Loci to Pathways and Beyond**

The multiple efforts in mapping MS susceptibility loci have succeeded in identifying more than 200 independent associations. Without an elaborated analysis, the overall flavor of the associations can be assumed to be heavily immunological. Furthermore, several associated SNPs have been shown to be expression quantitative trait loci (eQTL) in different cells of the immune system. For example, associated variants that are eQTLs for genes such as IRFIM3, CD37, and CD6 in CD4+ T cells strongly suggest that they have an effect in modulating adaptive immune responses. Furthermore, eQTL variants in or near CLECL1, RGS1, and MERTK in CD14+ monocytes suggest that the innate immune system is also involved. Notably, variants in genes expressed in the central nervous system (CNS) are less common in these studies, thus
supporting a model in which genetic susceptibility to MS is mediated primarily by sustained dysregulation of immune responses over several decades before clinical symptoms appear.

However, as described above, the functional consequences have been described for only a handful of loci. Another way to contribute to the interpretability of genetic associations is to conduct pathway analysis. In its simplest form, pathway analysis is the search for enrichment of associated signals at or near genes with similar or complementary functions. Analogous to how large gene sets from transcriptomics data are analyzed, associated polymorphisms can be mined for their effects on genes from common predefined processes, such as gene ontologies or biological pathways. The main challenge with GWAS signals is that the affected gene(s) is not always obvious. While some analyses focus on the closest gene to the risk-associated SNP, it has now become clear that conserved motifs from promoters, enhancers, and regulatory elements can be far away from the target gene(s). Modern approaches will likely incorporate this type of information from projects such as REP [56], Encyclopedia of DNA elements (ENCODE) [71,72], and the International Human Epigenome Consortium (IHEC) [73]. Once the SNP-to-gene mappings are performed in a biologically meaningful way, enrichment of the putative genes in specific pathways can be evaluated [74,75].

Instead of performing enrichment analysis with canonical pathways (an approach that could lead to biased results, based on our current knowledge) it was proposed to merge the statistical evidence of association with the biological evidence of interaction at the protein level by virtue of a protein interaction network (PIN) [76–79]. The main question posed by this approach is whether genes known to be regulated by risk SNPs are more likely to interact among them than what would be expected by chance. A significantly higher number of interactions would indicate that those genes are likely participating in a common biological process (independently of whether that pathway was described in a database). This approach was termed ‘PIN-based pathway analysis’ (PINBPA) and a user-friendly analysis tool as been created (as an App for the software Cytoscape), which currently has more than 4000 users worldwide [80,81].

To provide a more accurate representation of the overall risk posed by MS susceptibility loci, we propose an approach that integrates PINBPA with cell-specific regulatory signals (Figure 3, Key Figure). Furthermore, this approach can also be used to compute personalized risk profiles, an area of much interest within the precision medicine community.

Concluding Remarks
It is hard to argue against the contribution of common genetic variation to the susceptibility of MS. While discovery of non HLA-related loci lagged for some time, MS is one of the great successes of the GWAS era in human genetics. The impressive pace of discovery over the past 10 years highlights the need for collaboration, because the number of cases necessary to achieve statistical power is only possible in the context of international consortia.

While additional susceptibility GWAS focusing on common variants will likely be more sporadic in the future, there is still much to be done (see Outstanding Questions). As efforts on precision medicine evolve, we anticipate that the study of non-European populations will take a more prominent role. Also, studies of rare variants will likely be undertaken, either by whole-genome or exome sequencing, or targeted resequencing efforts. Similarly the contribution of genetic variation to endophenotypes has not yet been fully explored and, with the advent of wearables and other digital technology that can precisely sense and record internal variables (gait, vision, MRI, micro-movement detection, etc.), we anticipate that these studies will see much development.

Finally, the development of affordable and massively parallel sequencing technologies has enabled the exploration of genomes other than our own, giving birth to the field of
metagenomics. It is expected that, in the near future, large studies will be carried out to characterize the genomes of all microscopic entities that coexist with humans (e.g., bacteria, Archaea, viruses, etc.) and their potential relationship to disease triggering or perpetuation.

In summary, the genetics architecture of MS has been largely unraveled by international, collaborative efforts. The next frontier appears to be the identification of their functional consequence, the pathways affected by these variants, and finally, the integration of multimodal data sources into a subject’s medical record, a process that could bring the promise of precision medicine in MS one step closer.
Acknowledgments

We thank the National Institute of Health’s (NIH) Institute for Neurological Diseases and Stroke (NINDS) (grants R01NS088155 and R01NS26799). S.E.B. is the Heidrich Family and Friends Endowed Chair in Neurology at UCSF. J.R.O. is the G. Zimmerman Endowed Chair in Neurology at UCSF.

Resources

www.roadmapepigenomics.org
www.cytoscape.org

References


86. Bertrams, J. et al. (1972) HLA-antigens and multiple sclerosis. Tissue Antigens 2, 405–408


