2014-08-01

On the Allelic Architecture of Multiple Sclerosis in Sardinia

Athena Hadjixenofontos

University of Miami, hadjixenofontos@gmail.com

Follow this and additional works at: http://scholarlyrepository.miami.edu/oa_dissertations

Recommended Citation

http://scholarlyrepository.miami.edu/oa_dissertations/1272

This Open access is brought to you for free and open access by the Electronic Theses and Dissertations at Scholarly Repository. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of Scholarly Repository. For more information, please contact repository.library@miami.edu.
ON THE ALLELIC ARCHITECTURE OF MULTIPLE SCLEROSIS IN SARDINIA

By
Athena Hadjixenofontos

A DISSERTATION

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Coral Gables, Florida
August 2014
A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

ON THE ALLELIC ARCHITECTURE OF MULTIPLE SCLEROSIS IN SARDINIA

Athena Hadjixenofontos

Approved:

________________                _________________
Jacob McCauley, Ph.D.         Margaret Pericak-Vance, Ph.D.
Associate Professor of Human Genetics         Dr. John T. Macdonald Foundation
                                      Professor of Human Genetics

________________                 _________________
William Scott, Ph.D.             M. Brian Blake, Ph.D.
Professor of Human Genetics         Dean of the Graduate School

________________
Alberto Pugliese, Ph.D.
Research Professor of Medicine
Multiple Sclerosis (MS) is a demyelinating disease of the central nervous system with autoimmune etiology. It affects approximately 2.3 million people worldwide, but prevalence is distributed unequally with countries closer to the equator manifesting a lower prevalence of MS. The Italian island of Sardinia is an exception, with prevalence rates that are among the highest in the world. Sardinia is inhabited by a unique, isolated population that was founded approximately 10,000 years ago. The reasons for this enrichment of MS cases in Sardinia are unknown.

Like most complex diseases, MS has both genetic and environmental components of susceptibility. To date, research has uncovered the identity of 114 Single Nucleotide Polymorphisms (SNPs) which tag loci that explain approximately 27% of the genetic factors that drive MS susceptibility, in populations of Northern European ancestry. With the exception of the effect exerted by polymorphisms in the Human Leukocyte Antigen \( DRB1 \) gene, these genetic susceptibility alleles have small to moderate effect sizes (Odds Ratio range 1.03 to 1.34) and are largely common in the population (Risk Allele Frequency range 0.09 to 0.95). There are multiple reasons to explore the hypothesis that the Sardinian population may be enriched for the risk alleles that drive MS susceptibility, such as the high prevalence of MS and predictions made by population genetics theory with regard to the genetic landscape of isolated populations.
Past studies in the genetics of MS in Sardinia have uncovered regions of the genome with possible roles in MS pathogenesis that display little overlap with regions identified in other populations. In the present study, I examined the presence of established MS-associated SNPs in a dataset of 19 multiplex Sardinian families. Although the Northern European-derived risk variants are present in Sardinians, these are able to differentiate patients from unaffected Sardinian individuals only when considered cumulatively, with the use of a weighted genetic burden score.

The presence of multiple MS cases in the same family afforded us the opportunity to search for genetic variation that affected relative pairs may share from a common ancestor. Five regions with suggestive amounts of allele sharing were detected (logarithm of the odds (LOD*) score ≥ 1); fine-mapping underneath these linkage peaks identified four genes that may be relevant in MS pathogenesis in Sardinia (EPHA7 on 6q16.1, JAZF1 on 7p15.1, KLRC2 on 12p13.2 and CD226 on 18q22.2). Interestingly, the chromosome 12 peak spans the natural killer cell gene cluster at that location.

I therefore used whole exome sequencing data of the affected individuals from 5 of the Sardinian multiplex families to search for rare, nonsynonymous variants. I identified two variants in IKZF1 at 7p12 and MANBA at 4q24, two genes that are implicated in MS via the established associations. These variants are conserved and predicted to be probably damaging to the protein product. I also found a range of variants in the genes underneath the linkage peaks, highlighting the importance of cumulative assessments of the burden of rare and common variants in disease.

In total, these data indicate that the overall MS susceptibility landscape in Sardinia is not markedly different from that of outbred European populations, and likely
includes both common and rare risk alleles. However, these data also highlight the utility of multiplex families from an isolated population in the initial identification of possible risk alleles. Replication in large population samples is required to assess the relevance of the identified variants in MS pathogenesis.
To my parents, Savvas and Cleo
Acknowledgements

I owe a debt of gratitude to my advisor, Dr. Jacob L. McCauley, for his years of excellent mentorship. He has provided me with sound guidance and ample opportunities for professional enrichment that have helped me to grow as both a scientist and a person. I also wish to thank my committee members, Drs. William K. Scott, Alberto Pugliese, and Margaret Pericak-Vance, for their guidance and encouragement as I completed my research.

I am also grateful to Dr. Luisa Bernardinelli for giving us the opportunity to work with her unique dataset of multiplex Sardinian families. In addition, discussions of the analyses with members of the International Multiple Sclerosis Genetics Consortium (IMSGC) have been monumental in shaping my analytical approaches. Specifically, Dr. Pierre Gourraud based at the University of California San Francisco was instrumental in the analysis of genetic burden. Most of the analyses would not have been possible without the computing infrastructure and support provided by the Center for Computational Science at the University of Miami.

My scholarship and research were supported financially by the Programs in Biomedical Sciences at the University of Miami, Miller School of Medicine, the fellowship provided by the Dr. John T. Macdonald Foundation Department of Human Genetics at the University of Miami Miller School of Medicine, and by grant RG 4680A1/1 from the National Multiple Sclerosis Society (NMSS). Finally, I gratefully acknowledge the resources provided by the Biorepository, Genotyping and Bioinformatics Cores at the John P. Hussman Institute for Human Genomics.
# Table of Contents

List of Figures ............................................................................................................. vii

List of Tables .............................................................................................................. xi

Chapter

1 Motivation ............................................................................................................... 1
  1.1 The immunopathogenesis of multiple sclerosis ........................................ 1
  1.2 Clinical and epidemiological characteristics of multiple sclerosis .......... 3
  1.3 Genetic and environmental susceptibility to multiple sclerosis ............ 5
  1.4 The allelic architecture of complex disease .............................................. 8
  1.5 The Sardinian population isolate .............................................................. 13
  1.6 The genetics of multiple sclerosis in Sardinia ......................................... 17
  1.7 Avenues for exploration ............................................................................ 18

2 Common Multiple Sclerosis Susceptibility Alleles in Sardinia ..................... 20
  2.1 Polygenic burden scores and their relevance in diverse populations....... 20
    2.1.1 Assessment of cumulative genetic burden in polygenic disease .... 20
    2.1.2 Model assessment tools..................................................................... 21
    2.1.3 The overlap of genetic susceptibility between populations ........... 23
    2.1.4 Motivation ......................................................................................... 24
  2.2 Experimental design................................................................................... 25
    2.2.1 Dataset description ............................................................................ 25
    2.2.2 Genotyping and quality control procedures ...................................... 26
    2.2.3 Statistical analysis............................................................................. 28
  2.3 Sardinian cases carry MS-associated alleles discovered in Northern Europeans ............................................................. 32

3 Allelic Sharing in Sardinian Affected Relative Pairs with Multiple Sclerosis 38
  3.1 Contextual framing of the linkage scan .................................................... 38
    3.1.1 Linkage analysis of complex traits .................................................... 38
    3.1.2 Past linkage scans in multiple sclerosis ............................................ 41
    3.1.3 Mapping MS regions in Sardinia ...................................................... 43
    3.1.4 Motivation......................................................................................... 45
  3.2 Methodological approach........................................................................... 46
    3.2.1 Selection of the linkage map ............................................................ 46
    3.2.2 The multiplex Sardinian pedigrees .................................................. 50
3.3 Non-parametric linkage results ................................................................. 54
3.4 Association hits in the Sardinian linkage peaks ...................................... 58
3.5 Fine-mapping of the linkage peaks using a family-based association test ....................................................... 58

4 Exome Variants in Large Sardinian Pedigrees of Multiple Sclerosis .......... 66
4.1 The case for rare variants in disease susceptibility .................................. 66
   4.1.1 Population genetic theory and rare variants ...................................... 66
   4.1.2 Examples of rare variants in complex disease ............................... 69
   4.1.3 Family study designs in the search for rare variants ....................... 74
4.2 Methodological approach ...................................................................... 76
   4.2.1 Characteristics of the 5 sequenced pedigrees .................................... 76
   4.2.2 Exome sequencing, alignment and processing ................................. 78
   4.2.3 Quality control and annotation of exome sequencing data ............... 80
4.3 Coding variants in regions shared between affecteds .............................. 83
   4.3.1 Rare variants shared identical-by-descent in MS-associated regions .......................................................... 83
   4.3.2 Common and rare variants under linkage peaks ............................. 92
      4.3.2.1 A closer look at functional classes ........................................... 92
      4.3.2.2 The minor allele frequencies of variant under the linkage peaks ......................................................... 97
      4.3.2.3 Genes implicated by rare, probably damaging variants .......... 102
      4.3.2.4 Genes implicated by rare, probably damaging variants that have risen in frequency in Sardinia ................................. 104
      4.3.2.5 Variants implicated by fine-mapping under the linkage peaks .......................................................... 105
   4.2.3 Summary ........................................................................................ 109

5 Discussion .................................................................................................. 111
5.1 Summary of findings ............................................................................. 111
5.2 Synthesis ............................................................................................... 124

References ................................................................................................ 127
List of Figures

Figure 1.1. Prevalence of multiple sclerosis (MS) in Europe, showing the latitude gradient of prevalence. Note that Sardinia is colored with the average prevalence rate in Italy, and not the prevalence rate in the Sardinian population isolate. Adapted from Atlas of MS © MSIF 2013. .................................................. 5

Figure 1.2 A visualization of the polygenic nature of MS as assessed by the en masse analysis. The y-axis displays a measure of variance explained by each single nucleotide polymorphism (SNP) set, calculated as the difference in Nagelkerke’s Pseudo-$R^2$ between a logistic regression model that contains the polygenic SNP set score and the covariates (full model) versus a model with the covariates alone (reduced model). Each of these sets of SNPs shows highly significant association to disease status with p-values that range between 3.14E-9 and 6.21E-21, as assessed by a 1-df likelihood ratio test between the full and reduced models. The $p < 0.2$ threshold includes 21% of all SNPs in the dataset. The apparent plateau is likely to be due to limited power rather than an absence of true effects in the p-value range between 0.2 and 0.5. Adapted from (1).......... 10

Figure 1.3 The posterior probability distributions of genotypic relative risk (GRR) and minor allele frequencies (MAFs) of the disease-associated SNPs inferred from the polygenic analysis results for four diseases (a) Rheumatoid Arthritis (excluding known risk loci), (b) Celiac Disease with the major histocompatibility complex (MHC) alleles removed, (c) Myocardial Infarction/Coronary Artery Disease, and (c) Type II Diabetes. Adapted from (2)................................................................. 12

Figure 2.1 The distribution (a) of the odds ratios for the 113 MS-associated SNPs included in the burden score, and (b) the risk allele frequencies for the 110 non-MHC MS-associated single nucleotide polymorphisms in unaffected population samples. RAF = Risk Allele Frequency. .......................................................... 29

Figure 2.2 The distribution of the MHC-inclusive genetic burden score across the various groups (a) using the ORs as weights, and (b) un-weighted. The p-values shown correspond to the burden score coefficient from the Generalized Estimating Equations (GEE) model between the groups flanking it. ...................................................... 33

Figure 2.3 The distribution of the genetic burden score that does not include MHC effects across the various groups (a) using the ORs as weights, and (b) un-weighted. The p-values shown correspond to the burden score coefficient from the GEE model between the groups flanking it. ...................................................... 33

Figure 2.4 Reciever Operating Characteristic (ROC) curves for the MHC-inclusive burden score (a) population membership of MS cases (model 2), and (b) disease status (model 1). The Area Under the Curve for each of the models is shown at the bottom. The color codes represent either the weighted (blue) or unweighted (red) burden scores. .. 34
Figure 2.5 There is no significant relationship between the extent of clustering of cases in a family and the mean (a) weighted or (b) unweighted burden score. The multiple regression lines are plotted for the mean family size (21.8 individuals).

Figure 2.6 The distribution of multilocus genotypes in (a) cases from the two populations, and (b) Sardinian cases and controls. The numbers represent the genotype at each SNP such that a 2 represents the homozygotes for the major allele, a 3 represents heterozygotes, and a 4 the homozygotes for the minor allele. Note that the Sardinian cases belong to 19 large pedigrees and that the 4 SNPs that make up the multilocus genotypes are different in the two panels.

Figure 3.1 A diagram showing the quality control steps that resulted in the SNP-set from which the linkage map was created (99,724 SNPs). The number of SNPs shown is the number that remains after all markers that do not satisfy each quality control criterion are dropped. HWE = Hardy-Weinberg Equilibrium.

Figure 3.2 Diagrammatic representation of the algorithm implemented in MASEL and used to select the final linkage SNP set (11,605 SNPs).

Figure 3.3 The average information content for each chromosome, calculated across all families. The x-axis represents genetic distance measured in centiMorgans (cM).

Figure 3.4 Multipoint non-parametric linkage analysis performed using MERLIN (3). Each graph represents one chromosome; the x-axis denotes the genetic length of the chromosome in centiMorgans (cM), and the y-axis denotes the Kong and Cox linear model LOD* score on a scale of 0 to 2. The MHC extends between 53 and 55.5cM on chromosome 6. The red dots at the top represent the physical positions of the current MS-associations.

Figure 3.5 Per-family contributions to the linkage results. Only the chromosomes reaching logarithm of the odds (LOD*) ≥ 1 are plotted. Note that the LOD* scale extends from 0 to 0.8. The legend applies to all chromosomes. The 24 subpedigrees are annotated in the legend with a numeric suffix indicating the subpedigree to which they belong. For example, family 21 was split into subpedigrees 211 and 212, and so on.

Figure 3.6 The results of a family-based test of association (CAPL) inside the support intervals of the linkage peaks that reach LOD* >1. The nominal significance threshold is at -log(P-value) = 3, and the Bonferroni-corrected significance threshold for all 1716 markers is at -log(P-value) = 11.38. Only markers with variance of the test statistic > 5 are plotted. The most significant marker is plotted as a diamond, and the \( r^2 \) between this marker and the remaining ones in the region is indicated by the color of the point, according to the legend. Some of the genes inside the regions are not plotted, as indicated by the note next to each plot. LocusZoom was used to create these plots (4).

Figure 4.1 Summary metrics on the depth of coverage, obtained using Picard (http://picard.sourceforge.net).
Figure 4.2 Diagrammatic overview of the various filtering approaches taken to explore the exome data for variants related to MS susceptibility. IBD = identical-by-descent, both populations = SardiNIA and Exome Variant Server European American datasets. 84

Figure 4.3 Visualization of the segments shared identical-by-descent (IBD) in each subpedigree using Olorin. The height of the bars next to each chromosome represents the number of cases that share that region IBD. The colors represent regions that are shared between N affecteds, where N is a number specified by the user as indicated by Table 4.4. For example, for family 611, 4 out of 6 affecteds with exome sequencing data share any regions IBD, these regions are colored green. The height of the blue bars indicates successively lower thresholds of sharing. For example, for family 611, any 3 out of the 6 affecteds share the regions indicated by the tall blue bars, and any 2 of the 6 share the regions indicated by the short blue bars. Note: Subpedigree 52 showed no shared regions between the two affected individuals. Subpedigree 23_602 is excluded from this figure. 87

Figure 4.4 The (a) alternative allele frequency and (b) conservation scores (GERP) profiles of the 144 variants that are shared IBD in affected individuals in any of the Sardinian families, located in genes that are implicated by the MS-associated loci. 89

Figure 4.5 (a) The Genome Variant Server (GVS) functions of 11,444 variants out of the 12,865. Note that this is presented for descriptive purposes only; the frequencies of variants in each functional class are not corrected for the representation of that functional class in the exome, and (b) the distribution of the probability of damaging effects of 1,855 nonsynonymous variants out of the 23,865 according to PolyPhen2; 348 of these 1,855 variants are annotated as probably damaging; these are distributed among 223 genes/open reading frames. 93

Figure 4.6 The distribution of alternative allele burden for common and rare variants in the 31 cases from 5 multiplex Sardinian multiple sclerosis families in the gene families under the linkage peaks for (a) the CLEC cluster on chromosome 12, and (b) the serpin cluster on chromosome 18 plus SERPINE1 on chromosome 7. 96

Figure 4.7 Bootstrap tests of the mean minor allele frequency of the variants inside the support intervals of the linkage peaks against random samples of the same number of variants from other regions of the clean exome. The red points mark the observed mean SardiNIA MAF for the variants under the linkage peaks = 0.176 (bootstrap P-value = 1E-04), and the observed variance = 0.022 (bootstrap P-value = 0.2025). The distributions are built using 10,000 re-samples drawn with replacement. 99

Figure 4.8 The difference in mean MAF between (a) SardiNIA and (b) Exome Variant Server (EVS) European American datasets, for the variants under the linkage peaks, was tested against random samples from regions outside the linkage peaks. The variants under the linkage peaks have similar (c) differences in MAF mean between the two populations (bootstrap P-value = 0.78) and (d) differences in MAF variance between the two populations (bootstrap P-value = 0.33) to variants outside the linkage regions. The dataset
was restricted to variants that had non-missing allele frequencies reported for both populations. ................................................................. 100

Figure 4.9 Conservation and damage prediction for the variants under the linkage peaks. (a) the mean GERP score from 12,472 variants is similar to random samples from regions outside the linkage peaks (bootstrap $P$-value = 0.75), (b) similar results are observed for the PhasCons scores of 12,855 variants (bootstrap $P$-value = 1), and for (c) the number of variants that are annotated as probably damaging (348 under the linkage peaks) (bootstrap $P$-value = 0.28). ................................................................. 101

Figure 4.10 Characterization of the missense variants under the linkage peaks that are annotated as rare or absent in both the SardiNIA and EVS EA datasets. (a) 19 genes carry > 2 of these 232 variants, and (b) the GERP score distribution of all 232 variants. ... 103

Figure 4.11 Characterization of the 63 variants in 58 genes that are rare in the EVS EA dataset, have a higher frequency in Sardinia, and are annotated as missense and probably damaging by PolyPhen2. (a) Most of the variants with a higher MAF in SardiNIA do not show large allele frequency differences. The variants with the highest differences are in the MYC17 gene. (b) The distribution of the GERP scores. .............................................. 104

Figure 4.12 The burden of alternative alleles for both common and rare variants in the genes implicated by fine-mapping underneath the linkage peaks for 31 affecteds from 5 multiplex Sardinian multiple sclerosis pedigrees. (a) the EPHA7 gene on chromosome 6, (b) JAZF1 on chromosome 7, (c) KLRC2 on chromosome 12, and (d) CD226 on chromosome 18. ................................................................. 108
List of Tables

Table 2.1 Clinical characteristics of the Sardinian and US cases. PR = Progressive Relapsing, CIS = Clinically Isolated Syndrome, SE = Standard Error. ........................ 27

Table 2.2 The four SNPs that were used in constructing the multilocus genotypes for US and Sardinian cases. Note that risk refers to the risk of being a US case, and not the risk of disease. F = frequency, the subscripts are defined as follows: USC = US cases, SC = Sardinian cases, SP = Sardinian population (retrieved from the SardiNIA Medical Resequencing Project on the dbGaP repository), CEU = Centre d' Etude du Polymorphisme Humain (retrieved from dbSNP with respect to the HapMap CEU dataset). .......................................................................................................................... 31

Table 3.1 The Sardinian dataset .......................................................................................................................... 53

Table 3.2 Descriptive statistics on pedigree structure .......................................................... 53

Table 3.3 Established MS-associated risk alleles near linkage peaks .............................. 58

Table 3.4 Regions under the support intervals of linkage peaks with logarithm of the odds score (LOD*) ≥ 1 .................................................................................................. 59

Table 4.1 Overview of sequenced pedigrees ........................................................................... 77

Table 4.2 Human Leukocyte Antigen (HLA) haplotype status of affected individuals in the sequenced pedigrees ................................................................................. 78

Table 4.3 Available functional annotations for the exome sequencing data .......... 82

Table 4.4 Summary of variants in coding regions that are shared identical by descent (IBD) .............................................................................................................................. 85

Table 4.5 Variants that are rare in outbred Europeans and in genes near MS-associated loci that reside in regions shared IBD in the affected individuals of one or more Sardinian families ............................................................................................................................... 87

Table 4.6 Variants in regions shared IBD between affected individuals in three different families. The variants on chromosome 16 are relatively more interesting because they are shared between sub-pedigrees that originate from unrelated families. ........................................................................... 91
Motivation

1.1 The immunopathogenesis of multiple sclerosis

Multiple sclerosis (MS) is a disease that affects the central nervous system as a result of autoimmune demyelination of axons in the brain and spinal cord. The white matter of the brain is primarily composed of the axons of neurons that have their cell bodies in grey matter regions. These axons are enveloped by a myelin sheath the function of which is to accelerate the propagation of the action potential towards the synapse. Myelination is an ongoing process carried out by oligodendrocytes: a type of glia cell that has multiple processes enabling it to contact many axons and deposit myelin. MS lesions result from the infiltration of the blood-brain barrier by autoreactive, activated T-cells, B-cells,

plasma cells and macrophages leading to the demyelination of the axons and giving rise to a variety of symptoms depending on which region is demyelinated. For example, cortical lesions are associated with cognitive impairment (5), whereas demyelination of the optic nerve is associated with visual impairment and optic neuritis.

Lymphocytes of the adaptive immune response, primarily CD4+ T-cells are involved in the pathogenesis by releasing cytokines and immune mediators to attract macrophages and induce the release of further proinflammatory cytokines. The two major players in the encephalitogenic response are believed to be CD4+ T-helper 1 (Th1) cells which release interferon-γ (IFNγ) and Th17 cells. In contrast to CD8+ T-cells, they require contact with antigen presenting cells (APCs) carrying major histocompatibility complex (MHC) class II molecules on their cell surface for their activation, such as B-cells, dendritic cells and macrophages. CD8+ T-cells may be involved in MS pathogenesis in a more direct way, through the recognition of antigens carried by MHC class I molecules by their T-cell receptors (TCR). MHC class I molecules are expressed by most nucleated cells including neurons and oligodendrocytes. Subsequent activation of CD8+ T-cells leads them to recognize and attack cells of the central nervous system (CNS) (6). Not all populations of T-cells are proinflammatory; certain subpopulations of CD4+ T-helper cells (Th2) may prime remyelination in MS lesions and counter-regulate encephalitogenic responses (7-9). In addition, specific subtypes of CD8+ T regulatory cells may have effector functions that include eliminating pathogenic CD4+ T-cells (10).

Additionally, the detection of complement proteins and immunoglobulin G (IgG) deposits in MS lesions, and of clonally expanded B-cells in the cerebrospinal fluid (CSF) and brain lesions of MS patients, support a role for B-lymphocytes in the pathogenesis of
MS (11, 12). Indeed, B-cell activation marked by the production of oligoclonal bands (OCB) is a consistent finding in MS patients (13). Multiple lines of evidence suggest that B-cells in part exert their effects by influencing the T-cell response via antigen presentation, and cytokine and chemokine production (reviewed by (14)). B-cells can also differentiate into antibody-secreting plasma cells. These plasma cells are implicated in MS pathogenesis by releasing immunoglobulins that target autoantigens on glial cells and activating the complement system. Furthermore, plasma cells may produce autoantibodies that target myelin (for example, (15, 16), with conflicting evidence presented by (17)).

Finally, the participation of natural killer cells indicates the likely involvement of innate immunity in MS pathogenesis. Like T-cells, natural killer cells can also have opposing roles: protecting against autoimmunity through eliminating both APCs and autoreactive T-cells, or by secreting immune regulating cytokines. Depending on the type of natural killer cell and the microenvironment of the extracellular matrix, natural killer cells can also exacerbate autoimmune responses by secreting IFNγ and by contributing to the maturation of APCs (reviewed by (18, 19)).

1.2 Clinical and epidemiological characteristics of multiple sclerosis

The occurrence of a first demyelinating event that manifests as an episode of neurologic symptoms is described by the term Clinically Isolated Syndrome (CIS). Presentation of a single symptom may be indicative of the presence of a lesion in a single brain region, and it is therefore termed monofocal; in contrast, multiple symptoms may indicate multifocal lesions. The most common symptoms include attacks of optic neuritis, weakness on one or both sides, numbness of the face, body or limbs, spasticity, bladder and/or bowel
dysfunction, fatigue, dizziness or vertigo, pain and cognitive changes. Due to the many possible causes of the neurological symptoms, the formal diagnosis of clinically definite MS is one of exclusion (20), and it requires the presence of lesions on magnetic resonance imaging (MRI) scans as part of the diagnostic criteria (21). The global mean age at diagnosis is 29.9 years of age, but the diagnostic lag between the occurrence of the first episode and a definite diagnosis can extend to a few years. In addition, there is a sex bias in MS diagnosis, with females diagnosed twice as often as males. This ratio varies slightly between populations but the general observation that females develop MS more often than males holds true for all of them.

There are four major disease courses, defined by the temporal progression of disability. The majority of patients (85%) are classified into the relapsing remitting type of MS (RRMS), which is characterized by attacks or exacerbations followed by periods of remission. In Caucasians, approximately 80% of RRMS patients transition to the secondary progressive MS (SPMS) disease type, characterized by continuous worsening of symptoms, without relapses and remissions. Patients who initially present with primary progressive disease type (PPMS) (10%) are generally diagnosed a little later; PPMS follows a course of steady worsening of symptoms starting at disease onset. Finally, some patients (5%) are diagnosed with progressive relapsing MS (PRMS) that is characterized by progressive worsening and the intermittent occurrence of acute relapses with no remissions (22).

MS prevalence follows a latitude gradient whereby countries closer to the equator demonstrate lower prevalence (23, 24). The Mediterranean island of Sardinia, Italy, defies this latitude gradient with unusually high prevalence ranging between 143 to 262
per 100,000 people in different regions of the island, with central Sardinia demonstrating the highest rate (25, 26). This is in stark contrast to 40 to 70 cases per 100,000 people in the Italian penisula (reviewed by (27)), and rivals the highest prevalence rates seen across northern Europe.

![Figure 1.1](image)

**Figure 1.1** Prevalence of multiple sclerosis (MS) in Europe. Sardinia is colored with the average prevalence rate in Italy, and not the prevalence rate in the Sardinian population isolate. Adapted from Atlas of MS © MSIF 2013.

### 1.3 Genetic and environmental susceptibility to multiple sclerosis

MS is a complex, multifactorial disease with genetic and environmental etiology. The risk for first degree relatives is estimated to be between 10 and 20 fold higher than the general population, as measured by the recurrence risk in relatives parameter \( (\lambda) \) (28).

The genetic contribution to MS is also evident from twin studies. These studies show an approximate 25-30% concordance rate for monozygotic twins, compared to an approximate 4% concordance rate for dizygotic twins and siblings (29, 30). Adoptee and half-sib studies also support a strong genetic contribution to familial clustering of MS (for example, (31, 32)).

More recently, the landscape of genetic susceptibility to MS in samples of Northern European ancestry is estimated to include hundreds of moderate effects and possibly thousands of small effects, reinforcing the polygenic nature of this complex disease as
well as the need for integrated assessments of cumulative genetic burden (1). Currently, it is estimated that approximately 27% of the genetic component to MS susceptibility has been uncovered, including the MHC effects (33).

The first multiple sclerosis-associated genetic variation was uncovered using cell-culture methods in 1972. This study identified the Human Leukocyte Antigen (HLA) genes which lie in the MHC region of chromosome 6 as an important risk factor in the development of MS (34-36). The MHC is characterized by extensive stretches of linkage disequilibrium (LD) and extremely high levels of polymorphism (37). Therefore, although the initial associations were with HLA-A3, HLA-B7 and HLA-DR2, these were later understood to be reflections of extensive LD in the region. Subsequent molecular dissection using restriction fragment length polymorphisms (RFLPs) verified the association with the DR2 haplotype (referred to as DR15 using updated nomenclature) (38), while other studies determined that the class II genes were the primary drivers of that association (39, 40), and specifically the DRB1*1501 sub type of DR15 (41). Other MHC haplotypes and interactions between DRB1*1501 and other MHC genes have been confirmed to have a role in MS susceptibility (42)(43).

Between 2007 and 2011, seven independent Genome-Wide Association Studies (GWAS) and a meta-analysis uncovered 23 genetic loci associated with MS outside the MHC region (43-49). A second meta-analysis identified another two loci (50), and a subsequent collaborative study between the Wellcome Trust Case Control Consortium 2 (WTCCC2) and the International MS Genetics Consortium (IMGSC) added another 32 loci (51) (reviewed by (52)). The list grew to 110 non-MHC loci in 2013, when the IMMSGC published the results of a high-density association study that used a custom
genotyping chip tiling autoimmune disease related-variation (33). Interestingly, scans for genetic loci that specifically influence disease progression have yet to be unequivocally identified.

Another interesting piece of the puzzle comes from the observation that variants identified in autoimmune diseases overlap. The advantages conferred by the shared pathogenesis of immune-mediated diseases in studying their combined genetics and understanding the underlying mechanisms of pathogenesis have been discussed in the literature (for example (53)). Indeed, a quantitative evaluation of the extent of this sharing across 7 immune-mediated inflammatory and autoimmune diseases (celiac disease, Crohn’s disease, MS, systemic lupus erythematosus, psoriasis, rheumatoid arthritis and type I diabetes) found that 44% of 107 single nucleotide polymorphisms (SNPs) associated with any one of these diseases were in fact shared across multiple diseases. Another interesting finding from the same study is that the variants shared across these diseases also preferentially cluster together in protein-protein interaction networks (54).

Considerable evidence exists for a number of environmental factors, such as cigarette smoking, the plasma concentration of the active form of vitamin D, childhood obesity, and Epstein-Barr virus (EBV) seropositivity, influencing susceptibility to MS (reviewed by (55, 56)). Additionally, these exposures may interact with the genetic background to modify susceptibility. For example, DR15 carriers with elevated antibody titers for an anti-EBV antibody may have as high as 9x higher risk for MS (57), an association which appears to be further enhanced by smoking (58). Molecular evidence suggests an
additional interaction mediated by the regulation of HLA-DRB1 expression by Vitamin D (59).

1.4 The allelic architecture of complex disease

A general overview of the effects that constitute the genetic susceptibility landscape to common, complex diseases is needed to provide the proper context for the present study. That variation in continuous traits can be explained by the independent segregation of multiple factors was proposed by Mendel himself, but then somehow got overlooked during the Biometricians versus Mendelians debate in the early 1900’s. In 1918, Fisher proposed the biometrical model that posits that a large number of genetic effects acting additively and inherited according to the laws of Mendelian inheritance could account for the observation of familial clustering for complex traits (60). The “threshold model” was further developed to describe multifactorial inheritance (61). According to the threshold model, an individual’s liability for a binary trait is directed by an underlying graded attribute. Beyond a certain point on the liability curve (the threshold) the accumulated risk is enough for the individual to manifest a binary phenotype. Two important parameters in defining this space are the number of effects and their associated effect sizes. A number of different statistical approaches have been used to measure the variance in complex phenotypes that is explained by varying sets of effects, affording a – noisy – glimpse into polygenic space.

Early successes in mapping the largest genetic effects in diseases such as MS, combined with the amount of heritability that they explained, set a precedent for the expectation that similar discoveries would follow. It has been suggested that two of the reasons that complicated the identification of further susceptibility genes were the modest
effect sizes attributed to individual loci and the statistical consequences that follow from the enormous size of the genome (62, 63). The emerging genetic landscape of complex diseases is likely comprised of multiple, at-best modest effects giving rise to the extensive use of the metaphor “looking for needles in a very large haystack” (64). Indeed, en masse analyses - where the top associated markers that fall below a certain $P$-value threshold are grouped into a polygenic score that is then assessed with respect to the variance that it explains in the phenotype – including MS, schizophrenia, heart rate, body mass index and height – have produced tangible evidence that common SNPs account for the majority of the remaining heritability but their moderate effect sizes do not allow them to survive stringent multiple testing corrections in GWAS discovery efforts.

In a seminal paper by the International Schizophrenia Consortium, Purcell et al. tested the hypothesis that increasing proportions of small and moderate effects (Genotypic Relative Risk (GRR) $\geq 1.05$) will be detected at increasingly liberal significance thresholds (65). This approach was duplicated in multiple sclerosis, by constructing aggregate quantitative scores for each individual in the sample based on SNP sets defined by varying the significance threshold and testing the score that was derived from the discovery dataset in a separate target dataset (Figure 1.2) (1).

Others have calculated the additive genetic variance that is due to common SNPs through linear mixed-effects modeling, where the random effect represents the polygenic component of the variation (66, 67). A third approach, dependent on the assumption that the distribution of effect sizes of yet undiscovered truly associated SNPs is similar to that of validated associations, estimates the number of additional SNPs with similar effect sizes by using a power correction (68). The application of these methods to complex
phenotypes supports the hypothesis that additional true associations with common SNPs remain to be discovered for most of the phenotypes examined, including MS.

**Figure 1.2** A visualization of the polygenic nature of MS as assessed by the *en masse* analysis. The y-axis displays a measure of variance explained by each single nucleotide polymorphism (SNP) set, calculated as the difference in Nagelkerke’s Pseudo-$R^2$ between a logistic regression model that contains the polygenic SNP set score and the covariates (full model) versus a model with the covariates alone (reduced model). Each of these sets of SNPs shows highly significant association to disease status with p-values that range between 3.14E-9 and 6.21E-21, as assessed by a 1-df likelihood ratio test between the full and reduced models. The p < 0.2 threshold includes 21% of all SNPs in the dataset. The apparent plateau is likely to be due to limited power rather than an absence of true effects in the p-value range between 0.2 and 0.5. Adapted from (1).

Stahl et al. went a step further in characterizing polygenic space by estimating the posterior distribution of the GRR distribution as a function of minor allele frequency (MAF) inside a Bayesian framework (2). Their results in four different complex diseases are consistent with the majority of the effects lying in the range of GRR = 1 to approximately 1.05, and with larger effect sizes seen for SNPs in the lower end of the
MAF spectrum (Figure 1.3). Further modeling, using haplotypes built on the basis of the 1000 Genomes Project data (69), defined the boundaries of allelic heterogeneity that may underlie observed (potentially “synthetic”) associations. The models that are consistent with the results of the Bayesian posteriors harbor 1 to 4 common (MAF > 5%) causal variants, 8 to 16 rare (MAF < 1%) causal variants, or a combination of the two per associated locus (2). For complex phenotypes in general, it seems likely that there are few loci (i.e. on the order of ten loci) that harbor rare variants, since a larger number of these would result in the variance explained exceeding the heritability of the disease. The effect sizes of these the variants that are consistent with the disease parameters for the 4 diseases that were modeled are small. Specifically, the causal variants are predicted to have GRR of 1.04, 1.1, 1.5 and 3.5 for MAF = 0.5, 0.05, 0.01, and 0.001, respectively. Importantly, this does not exclude the possibility of private variants segregating through families, since these would not cause SNP associations in GWAS studies and would therefore not fall under the causal SNP models explored in this paper.

A systematic exploration of the architectures that are consistent with epidemiological data and with the results of past linkage and GWAS scans concluded that models in which rare variants explain either little 20% or most (>80%) of the heritability for Type II Diabetes (T2D) are consistent with the observed results (70). This result is not inconsistent with Stahl et al. since the two studies are taking markedly different approaches. It does however indicate that results are dependent on the parameters that were modeled. In order to demystify the identities of the remaining effects, whether they are common with moderate GRRs or rare with larger GRRs or most likely, a combination of the two, extremely large samples will be required (>250,000 individuals) (70).
Figure 1.3 The posterior probability distributions of genotypic relative risk (GRR) and minor allele frequencies (MAFs) of the disease-associated SNPs inferred from the polygenic analysis results for four diseases (a) Rheumatoid Arthritis (excluding known risk loci), (b) Celiac Disease with the major histocompatibility complex (MHC) alleles removed, (c) Myocardial Infarction/Coronary Artery Disease, and (d) Type II Diabetes. Adapted from (2).

Furthermore, an important defining parameter of the overall polygenic landscape of many complex phenotypes including MS is captured by the relationships between identified risk loci. It is likely that despite the multiplicity of effects that are driving
disease risk, a smaller number of biological pathways are implicated. Multiple functional
groupings have been described including those based on tissue-specific expression,
physical protein-protein interactions, descriptions in the literature, and molecular
biological pathways. In MS, the latest pathway analysis included 57 non-MHC
susceptibility loci and a dataset of 15,000 cases and 30,000 controls (71). Indeed,
products of associated genes are more likely to physically interact with each other, and
belong to related pathways, thereby effectively capturing informative trends from
univariate, single-locus analyses.

1.5 The Sardinian population isolate

The genetic architecture of any trait is shaped by multiple population genetic forces such
as mutation, random genetic drift, and selection. Population isolates are defined by their
specific demographic history, arising from a generally limited number of migrating
founders and shaped by subsequent genetic isolation. Multiple forces that shape the
genetic parameters of a population have pronounced effects in isolates. Specifically,
founder effects decrease variability and random genetic drift that drives alleles towards
extinction or fixation has the opportunity to do so in a shorter time frame. It follows that
the size of the founder population and the rate of expansion are important parameters in
defining the magnitude of genetic drift and the resulting susceptibility landscape for
polygenic diseases.

Isolated populations have been especially attractive for mapping trait loci for
monogenic phenotypes. One of the major reasons for this is the increased probability that
individuals will share alleles derived from a handful of ancestral haplotypes. This relative
increase in homogeneity and reduction in genetic complexity and variation has enabled
the discovery of disease alleles for rare, recessive diseases from very small numbers of affected individuals. For example, the Finnish isolate has proven useful in mapping genes for rare Mendelian phenotypes (72), but success has been elusive for common, complex diseases. This demonstrates that whether or not an isolated population will provide an advantage depends on the hypothesis that is being examined (73). Indeed, the patterns of LD observed in both the Sardinian and Finnish population isolates suggest that these populations will not offer an advantage over heterogeneous populations under the common-disease common-variant hypothesis (74). On the other hand, the reduction in genetic heterogeneity may still prove useful in the study of complex disease genetics (75). Furthermore, the use of isolated populations has uncovered Mendelian subtypes of otherwise more complex diseases, such as Hirschsprung disease in the Mennonites and nonsyndromic hearing loss in the Bedouins (76-78).

The demographic history of a population determines many of the parameters that are of interest in disease and population genetics. Sardinia is an autonomous region in Italy separated from Corsica by the straits of Bonifacio to the north. It is composed of a main island and many smaller islands. According to the National Institute of Statistics in Italy, the current population of Sardinia is 1,675,411 inhabitants. Archeological evidence such as stone instruments indicates the presence of humans on the island during the lower Paleolithic period. Human remains from the upper Paleolithic period confirm that the island was inhabited about 14,000 years ago. During the Neolithic period, material aspects of the culture suggest an active exchange with civilizations in the Aegean Sea, while the architectonical structure of Bronze Age monuments termed the “nuraghi” suggests influences that originate from the Balearic Islands and Corsica. Other Bronze
Age metalworking objects demonstrate similarities to those found in France and the Iberian Peninsula during the same time.

Despite the extensive apparent contact of Sardinia with other populations and the conquests by the Phoenicians, then the Carthaginians and subsequently the Romans, the linguistic and demographic histories of Sardinia argue for a limited impact on the introduction of outside genetic influences (79), and for the existence of isolated sub-regions on the island.

On the basis of mitochondrial haplogroup analysis and in contrast with mainland populations, the population underwent two demographic expansions, one right after the Last Glacial Maximum (20,000 to 16,000 years ago), and a more recent one during the early Neolithic period (12,500 to 2,900 years ago) (80). However, central Sardinia where the Nuoro province is located indicates shared ancestry with the Iberian Peninsula based on their sharing of mitochondrial haplogroup V, supporting a more prominent role for a Paleolithic background in this sub-isolate (81).

These population expansions were relative, as Sardinia showcases a low population density throughout its history. At around 1700 C.E. the total population reached the considerable size of 300,000 individuals (82); translating to 6 inhabitants per square km providing an indication of the internal isolation of Sardinia. Indeed, the neo-Latin Sardinian language is subdivided into ten main variants, and a further 30 subdivisions. This microgeographic linguistic heterogeneity is mirrored by observations of genetic heterogeneity within the island and the presence of multiple sub-isolates (83, 84).

Multiple observations support a large role for genetic drift in shaping the genetic characteristics of this population. These include the low population density throughout
Sardinian history, the internal isolation due to the presence of natural boundaries, the high levels of consanguinity and the occurrence of major bottlenecks such as the Black Plague that reduced the population by half around the year 1348 C.E. The effects of drift are observed through many classical genetic markers which show dramatically different allele frequencies in Sardinia as do the disease alleles for thalassemia and glucose-6-phosphate dehydrogenase deficiency (reviewed by (85)). It has been suggested by Haldane that these polymorphisms may provide a selective advantage against malaria (summarized in (86)), a disease that has historically been present in Sardinia.

Studies of the HLA region confirm the unique status of Sardinians among neighboring populations. The HLA genes have also been used in defining internal heterogeneity in Sardinia and the differentiation of this population, but they are limited by the complexity of forces that act on them (87). A study of Y-chromosome variants that do not appear to be subject to selection revealed a Y-chromosome haplogroup (I-M26) that is rare or absent outside Sardinia with a frequency of 0.37 within the island, indicating a founder effect and supporting the pre-Neolithic settlement of the island followed by little gene flow from other populations (82, 88, 89). Analysis of low-pass whole genome sequencing data of the Y-chromosome provides additional support for these as well as the Neolithic expansion findings and confirms a putative age for coalescence with other human populations of approximately 180,000 to 200,000 years ago (89). Of special interest is the high degree of isolation and genetic differentiation observed for the province of Nuoro, the isolated, mountainous province that serves as the site of ascertainment for the present study, on the basis of Alu insertions (90) as reviewed by (85).
Finally, consanguinity has also played a part in shaping the genetic landscape of the Sardinian population. Until recently, the levels of endogamy, i.e. marrying within the commune, in various sub-regions of the province of Nuoro were as high as 90% (91, 92). The rate of consanguineous marriages in the Nuoro communes ranged from 13.87% to 19.9%, compared to a Sardinia-wide average of 4.44%. The same study calculated a range of inbreeding coefficients between 3.5E-10 and 4.69E-10, comparable to the genetic distance between third cousins (3.9E-10) (91, 92). This is similar to estimates from the north-east coast of Sardinia, although in some regions the inbreeding coefficient drops to 1E-10 (92).

In summary, the Sardinian population is a unique population with pre-Indo-European origins that is differentiated from the continental Italian and European populations of present day. It is also characterized by internal fragmentation into numerous sub-isolates, with forces such as founder effects and genetic drift appearing to have played a major role in shaping the genetic landscape that characterizes this population.

1.6 The genetics of multiple sclerosis in Sardinia

The population history of Sardinia combined with the high prevalence of multiple autoimmune diseases suggest multiple advantages in carrying out genetic studies in this population (93). For example, on the basis of linkage disequilibrium (LD) patterns on the X chromosome, it was determined that in Talana, one of the isolated sub-populations in Sardinia, 80% of inhabitants were direct descendants of just 8 paternal and 11 maternal lineages (94). However, linkage and GWAS studies in this population have suggested a few potentially relevant regions outside the MHC region, but no major effects (47, 95-97).
While the clinical phenotype is indistinguishable between Sardinian and Northern European cases, the relevance of the established MS risk alleles in Sardinia is unknown. The largest genome-wide association study (GWAS) in Sardinia uncovered \textit{CBLB} on chromosome 13 as a possible target of disease predisposing mutations (47), an association that has yet to be replicated. Furthermore, the haplotypes of HLA genes on chromosome 6p21 that are associated with MS are known to differ in the Sardinian population, where DR3 and DR4 account for the main risk effect attributed to this region (98-101). Despite having the strongest effect, HLA DRB1 alleles make a relatively small contribution ($\lambda_{S\text{HLA}} = 1.6$ to 1.78) to familial aggregation of MS as found by studies in UK and Sardinian families (Marrosu, 2000).

Lastly, the dataset of extended, multiplex Sardinian families used in the present study has previously been explored for the association of a candidate gene, Amiloride-sensitive Cation Channel Neuronal 1 (\textit{ACCN1}) to multiple sclerosis, with positive results (multiple testing corrected $P$-value = 0.002) (97).

\textit{1.7 Avenues for exploration}

Accumulation of genetic susceptibility, the homogeneity of environmental factors and their interplay may explain the high prevalence observed in Sardinia, not only for MS but also for other autoimmune diseases (102). I hypothesize the involvement of the known common variants, identified in other populations with European ancestry, in the pathogenesis of MS in the high-risk Sardinian population isolate.

Furthermore, the familial clustering observed in the extended pedigrees used in the present study may be due to the presence of variation with relatively large effects than previously identified in other European populations. This dataset affords improved
power to detect these relative to sib-pair studies because the affected relative pairs in this
dataset are more distantly related than those used in previous investigations. I therefore
sought to identify these effects by defining regions of sharing between affected relative
pairs and examining both the common and rare variation in those regions using multiple
different approaches, as outlined in Chapters 3 and 4.
Common Multiple Sclerosis Susceptibility Alleles in Sardinia

2.1 Polygenic burden scores and their relevance in diverse populations

2.1.1 Assessment of cumulative genetic burden in polygenic disease

Our current understanding of polygenic disease rests on the hypothesis that an accumulation of susceptibility alleles leads to increasing disease risk in individuals (60). These risk alleles are sampled from the population so that a population with many high-risk individuals is likely to manifest a higher prevalence of the disease than a population with fewer high-risk individuals. For example, variants that are annotated as possibly damaging are not observed at a heavier load in European American individuals than in

---

African American individuals, despite the observation that on the population level, European Americans carry more deleterious variation than African Americans (103). In general, examining the risk alleles in aggregate thereby taking into account their cumulative effect is likely to be useful in characterizing the genetic burden for MS-associated alleles in individuals from the Sardinian population.

Cumulative genetic burden scores are most commonly used for the purpose of disease prediction in both research and clinical settings, with varying degrees of success. For example, genetic risk scores have been used to predict seropositive rheumatoid arthritis (104), early onset myocardial infarction (105) and multiple other complex phenotypes. Risk scores have also been used in attempting to differentiate between early and late onset cases, for example in Crohn’s Disease (106), and for prediction of secondary disease characteristics (107). The use of endophenotypes has also proved useful: for example, a weighted risk score using loci for obesity, distinguished individuals based on their accumulation of adipose tissue (108). Additional clinical factors are often used where the intention is clinical disease prediction, with consistent improvements in predictive ability.

Additionally, pathway-wide genetic load scores have been used in discovery studies. For example, a load of 5 or 6 in the renin-angiotensin-aldosterone system is associated with 2.3 fold increase in CHD risk in cases of familial hypercholesterolemia, relative to patients with a load of 0 or 1 (109).

2.1.2 Model assessment tools

Polygenic burden scores are often used in logistic regression models where the cumulative score is used as the predictor and the binary disease state is modelled as the
outcome. The predictive ability of such a binary classifier is often assessed using the area under the curve (AUC) of receiver operating characteristic (ROC) curves. ROC curves plot the true positive rate (sensitivity) as a function of the false positive rate (1-specificity). They are named after their original application during World War II in radar detection of enemy objects in battlefields. As a reference point, an AUC of at least 0.9 is considered to be reflective of a practically useful classifier. In MS, the AUC that describes the ability of polygenic burden scores to differentiate cases from controls has increased from 0.63 using just 16 MS-associated loci \((110)\), to 0.75 using 56 MS-associated SNPs. Both of these estimates include the effects of the MHC and female sex. Before the list of 56 was available, an attempt to include nominally significant effects (uncorrected genome-wide P-value < 0.05) resulted in a polygenic score of 350 SNPs, with an AUC of 0.76 \((111)\). In comparison, a model that distinguishes between MS probands and unaffected siblings performs slightly worse (AUC = 0.57 for 16 loci, and AUC = 0.59 for 56 loci) \((112, 113)\). This observation may indicate that controlling for genetic and environmental risk factors that siblings have in common results in lower predictive ability.

Additionally, the behavior of MS genetic burden (MSGB) scores in multiplex families has been assessed. Previous versions of the MSGB score, incorporating variants in 13, then 16 and subsequently in 64 loci, were demonstrated to be higher in cases from multicase US MS families than in sporadic cases \((112-114)\). This observation was hypothesized to be driven by the higher frequency of a subset of the effects in multiplex families: for example, a risk variant in the interleukin 7 receptor alpha chain \((IL7R)\) gene is found in 80% of familial cases compared to 75% of sporadic cases \((114)\).
Despite their currently limited clinical predictive ability, genetic burden scores are useful in stratifying individuals based on their risk. This application is becoming attractive in current discovery efforts, since burden scores based on the current roster of susceptibility loci can identify high-risk controls and low-risk cases. These two groups are expected to be enriched in the genetic (risk or protective) factors that remain to be discovered. Therefore, focusing on these groups of patients can increase statistical power by increasing the apparent effect size.

2.1.3 The overlap of genetic susceptibility between populations

In calculating burden scores, researchers have used different sets of SNPs for different populations. For example, two different sets of minimally overlapping SNPs were used in predicting incidence of coronary heart disease (CHD) in blacks and whites, generating different results in the predictive ability of the models (115). It is possible that these sets reflect differences in the tag SNPs that are associated with the CHD in each population and not necessarily differences in the causal, unobserved markers. These tag SNP differences are expected as they reflect differences in the linkage disequilibrium (LD) patterns between the two populations.

Additionally, the choice of the weights used in calculating weighted burden scores indirectly depends on differences in LD between the two populations. This is because the Odds Ratios (ORs) that are often used to weigh the presence of the risk allele depend on the strength of the LD between the tag SNP and the causal polymorphism that is suspected of having a direct role in pathogenesis.

The extent of LD and therefore the relevance of the Northern European-derived OR is an important consideration in the present study. Although one might suspect that
the demographic history of Sardinia may lead to longer stretches of LD, studies have shown that LD in Sardinians is no more extensive than in outbred European populations (116, 117). This may be explained, at least in part, by the age of the isolate.

Multiple lines of evidence suggest that it is unlikely that all genes relevant to MS in Northern European populations are important in Sardinia. Linkage and association screens in the Sardinian population have proposed novel and potentially population-specific regions of interest. An alternative explanation for this observation may be that initial efforts to map MS susceptibility in both populations have been severely underpowered given the small effect sizes of the underlying risk variants. At the same time, it is possible that different polymorphisms in the same regions will be important in different populations, due to differences in allele frequencies that are shaped by demographic history. An example of this is the HLA region which is relevant in both Sardinians and Northern Europeans, but for which the relevant haplotypes differ in the two populations (118).

2.1.4 Motivation

In the present study, an MS genetic burden score is used for the purposes of characterizing the risk of MS cases from the Sardinian population with respect to the MS-associated risk alleles discovered in the Northern European population. The topic of the relevance of established risk factors in diverse populations is often framed in the context of bridging health disparities. In contrast, the focus is on evaluating the risk landscape in an isolated population that offers certain advantages in the search for genetic susceptibility. Therefore, the importance of these experiments lies in guiding the design of future studies in the Sardinian population to target areas of the susceptibility space, as
defined by the genetic architecture of the disease, that are consistent with these observations.

2.2 Experimental design

2.2.1 Dataset description

The dataset consists of samples from two distinct populations. Sardinian families were collected following the identification of the proband through an MS case registry that has operated in Nuoro, Central Sardinia, since 1996. A total of 19 families of genetically connected individuals are included in this current dataset, comprising 820 individuals, 84 of whom are MS affected. The average pedigree includes 51 individuals, with an average of six generations represented in each pedigree (see Chapter 3.2.2 for more information on pedigree structure). The average kinship coefficient (r) between cases is 0.0206 (for reference, second cousins r = 0.0313). Only clinically definite MS affected individuals, diagnosed via the Poser and McDonald criteria, are included in the study.

DNA samples from 177 unrelated MS cases (referred to hereafter as “US cases”) of Northern European descent were ascertained through the University of Miami Health System’s designated MS Center of Excellence, serving the greater Miami area and through patient outreach efforts in the same geographical area. Clinical disease characteristics for the US cases were obtained through a review of the medical records by board-certified neurologists. The clinical characteristics, including age at onset, the distribution of disease course and the female to male ratio for the two populations are shown in Table 2.1. Finally, the 1967 white US control samples were obtained from the Vanderbilt University DNA Biorepository as described in (33).
2.2.2 Genotyping and quality control procedures

DNA samples were available for 326 individuals from these 19 families, including samples from 75 MS cases. DNA samples from 514 individuals were received at the University of Miami, and processed at the Biorepository Core facility within the Center for Genome Technology (CGT) at the John P. Hussman Institute for Human Genomics (HIHG). A total of 88 samples did not pass the quality control (QC) criteria and were therefore excluded from genotyping; 21 of these had low DNA quality (EQ score ≤ 3) and the remaining 67, extracted from buffy coat and cell straw sources, were too low in concentration (< 0.06 μg) to provide the protocol required amount of DNA (1ug) for genotyping. Genotyping on the ImmunoChip was performed on 426 samples by the Genotyping Core facility at CGT. A total of 196, 524 SNPs are tiled on the ImmunoChip, which is a custom Illumina Infinium HD genotyping array. Excluding the MHC, 189 regions that were previously associated with autoimmune disease are deeply interrogated on the chip, which was designed to fine-map established associations and replicate previous results (33). Genotype data generated from the ImmunoChip was used to calculate the genetic burden score attributed to MS-associated alleles.

Three samples showed sex discrepancies in the manifest versus their genetic profiles. Two of them were resolved after investigation of the original records. Relationship check procedures that were performed as part data of sample quality control identified three samples from the pool of 99 unrelated population controls that were genetically deemed to be related and therefore dropped. One of them was distantly related to individuals in family 9, and the other two were siblings of other unrelated control samples. In the case of the siblings, the sample with the highest call rate was kept in the
analysis. The relationship checks were performed using Graphical Representation of Relationships (GRR), a Windows-based application developed by the Abecasis lab within the Center for Statistical Genetics at the University of Michigan (119). One sample had a low genotyping call rate (< 95%), and an additional two samples were dropped because they were determined to be duplicates. A total of 421 Sardinian samples were included in the clean dataset, 94 of which were the unrelated population controls ascertained from the same region. The remaining 252 controls belong in one of the 19 pedigrees, for a total of 346 controls. Sample quality control (QC) procedures for the 177 US cases, 1967 US controls were carried out as previously described (33).

<table>
<thead>
<tr>
<th></th>
<th>Sardinian cases (N = 75)</th>
<th>Caucasian cases (N = 177)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease Course (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relapsing Remitting</td>
<td>0.41</td>
<td>0.72</td>
</tr>
<tr>
<td>Secondary Progressive</td>
<td>0.20</td>
<td>0.16</td>
</tr>
<tr>
<td>Primary Progressive</td>
<td>0.07</td>
<td>0.04</td>
</tr>
<tr>
<td>Unknown</td>
<td>0.32</td>
<td>0.04</td>
</tr>
<tr>
<td>Other (PR and CIS)</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td>Age at onset (years (SE))</td>
<td>28.5 (1.08) range (11,51)</td>
<td>35.6 (0.82) range (12, 60)</td>
</tr>
<tr>
<td>F:M ratio</td>
<td>2.00</td>
<td>3.75</td>
</tr>
<tr>
<td>N of cases per family (median (range))</td>
<td>5 (2,11)</td>
<td>NA</td>
</tr>
<tr>
<td>Symptoms at onset (%)</td>
<td>N=64</td>
<td>N=177</td>
</tr>
<tr>
<td>Sensory</td>
<td>0.56</td>
<td>0.49</td>
</tr>
<tr>
<td>Motor</td>
<td>0.50</td>
<td>0.21</td>
</tr>
<tr>
<td>Cerebellar</td>
<td>0.19</td>
<td>0.07</td>
</tr>
<tr>
<td>Visual</td>
<td>0.27</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Table 2.1. Clinical characteristics of the Sardinian and US cases. PR = Progressive Relapsing, CIS = Clinically Isolated Syndrome, SE = Standard Error.
2.2.3 Statistical analyses

I calculated a log-weighted additive genetic burden score that includes a term for each of the tag SNPs that have been established as associated with MS:

\[ GB = \sum_{i=1}^{s} w_i X_i \]

where \( i \) is the SNP, \( w_i \) is the weight for SNP \( i \) \( (w_i = \log(\text{OR}) \) for the weighted score or \( w_i = 1 \) for the unweighted score) and \( X_i \) is the number of risk alleles at that SNP. A total of 110 SNPs outside the MHC were incorporated; these were either fine-mapped or identified by the largest study of MS genetic susceptibility to date which analyzed data on 80,094 individuals of European ancestry (33). These 110 SNPs tag effects in 103 distinct loci, with genotype data on 109 available within this dataset. Importantly, the ORs that were used to weigh each term are representative of this discovery dataset, and not necessarily of the Sardinian population. An additional four SNPs tag effects in the HLA genes (rs2523822 for HLA-A*02:01, rs1265754 for DRB1*03:01, rs3135388 for DRB1*15:01, and rs3763308 for DRB1*13:01) (51). In total, the burden scores incorporated information on 113 SNPs, 109 non-MHC SNPs and 4 MHC SNPs.

I then calculated the frequency differences using those reported in Supplemental Table 8 of the IMSGC ImmunoChip analysis (33) for the derivation population (based on the UK controls) and the individuals in the dbGaP SardiNIA dataset for the Sardinian population. The means and variances of the risk allele frequencies were not significantly different in the two populations of unaffected individuals (mean\(_{\text{SardiNIA}} = 0.53\), mean\(_{\text{UK controls}} = 0.51\), \( t \)-test \( P \)-value = 0.65, ratio of variances = 0.82, F-test \( P \)-value = 0.31). The
distribution of the risk allele frequency differences between the two populations in these 110 tag SNPs is shown in Figure 2.1(b).

The MS genetic burden score was calculated in four different ways: with and without the inclusion of four SNPs that tag susceptibility at the MHC based on the

![Graph showing distribution of odds ratios and risk allele frequencies](image)

**Figure 2.1** The distribution (a) of the odds ratios for the 113 MS-associated SNPs included in the burden score, and (b) the risk allele frequencies for the 110 non-MHC MS-associated single nucleotide polymorphisms in unaffected population samples. RAF = Risk Allele Frequency.

Northern European haplotypes, and with and without the use of the log of the OR as a weight for each term. Figure 2.1(a) shows the distribution of the ORs. The un-weighted genetic burden score represents a simple count of MS risk alleles carried by an individual. ORs and 95% confidence intervals (CI) were obtained by logistic regression via generalized estimating equations (GEE) as implemented by PROC GENMOD in SAS software, of the SAS System for UNIX. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA. Individuals were clustered by family, and an independence matrix was used as the
The initial covariance structure. The covariance matrix was iteratively updated using the observed correlations between relatives.

Two outcomes were modeled: disease status and population membership. Model 1 examines the hypothesis that the established MS-associations can differentiate Sardinian cases from Sardinian control. Model 2 targets population differences by testing the hypothesis that Sardinian cases carry a different burden than US cases. The same model was used to test the hypothesis that Sardinian controls carry a different burden than US controls (model 3). The predictive power of the models was assessed with the AUC of ROC curves. Statistical significance was assessed at $\alpha = 0.05$ for the following models:

$$\text{Logit}(\text{MS}) = \beta_0 + \beta_1 \text{burden score} + \beta_2 \text{sex} \quad \text{Model 1}$$

$$\text{Logit}(\text{population}) = \beta_0 + \beta_1 \text{burden score} + \beta_2 \text{sex} \quad \text{Models 2 and 3}$$

In order to further examine whether the combination of MS risk alleles driving the genetic burden is different in Sardinian cases compared to US cases, I carried out logistic regression analyses using single SNPs as the predictors and used the top hits in constructing multilocus genotypes. The single marker tests were also corrected for the family structure using GEE, modeling either population membership or disease status as the binary outcome. In order to account for multiple testing, significance was assessed at a $P$-value of $4E-4$, the Bonferroni corrected threshold for 113 tests. I then used the 4 most highly significant markers from the single marker tests to construct multilocus genotypes (Table 2.2). These reflect the combinations of risk alleles that are carried by each individual, and not the phased haplotypes. I used McNemar’s test (implemented in the R function mcnemar.test) as a non-parametric assessment of the difference in the
proportions of US cases and Sardinian cases that carry all 4 of the most significant SNPs in the homozygous risk allele state. Because none of the 113 SNPs survive the multiple testing correction in the GEE single marker tests with disease status as the outcome, I used the 4 most significant markers ($P$-value $\leq 0.05$) in constructing multilocus genotypes for Sardinian cases and controls. I then assessed their association with case status, examined the distribution of the combinations of risk alleles in Sardinian cases and controls, and tested the difference in the proportion of Sardinian cases and controls who are homozygous for the risk allele at all 4 SNPs using McNemar’s test for correlated proportions.

<table>
<thead>
<tr>
<th>Chr.</th>
<th>rsID</th>
<th>P-value</th>
<th>OR (95% CI)</th>
<th>Risk allele</th>
<th>$F_{USC}$</th>
<th>$F_{SC}$</th>
<th>$F_{SP}$</th>
<th>$F_{CEU}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>rs11719975</td>
<td>0.0001</td>
<td>3.22 (1.78, 5.80)</td>
<td>C</td>
<td>0.26</td>
<td>0.09</td>
<td>0.09</td>
<td>0.25</td>
</tr>
<tr>
<td>6</td>
<td>rs2523822</td>
<td>0.0001</td>
<td>0.39 (0.24, 0.63)</td>
<td>A</td>
<td>0.76</td>
<td>0.56</td>
<td>0.55</td>
<td>0.64</td>
</tr>
<tr>
<td>6</td>
<td>rs72928038</td>
<td>0.0001</td>
<td>0.42 (0.27, 0.65)</td>
<td>G</td>
<td>0.83</td>
<td>0.71</td>
<td>0.81</td>
<td>0.86</td>
</tr>
<tr>
<td>12</td>
<td>rs12296430</td>
<td>0.0004</td>
<td>0.41 (0.25, 0.67)</td>
<td>G</td>
<td>0.79</td>
<td>0.62</td>
<td>0.70</td>
<td>0.84</td>
</tr>
</tbody>
</table>

**Table 2.2** The four SNPs that were used in constructing the multilocus genotypes for US and Sardinian cases. Note that risk refers to the risk of being a US case, and not the risk of disease. $F$ = frequency, the subscripts are defined as follows: USC = US cases, SC = Sardinian cases, SP = Sardinian population (retrieved from the SardiNIA Medical Resequencing Project on the dbGaP repository), CEU = Centre d'Etude du Polymorphisme Humain (retrieved from dbSNP with respect to the HapMap CEU dataset).

Finally, in order to assess the effect of the degree of clustering in a family on the accumulation of genetic burden, I built linear regression models where family-wise mean burden scores were used to predict the number of cases in that family. Family size was used as a covariate.

Number of cases = $\beta_0 + \beta_1$ mean burden score + $\beta_2$ family size  

**Model 4**
2.3 Sardinian cases carry MS-associated alleles discovered in Northern Europeans

The unweighted burden score that includes the Northern European HLA tag SNPs is significantly higher in Sardinian cases than it is in any of the other three groups in this dataset (Figure 2.2). The difference between Sardinian and US cases is eliminated when weights are applied. This is likely due to the known fact that different HLA haplotypes capture MS risk in Sardinia and that the HLA effects are the most heavily weighted, so that when US cases carry those risk alleles their burden is increased to levels that compare to the burden in Sardinian cases that is primarily attributable to effects other than HLA.

Under all comparisons, Sardinian cases carry significantly higher burden than Sardinian controls, indicating the discriminatory ability of the genetic burden score to differentiate cases from controls in Sardinia. Sardinian controls are also more heavily loaded than US controls (Figures 2.2 and 2.3). This observation may reflect their membership in multiplex families or it could be a characteristic of the Sardinian population as a whole.

Although it is well established that the HLA haplotypes that drive MS risk are different in the Sardinian population, the results are not driven by these differences. As expected, the genetic burden for Sardinian cases does not increase significantly when HLA alleles are included, reflecting the relative absence of Northern European HLA alleles in Sardinians. In addition, when HLA effects are ignored Sardinian controls and US cases have similar burden for the remaining risk alleles (Figure 2.3). This observation points to the missing heritability and confirms that cumulative effect of the remaining known alleles is not sufficient to explain case status, a hypothesis that is further
**Figure 2.2** The distribution of the MHC-inclusive genetic burden score across the various groups (a) using the ORs as weights, and (b) un-weighted. The p-values shown correspond to the burden score coefficient from the Generalized Estimating Equations (GEE) model between the groups flanking it.

**Figure 2.3** The distribution of the genetic burden score that does not include MHC effects across the various groups (a) using the ORs as weights, and (b) un-weighted. The p-values shown correspond to the burden score coefficient from the GEE model between the groups flanking it.
The ROC curves are shown in Figure 2.4. The assessment of the predictive ability of models 2 and 3 reveal limited capability to distinguish between cases and controls as well as between cases from the two populations, using the currently available genetic burden score (AUROC = 0.56 to 0.68).

Multiplex families are known to carry a higher MS burden, using previous versions of the score (112). There is no significant relationship between the number of MS cases in a family and the mean genetic burden score in that family, a result that supports the hypothesis that population membership and not family structure accounts for these observations (weighted burden score P-value = 0.96, unweighted P-value = 0.66).
and Figure 2.5). As expected, family size is a significant predictor of the extent of clustering (weighted genetic burden score $P$-value = 0.002, OR (95% CI) = 1.21 (1.09, 1.37), unweighted $P$-value = 0.003, OR (95% CI) = 1.22 (1.08, 1.36)). Alternatively, the accumulation of risk in multiplex families may not be reflected in the extent of the clustering. In that case, clustering itself and not the extent of clustering may be the variable that has an effect in the accumulation of burden. On one hand, it may seem unlikely that clustering as a binary predictor would have an effect given the quantitative way in which polygenic burden accumulates. On the other hand, crossing the susceptibility threshold to disease is a binary outcome, so it is conceivable that above a critical value of burden, multiple cases would cluster in a family given enough opportunity to do so (i.e. enough meioses).

![Figure 2.5](image.png)

**Figure 2.5** There is no significant relationship between the extent of clustering of cases in a family and the mean (a) weighted or (b) unweighted burden score. The multiple regression lines are plotted for the mean family size (21.8 individuals).
Using single marker tests in a GEE framework, four markers stand out as surviving the multiple testing threshold (Table 2.2), one of which tags the HLA-A*02:01 risk haplotype. The reduced diversity seen in the Sardinian samples over all of the observed multilocus genotypes is likely affected by the relatedness in this sample (Figure 2.6).

**Figure 2.6** The distribution of multilocus genotypes in (a) cases from the two populations, and (b) Sardinian cases and controls. The numbers represent the genotype at each SNP such that a 2 represents the homozygotes for the major allele, a 3 represents heterozygotes, and a 4 the homozygotes for the minor allele. Note that the Sardinian cases belong to 19 large pedigrees and that the 4 SNPs that make up the multilocus genotypes are different in the two panels.
In addition, I performed a non-parametric test for unequal proportions which does not assume independence between observations to compare the proportions of cases with homozygous genotypes for the risk allele at all four SNPs (denoted as 4444). Significantly more Sardinian cases carry the 4444 multilocus genotype than US cases (McNemar’s chi-squared test $P$-value = 0.01). Interestingly, the multilocus genotypes that represent homozygotes for the non-risk allele at more than one SNP are not observed. However, this observation is possibly biased by the a priori selection of the markers with the most highly significant frequency differences between Sardinian and US cases.

A similar analysis comparing Sardinian cases to Sardinian controls also reveals differences in the overall distributions of the multilocus genotypes between these two groups (Wilcoxon signed-rank test $P$-value = 4.41E-13). In addition, significantly more Sardinian controls are homozygous for the risk allele at all of the SNPs used in the multilocus genotype (McNemar’s chi-squared test $P$-value = 0.008). This observation is contrary to what would be expected if the markers used in the multilocus genotypes tagged disease status and it likely points to the relatively weak effect sizes of the risk alleles that are associated with MS.

The substantial levels of noise indicated by the proportion of controls that carry the risk alleles in homozygous or heterozygous states (Figure 2.6) may indicate that the risk alleles that effectively distinguish Sardinian cases from controls belong to a different set – one that has limited overlap with the set of MS-associated markers in individuals of Northern European ancestry. This observation motivates the search for additional MS loci in this population (Chapters 3 and 4).
3

Allelic Sharing in Sardinian Affected Relative Pairs with Multiple Sclerosis

3.1 Contextual framing of the linkage scan

3.1.1 Linkage analysis of complex traits

Linkage analysis tracks the co-segregation of marker alleles with the disease phenotype of interest. It identifies the marker locus that is in linkage disequilibrium (LD) with the unknown disease locus by examining the frequency of recombination between the two loci. Parametric linkage analysis requires the specification of a number of parameters, including the mode of inheritance. Therefore, it is best suited to Mendelian phenotypes.

Non-parametric linkage (NPL) approaches have been developed for the genetic analysis of complex traits that are heritable to a certain degree, but which show no clear patterns of segregation. Heritability can be estimated in a number of ways, for example
by comparing concordance between monozygotic and dizygotic twins, as well as using siblings that are reared apart to account for shared environment. As mentioned in Chapter 1, one measure of familial aggregation is $\lambda_S$ (120). It measures the increased risk in siblings of affected individuals relative to the general population. Aside from relatively modest heritability, complex traits often present special challenges such as the large number of loci underlying the trait, the presence of phenocopies (non-genetic cases), genetic heterogeneity, and reduced penetrance. Non-parametric methods have the advantage of not having to specify an inheritance model; instead, they are based on allele sharing. Initially, Penrose proposed the sib-pair method, where linkage between two traits is investigated by examining the relative frequencies of pairs of siblings who are concordant or discordant for both traits (121). The sib-pair method was also restricted to pairs of affected siblings (ASP). ASP methods measure the frequency of affected offspring that share marker alleles identical-by-descent (IBD). For an example of how IBD sharing is measured, consider a polymorphic marker for which both parents are heterozygous with a total of four different marker alleles. Their affected offspring may then share 0, 1 or 2 pairs of alleles IBD according to the proportions 1:2:1, respectively, under the null hypothesis of no linkage. Deviations from this expectation can be measured with a number of different allele sharing statistics. These statistics bypass the reduced penetrance issue by only considering siblings who are known to have received the disease allele. Linkage is indicated by a deviation towards higher numbers of alleles shared IBD compared to the levels of sharing that are expected based on the degree of relationship between the siblings.
Methods that use distant relative pairs rely more heavily on alleles that are shared identical-by-state (IBS) \((122, 123)\). IBS sharing examines whether alleles are the same, ignoring whether or not they are copies of a single ancestral allele. The expectation of sharing is adjusted with the appropriate kinship coefficient as well as marker allele frequencies. Risch et al. as well as Bishop and Williamson examined the power of different pairs of affected relatives under a range of risk ratio \((\lambda)\) values, using IBD and IBS methods respectively \((124, 125)\). Their results show that for diseases with higher \(\lambda\) values more distant relative pairs offer greater power than sibling pairs. For example, for a disease with \(\lambda = 5\), examining linkage with a fully polymorphic marker at \(\theta = 0.10\) \((\theta\) represents the recombination fraction between the marker and the disease loci), power is notably higher for grandparent-grandchild pairs (98%) than it is for sibling pairs (30%).

Kruglyak et al. proposed and implemented an IBD sharing statistic which performs multipoint calculations, enabling the use of all of the descent information from a pedigree \((126)\). Kong and Cox then proposed an improvement to the method, based on a likelihood ratio test, to more accurately calculate the variance of the linkage statistic and provide improved evaluations of statistical significance \((127)\).

Finally, an important parameter in any linkage study is information content. This is defined as the amount of linkage information that can be extracted from a given pedigree. Its importance is demonstrated by the linear increasing relationship between the expected LOD score and information content at a region, and it is the reason that maps of biallelic markers are able to extract less information than those based on more polymorphic markers such as microsatellites.
3.1.2 Past linkage scans in multiple sclerosis

The first linkage scans in multiple sclerosis were done in Finnish families, prompted by the remarkably high prevalence of MS in the Finnish population isolate as well as the fact that in some areas of Finland, over 30% of MS cases are familial. Initial reports implicated HLA DQA1 (relative risk (RR) = 3.8) and a tetranucleotide repeat in the myelin basic protein gene (MBP) on chromosome 18q (RR = 3.3) (128). A subsequent whole genome scan performed in 100 sibling pairs of European descent in Canada reported five regions (on chromosomes 2, 3, 5, 11, and the X) with maximum linkage scores > 1 (129). An additional two whole genome linkage scans were published in the same issue of Nature Genetics. These scans report similar results with respect to the magnitude of the effects. Haines et al. reported 19 regions of interest with LOD scores around 1 using 438 affecteds in both sibling pairs and/or extended families with European ancestry (130). Sawcer et al. reported one principal region of interest other than the MHC, on chromosome 17q22, and another four provisional regions (on chromosomes 1, 5, 7, and 14) again with the use of affected sibling pairs (131). Although the 17q22 region was replicated the following year in a dataset of 16 Finnish families (132), in general, there was very little overlap in the regions identified in the initial linkage scans. In total, 59 regions were identified as potentially interesting by the four genome-wide linkage studies performed between 1993 and 1997, but only 15 of those were detected by more than one group.

Another region on chromosome 19q13 was replicated by some of the studies and further explored in a larger US dataset of 150 affected sibling pairs and 84 affected relative pairs (133). In this study, the two-point LOD score for 19q13 increased from 1.47
to 2.01. An association that maps to the 19q13 was observed in the ImmunoChip experiment (rs8107548, $P$-value = 1.98E-6, OR = 1.09 (1.05, 1.13)) (33).

In a linkage study of 40 families (37 ASP) from continental Italy, another one of the 19 regions reported by Haines et al. was replicated, residing close to the centromere of chromosome 10 (10cen) (134). The same study reported another eight regions with non-parametric linkage LOD scores above 0.7 (1q42, 1q44, 2q36, 5q33, 6pter, 6q22, 10cen, and 15q21). A region on 2p22 was implicated in a separate Italian dataset using a family based association test (135). The differences in the results from the mainland Italian population are not surprising if known differences in the allele frequencies are taken into account; for example, the frequency of HLA DRB1*15:01 in MS cases of Northern European ancestry is approximately 60%, compared to 30% in Italian cases.

By 2003, whole genome linkage screens were completed in nine different populations (American, British, Canadian, Australian, Finnish, Italian, Sardinian, Turkish and Scandinavian). A meta-analysis of all of the nine datasets identified, aside from the MHC at 6p21, six regions that may harbor ubiquitous susceptibility loci. These are 2p14, 10p15, 11pter, 16p13, 17q21 and 22q13 (136).

Prompted by the variable information content across linkage studies for the regions that showed suggestive linkage, Akesson et al. (137) attempted to refine the region on the long arm of chromosome 10 by increasing the density of the markers in a pooled dataset that consisted of the previously scanned UK, Sardinian, Italian and Nordic (138) affected sibling pairs. Although little changed in the individual country datasets, the combined results increased the peak maximum LOD score (MLS) to 2.5 in the 10p15 region, which includes two interleukin receptors (IL15-RA and IL2-RA). This success is
probably attributable to the increased power of the combined dataset, rather than the increased density of the markers.

In 2005, the International MS Genetic Consortium published a meta-analysis of 730 families with Northern European ancestry (Australian, Scandinavian, British and American) with significant improvements in the genotyping rate and average information content of the linkage map that they used relative to the 2003 meta-analysis (139). An additional strength of this study was that the sample size allowed stratified analyses which can reveal additional loci in the presence of locus heterogeneity. A stratified analysis generated the highest LOD score seen in the HLA DRB1*15:01 negative group on 2p25 (LOD = 1.61), a region that showed no evidence of linkage in the combined dataset. The highest LOD score in the HLA DRB1*15:01 positive group, outside of the MHC, was 2.18 on 5q33. This study also reports an ordered subset analysis (OSA) in the same dataset using the HLA region on chromosome 6 as the covariate. Ranking families according to a covariate – clinical characteristics are also often used – can be useful in generating subset-specific LOD scores and accounting for known effects. The OSA revealed four potentially interesting regions (12q24, 19p13, 1q43, and 7q21), although only 19q13 survived the multiple testing correction. The regions on 17q23 and 5q33 also show the highest MLS scores in the combined analysis. Although no regions outside the MHC jump out with unequivocal evidence for linkage, 63.9% of the genome excluding the MHC has a positive LOD score, relative to the expectation of 50%.

3.1.3 Mapping MS regions in Sardinia

Given that the disease allele frequency is an important determinant in the power of a linkage scan, and the expectation that this will vary between populations, there is good
reason to carry out linkage studies in a population where the disease allele frequency is suspected to be high. The differences in the HLA haplotypes that are associated with MS in the Sardinian and Northern European populations are direct observations of how the demographic history of a population can affect the frequencies of disease-relevant alleles (101).

In Sardinian linkage scans, the majority of the studies rely on the affected sibling pairs design. A candidate gene study in 28 Sardinian families (24 of them were sibling pairs) reported low linkage scores using two non-parametric methods. However, they did denote a subset of the candidate regions as promising (2p11, 5p12-14, 7p15 and 22q13) and noted that one of them (7p15) showed homology to the rat region harboring the experimental autoimmune encephalomyelitis locus (EAE) (96). EAE is one of the animal models of MS. It can be induced by immunization with antigens such as myelin basic protein (MBP) and its sub-peptides, spinal cord homogenates, and proteolipid protein (PLP) (reviewed by (140)). The extent to which overlap in the susceptibility regions between MS and EAE can be used as evidence in support of weak linkage results is questionable, as the use of EAE as a model for MS has received criticism (for example (141, 142)).

A whole genome linkage scan in 49 Sardinian families (43 of them were sibling pairs) revealed a distinct set of three regions (1q31, 10q23 and 11p15) of suggestive linkage in this population (95). The 10q23 region was one of the 19 regions reported by Haines et al. (130), and a gene in that region (hematopoietically expressed homeobox HHEX) has also been replicated by the Wellcome Trust Case Control Consortium II association study (rs7923837, $P$-value = 4.9E-9, OR = 1.1 (1.08, 1.11)) (51), and by the
ImmunoChip experiment performed by the International MS Genetics Consortium (rs7923837, \(P\)-value = 4.58E-8, OR = 1.11 (1.07, 1.14)) (33).

The largest linkage study using Sardinian MS cases genotyped 593 microsatellites in 175 families with a total of 413 affected individuals. This scan revealed three regions with a Kong and Cox LOD* score \(\geq 1\) (1q42.2, 18p11.21, and 20p12.3). This study also performed a combined scan for autoimmunity, taking advantage of the high prevalence of both Type I Diabetes and MS in Sardinia. The combined scan added another three loci to this list of chromosomal regions that potentially harbor MS susceptibility alleles (6q26, 10q21.1, and 22q11.22) (143).

Aside from linkage studies, trio families (proband and samples from both parents) have also been used in Sardinian association scans for multiple sclerosis genes. Coraddu et al., generated data on 3512 microsatellite markers in 235 trios and suggested a list of five regions (2q36, 6p21, 6p25, 7p12, 16p12) which were nominally significant with a \(\chi^2\) test in which control allele counts were adjusted using the parents, and which were also nominally significant in their Sardinian case-control dataset (144)(95).

3.1.4 Motivation

The consistent picture that arises in linkage studies of MS, is that of excess sharing compared to the levels expected by chance alone, but no unequivocal regions of linkage within or across populations. The regions of suggestive linkage that show up in one or a few underpowered studies reflect either type I errors (false positive results) or genuine susceptibility loci with modest effects. Nevertheless, the absence of unequivocal results is in line with the hypothesis that MS susceptibility is characterized by moderate to small effects that are below the sensitivity of the detection tools used so far. The results of the
genetic burden analyses in Sardinia (see Chapter 1) further support a polygenic architecture for MS in this population that may not be that different than in Northern Europeans – at least for common disease alleles. In addition, linkage studies in more distant relative pairs such as those present in this dataset may offer additional power in identifying regions of sharing with increased confidence, especially for diseases with large $\lambda$ values and for small $\theta$ (125).

Therefore, I performed a linkage scan in distant affected relative pairs from 15 multiplex Sardinian pedigrees with the intention of identifying regions of sharing that can be used to limit the search space in the whole exome sequencing studies (see Chapter 4) that interrogate the exome for both common and rare variation.

3.2 Methodological approach

3.2.1 Selection of the linkage map

The markers used to build the linkage map originate from the ImmunoChip SNP set. The ImmunoChip is a custom genotyping array that uses Illumina Infinium HD technology. It was built for the purposes of fine-mapping and performing deep replication of association results from the GWAS of a number of autoimmune diseases (145). The final set of markers included on the chip is comprised by 192402 autosomal, 1595 X-linked, 1735 Y-linked, 791 pseudoautosomal and one mitochondrial marker. The intention to fine-map previous genome-wide significant results across multiple immune-mediated diseases resulted in the inclusion 189 non-MHC regions (only two of them were on the X-chromosome) with a substantial density of markers compared to other regions of the genome. However, the uneven distribution of the original set of markers has not impacted the ability of the linkage map to uniformly interrogate the whole genome.
Genotype data on the set of 192402 autosomal markers was available on 421 Sardinian individuals. In order to connect the individuals in the pedigrees I added 509 un-genotyped family members with missing genotypes. Quality control was performed as described in Figure 3.1. Two individuals with duplicate samples were removed so that the one with the highest genotyping rate was kept in the dataset. SNPs with Mendelian inheritance errors in multiple individuals may indicate failed assays. A set of 111,962 Mendelian errors were detected in 71 out of the 386 trios available in this dataset, representing a smaller set of 8,127 SNPs which were dropped from analysis. Pedstats was used to verify that no Mendelian errors remained in the dataset. The genotyping rate statistics, allele frequencies and Hardy-Weinberg equilibrium test \( P \)-values were generated using PLINK v1.07 (146). In order to properly calculate missingness statistics the sets of un-genotyped relatives were specified as obligatory missing using the --oblig-missing and --oblig-clusters options in PLINK. A final set of 317 SNPs were flagged as having an unlikely configuration in the families using the error detection tool provided in the multipoint engine for rapid likelihood inference (MERLIN) package v. 1.1.2 (3). This tool examines stretches of shared genotypes by all the relatives in a pedigree to calculate the likelihood that a genotype is the result of a genotyping error. For example, a double recombination in a very small stretch of DNA will be flagged, as it may indicate genotyping error that would lead to false inference. These were also dropped from the dataset. The final SNP set used in the selection of the linkage map consisted of 99,724 SNPs.

The masel algorithm was used to select the linkage map (147). Figure 3.2 shows the steps: ranks are assigned to each SNP in the source set with each iteration. The source
set is defined as the set of SNPs from which the linkage map is selected. Each SNP is ranked according to their MAF, the number of SNPs that would be removed if that SNP was selected, and the variance of the distance between the SNPs in the linkage set that would result if that SNP was selected. The masel algorithm selected a set of 11,605 SNPs that comprise the linkage map.

Figure 3.1 A diagram showing the quality control steps that resulted in the SNP-set from which the linkage map was created (99,724 SNPs). The number of SNPs shown is the number that remains after all markers that do not satisfy each quality control criterion are dropped. HWE = Hardy-Weinberg Equilibrium.

Figure 3.2 Diagrammatic representation of the algorithm implemented in MASEL and used to select the final linkage SNP set (11,605 SNPs).
Finally, biallelic marker maps come with a cost in information content (for a quantification of this cost see (148)). Calculations of information content are informed by both the density of the linkage map and the level of polymorphism of the markers, as a measure of the fraction of inheritance information that is extracted by the map relative to a map of ideal density and polymorphism levels. The linkage information that can be extracted from the pedigrees along each chromosome is shown in Figure 3.3.

**Figure 3.3** The average information content for each chromosome, calculated across all families. The x-axis represents genetic distance measured in centiMorgans (cM).
3.2.2 The multiplex Sardinian pedigrees

A case register was established in 1995 in the Nuoro province of Sardinia, a mountainous region in the north east of the island. MS cases have been diagnosed as clinically definite
using the Poser and/or McDonald criteria (21, 149). The register captures clinical history information including MS diagnosis and course, age at onset, symptoms, comorbid diseases, genealogical information (extending to grandparents and first cousins) and blood samples. At the time of this study the register contained 706 patients, and 20,479 of their relatives and 112 unrelated population controls; however, DNA was available on a subset of these, consisting of 75 of the 90 cases, and 252 unaffected individuals. Additional investigation of the relationships and common ancestors allowed the linking of many of the smaller families into large multiplex pedigrees. This was especially true for families from the same village or from nearby villages. Standard relationship check procedures that rely on the proportion of the genotypes that are identical-by-state (IBS) between all pairs of individuals in a dataset revealed relationships that allowed the linking of pedigrees 23 and 60 into an even larger multiplex pedigree. Pedstats (Center for Statistical Genetics, University of Michigan) was used to check the integrity of the pedigrees and provide summary information on the pedigree structure (Table 3.2). Initially, disjoint groups were detected in four families (9, 56 and 100); these were due to subpedigrees that were connected by marriage and not sharing any known genetic relationship. Therefore, they were separated into those pedigrees; pedigree 9 was split into 901, 902, and 903, pedigree 56 was split into 5601 and 5602, and pedigree 100 was split into 10001 and 10002. This generated a total of 19 pedigrees with genetically connected individuals. Two of these pedigrees (903 and 10002) no longer contained any pairs of affected relatives after the split and were therefore ignored in the calculations of allelic sharing.
In order to perform exact likelihood calculations for a large number of markers in the linkage analysis, I had to limit the pedigree complexity to 24 bits. A python script was written to calculate the complexity of each of the 19 pedigrees according to the formula $2n-f$ where $n =$ number of non-founders and $f =$ number of founders. This was done using PedCut, an algorithmic approach to sub-dividing pedigrees that identifies subpedigrees which include the maximum number of individuals of interest who share a common ancestor (150). The affected individuals were designated as the individuals of interest in this procedure. The subpedigrees generated using the PedCut algorithm were manually edited to include genotyped, unaffected siblings which were left out since these could provide additional haplotype information, in the cases where the pedigree bit size allowed. Additionally, pedigree 3 was subdivided by PedCut into two subpedigrees, one with two (subpedigree 32) and one with three affecteds (subpedigree 31) – to the exclusion of one of the 6 affected individuals. Manual examination determined that the excluded individual could be included if one of the other affecteds in subpedigree 32 was reassigned to subpedigree 31, so that 31 would have a total of 4 affecteds, and 32 would have two affecteds. This increased the distance between the new pair of affecteds in subpedigree 32, reducing the probability of finding shared regions. However, the advantage to the increased distance is that if shared regions are identified, they would contribute more to the LOD* score (125). An affected individual was also excluded from both subpedigrees created for pedigree 5; however, it was not possible to manually rearrange the divisions so that this individual would be included without surpassing the maximum bit size. The extent of clustering of cases in the 19 original pedigrees is shown
in Table 3.1. After PedCut, a total of 505 individuals in 24 subpedigrees were analyzed (Table 3.2).

<table>
<thead>
<tr>
<th>Table 3.1 The Sardinian dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>21</td>
</tr>
<tr>
<td>26</td>
</tr>
<tr>
<td>44</td>
</tr>
<tr>
<td>45</td>
</tr>
<tr>
<td>58</td>
</tr>
<tr>
<td>61</td>
</tr>
<tr>
<td>81</td>
</tr>
<tr>
<td>901</td>
</tr>
<tr>
<td>902</td>
</tr>
<tr>
<td>903</td>
</tr>
<tr>
<td>5601</td>
</tr>
<tr>
<td>5602</td>
</tr>
<tr>
<td>10001</td>
</tr>
<tr>
<td>10002</td>
</tr>
<tr>
<td>23_60</td>
</tr>
<tr>
<td>Population</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3.2 Descriptive statistics on pedigree* structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individuals (N)</td>
</tr>
<tr>
<td>Founders (N)</td>
</tr>
<tr>
<td>Gender (N)</td>
</tr>
<tr>
<td>Families (N)</td>
</tr>
<tr>
<td>Family size (mean)</td>
</tr>
<tr>
<td>Generations (mean)</td>
</tr>
</tbody>
</table>

*These represent the family set that was used in the linkage analysis, after PedCut and manual editing separated the 19 pedigrees into 24 subpedigrees of < 24 bits.
3.3 Non-parametric linkage results

Non-parametric multipoint linkage analysis was performed on the combined dataset of 24 subpedigrees for each chromosome to allow the calculation of dataset-wide statistics. I calculated both the Whittemore and Halpern NPL statistics $S_{pairs}$ and $S_{all}$, using the options --pairs and --npl provided by MERLIN v1.1.2. In addition, per family statistics were calculated to identify the contribution of each family to the linkage signal (--perFamily). The Kong and Cox linear model (127) was used since it is designed to identify small increases in allele sharing among affecteds that are distributed across multiple pedigrees, in line with the multiple small effects expected for a complex disease such as MS. The results of the $S_{pairs}$ statistic are shown in Figure 3.4. In summary, six regions reached a LOD* score ≥ 1 with the $S_{pairs}$ statistic: two in 12p13, one on 18q21-22, one on 4q28, one on 6q15-16 and just one marker at 7q11. The regions on chromosomes 6, 7 and 18 were also detected with the $S_{all}$ statistic. Another two regions had LOD* > 1 using the $S_{all}$ statistic; these were on 11p11 and 7p12. Multiple families are contributing to the excess sharing observed in all regions where LOD* ≥ 1 (Figure 3.5).

The high prevalence of comorbidity of MS and other autoimmune phenotypes in the Sardinian population may weaken the linkage signal at general autoimmunity loci, if unaffected individuals who manifest a different autoimmune phenotype are coded as unaffected in the analysis.

Figure 3.4 Multipoint non-parametric linkage analysis performed using MERLIN (3). Each graph represents one chromosome; the x-axis denotes the genetic length of the chromosome in centiMorgans (cM), and the y-axis denotes the Kong and Cox linear model LOD* score on a scale of 0 to 2. The MHC extends between 53 and 55.5cM on chromosome 6. The red dots at the top represent the physical positions of the current MS-associations.
Figure 3.5 Per-family contributions to the linkage results. Only the chromosomes reaching LOD* ≥ 1 are plotted. Note that the LOD* scale extends from 0 to 0.8. The legend applies to all chromosomes. The 24 subpedigrees are annotated in the legend with a numeric suffix indicating the subpedigree to which they belong. For example, family 21 was split into subpedigrees 211 and 212, and so on.
3.4 Association hits in the Sardinian linkage peaks

In order to assess the overlap between the established MS-associations and excess sharing in this Sardinian dataset of affected relatives I identified the marker in the linkage map that is physically closest to the reported associated SNP ([33], Supplemental Table 8). The markers for which $S_{pairs} \geq 0.5$ are shown in Table 3.3.

<table>
<thead>
<tr>
<th>Chr.</th>
<th>hg19 position</th>
<th>Risk allele frequency</th>
<th>Gene</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>90976768</td>
<td>0.17</td>
<td>BACH2</td>
<td>intronic</td>
</tr>
<tr>
<td>12</td>
<td>6440009</td>
<td>0.4</td>
<td>TNFRSF1A</td>
<td>intronic</td>
</tr>
<tr>
<td>12</td>
<td>6503500</td>
<td>0.19</td>
<td>LTBR (dist=2768), CD27-AS1 (dist=44667)</td>
<td>intergenic</td>
</tr>
<tr>
<td>12</td>
<td>9905690</td>
<td>0.36</td>
<td>CD69</td>
<td>UTR3</td>
</tr>
<tr>
<td>18</td>
<td>56384192</td>
<td>0.77</td>
<td>MALT1</td>
<td>intronic</td>
</tr>
</tbody>
</table>

3.5 Fine-mapping of the linkage peaks using a family-based association test

In order to better define which gene regions are generating the signal seen in the linkage peaks I used a family-based association test that is robust to the presence of linkage as implemented in the Combined Association in the Presence of Linkage test (CAPL) ([151]). Correct inference of parental genotypes depends only on Mendelian probabilities and population allele frequencies when the marker and susceptibility loci are unlinked. However, in the presence of linkage the correlation between the marker genotypes of multiple affected siblings due to the presence of linkage can bias the inference of parental genotypes and lead to inflated type I error of the association tests (i.e. generate false positives).

The unrelated Sardinian population controls (as described in Chapter 1) were kept in the dataset for two reasons: because the information they provide can be used in the
CAPL framework, and because estimation of population allele frequencies is more accurate with a larger sample size. Five of the unrelated population controls were detected as outliers by CAPL and removed from the analysis. The families were defined in their original 19 pedigrees, in order to avoid underestimating the variance. CAPL extracts the most informative nuclear family from extended pedigrees based on internally calculated IBD measures. Given the small number of pedigrees included in this analysis, there is a danger of inflated type I errors; therefore the results should be interpreted with caution.

The regions were defined using the 1-LOD*-down support intervals around the maximum LOD* scores (152) (Table 3.4). A total of 4360 markers were extracted from the clean ImmunoChip dataset (99, 724 SNPs) within the support intervals of the 5 regions that achieved LOD* > 1 (chromosomes 4, 6, 7, 12, and 18). The variance of the test statistic depends on the allele frequency of the marker ($r^2 = -0.27, P$-value < 0.001), to the extent that allele frequency explains 7% of the variance in the test statistic variance.

<table>
<thead>
<tr>
<th>Region</th>
<th>Chr</th>
<th>ImmunoChip ID</th>
<th>cM position</th>
<th>hg19 position</th>
<th>LOD*</th>
<th>Max. LOD*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>rs1346715</td>
<td>131</td>
<td>126535676</td>
<td>0.390</td>
<td>1.363</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>rs1026047</td>
<td>143</td>
<td>140490091</td>
<td>0.295</td>
<td>1.363</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>rs755927</td>
<td>94</td>
<td>78925483</td>
<td>0.420</td>
<td>1.419</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>imm_6_99481481</td>
<td>108</td>
<td>99374760</td>
<td>0.413</td>
<td>1.419</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>rs745580</td>
<td>41</td>
<td>26008052</td>
<td>0.065</td>
<td>1.005</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>rs1468358</td>
<td>110</td>
<td>100853173</td>
<td>0.002</td>
<td>1.005</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>rs2535429</td>
<td>1</td>
<td>690473</td>
<td>0.228</td>
<td>1.202</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>rs11055338</td>
<td>33</td>
<td>13441584</td>
<td>0.105</td>
<td>1.202</td>
</tr>
<tr>
<td>5</td>
<td>18</td>
<td>rs17444197</td>
<td>79</td>
<td>51283376</td>
<td>0.182</td>
<td>1.157</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>imm_18_65719481</td>
<td>104</td>
<td>67568501</td>
<td>0.111</td>
<td>1.157</td>
</tr>
</tbody>
</table>
in this dataset. This is alleviated by considering only results with variance of the test statistic > 5 for which the correlation between allele frequency and variance of the test statistic is -0.04, which limited the results to 1716 markers. The mean allele frequency of these 1716 markers for the allele designated by CAPL as allele 1 is 0.54 (SD = 1.99, range 0.12 to 0.92).

Four of the 5 regions contain nominally significant markers (P-value ≤ 0.05); the region on chromosome 4 generated no nominally significant associations, although it is worth noting that it was relatively narrow and sparsely covered (Figure 3.6). The most significant SNP in the linkage regions on chromosomes 6 and 12 are intergenic, while the most significant SNPs on the chromosome 7 and 18 regions fall inside the JAZF1 and CD226 genes, respectively.

Coverage is biased towards loci that were intended to be fine-mapped by the ImmunoChip experiment, as well as SNPs with allele frequencies that allow the accurate calculation of the CAPL test statistic; therefore it is possible that potential MS susceptibility loci inside the linkage peaks are located in sub-regions other than the ones highlighted here. The results of the fine-mapping attempt are further used and discussed in Chapter 4, where they are used to prioritize variants in the subsequent exome sequencing studies.

In summary, the regions of suggestive sharing between affected relative pairs in these Sardinian multiplex pedigrees are implicating regions of the genome with genes that are strong biological candidates. Furthermore, the absence of regions of significantly excess sharing (defined by a LOD* > 3) suggests that MS-susceptibility in these
multiplex families is not driven by one or a few yet unidentified large effects, but likely mirrors the complex genetic susceptibility landscape that is driving risk in simplex cases.

**Figure 3.6** The results of a family-based test of association (CAPL) inside the support intervals of the linkage peaks that reach LOD* >1. The nominal significance threshold is at -log($P$-value) = 3, and the Bonferroni-corrected significance threshold for all 1716 markers is at -log($P$-value) = 11.38. Only markers with variance of the test statistic > 5 are plotted. The most significant marker is plotted as a diamond, and the r^2 between this marker and the remaining ones in the region is indicated by the color of the point, according to the legend. Some of the genes inside the regions are not plotted, as indicated by the note next to each plot. LocusZoom was used to create these plots (4).
4.1 The case for rare variants in disease susceptibility

4.1.1 Population genetic theory and rare variants

For a long time, genetic studies in complex phenotypes have been based on the common-disease common-variant (CDCV) hypothesis which predicts that the genetic risk for common diseases is largely due to common disease-predisposing alleles (153). Stated another way, the CDCV hypothesis predicts that there will be one or a few disease alleles (of high frequency in the population) at each susceptibility locus identified for a given disease. Although there isn’t enough evidence to prove or disprove the CDCV, advances in sequencing technology have enabled the interrogation of rare variation. Next-generation sequencing (NGS) technology has the capability to generate data on all of the
variation that is present on a DNA fragment, common and rare, without prior specification of the positions to be interrogated. These technological advances have enabled the research community to focus on examining the common disease-rare variant hypothesis (CDRV) \((154)\). The CDRV hypothesis predicts that multiple, rare alleles with large effects underlie the majority of the genetic risk to common complex diseases. The CDCV and CDRV hypotheses examine different areas of the susceptibility space and are therefore not mutually exclusive; in fact, for many complex diseases, evidence suggests that both common and rare variants have their roles in driving susceptibility (reviewed by \((155)\)).

Three of the main parameters that define the genetic architecture of a disease are the number of risk alleles, the total frequency of the risk alleles, and the penetrance of each combination of these alleles; the latter is arguably the most difficult to characterize. The CDCV and the CDRV hypotheses make different predictions with respect to these parameters. For Mendelian phenotypes, the total frequency of risk alleles is more clearly determined by the balance between mutation and selection: mutation serves to introduce disease alleles, while selection serves to remove them \((156)\). The effect of random genetic drift is not a major contributor. On the other hand, where selection is weak, drift and mutation will act to drive the allele frequencies away from equilibrium and the frequencies of the risk alleles will therefore vary widely between susceptibility loci (summarized by \((157)\)).

In fact, population genetic theory offers several models in which rare variants are consistent with the demographic history of the human population and the disease parameters of many of the common, complex diseases. For example, Pritchard has
developed a model that qualitatively describes complex disease loci, incorporating mutation, random genetic drift, and different levels of purifying selection at the disease loci (158). The results indicate that loci with high mutation rates and subject to weak purifying selection will likely contribute disproportionately to the genetic variance. Importantly, loci for which the mutation rate is on the high end of the predicted range and where the effects of purifying selection are weak will show extensive allelic heterogeneity leading to an abundance of rare variants (158). At these loci, the total frequency of risk alleles is likely to be high. Therefore, the extensive allelic heterogeneity that is predicted for complex disease loci supports not only the relevance of rare variants in disease susceptibility but also the use of methods that are robust to allelic heterogeneity, such as haplotype association and gene burden tests. It also provides support for family-based studies because the regions of haplotype sharing are expected to be fairly large, especially where the risk alleles are rare (158).

Coding variants are often classified into a number of functional categories based on their predicted effect on the protein product: nonsense, probably damaging, possibly damaging, missense, benign, and synonymous. Damaging variation is typically younger, because purifying selection does not allow it to persist. Therefore, rare variants – and especially damaging rare variants – are common, in the sense that sequencing more samples will lead to the discovery of more and more of these unique damaging variants; in other words, they are less likely to have been previously seen in other samples than benign variants (159).

The rate of deleterious mutation (defined here as those that are subject to purifying selection, whether or not they have a detrimental effect on protein
structure/function or a medical consequence) depends on both the rate of substitution and
the fraction of those substitutions that are deleterious. Bioinformatic approaches have
estimated the prevalence of mildly deleterious mutations at 70% of missense mutations
seen only once in 1,500 chromosomes. Purifying selection (selection coefficient in the
range of 0.001 to 0.003) acts on these mildly deleterious mutations enough to keep them
at low frequencies in the population (160). A high rate of mildly deleterious mutations
provides support that mutation-selection balance has a role in shaping the genetic
susceptibility landscape to common, complex diseases.

4.1.2 Examples of rare variants in complex disease

Multiple examples exist in the literature of rare variants with large effects on complex
phenotypes. In this context, “large” is defined relative to the effect sizes of risk alleles
identified through association studies, which given the currently available sample sizes
extend down to detection levels in the range of OR ~ 1.1.

Most examples of rare variants with roles in complex phenotypes have been
identified through candidate gene studies. One example is the observation that
nonsynonymous variants in three genes that cause Mendelian forms of low plasma high
density lipoprotein C (HDL-C) are much more common in the 5% of population samples
with the lowest HDL-C levels than in those above the 95% percentile. The
nonsynonymous, rare, coding variants for one of the genes (ABCA1, encoding the
adenosine triphosphate binding cassette transporter A1 required for the formation of HDL
particles after cholesterol leaves the cells) were present in 5% of the low HDL-C group
and 0.03% of the high HDL-C group. The other two candidate genes were apolipoprotein
A1 (APOA1; the major component of HDL) and lecithin cholesterol acetyltransferase
(LCAT; an enzyme that catalyzes the formation of cholesteryl esters in HDL). For these two genes, the rare variants were present in lower frequencies in both the low and high HDL-C groups: 2% of low HDL-C individuals carried nonsynonymous variants in LCAT in contrast to 0.03% of high HDL-C individuals, and 0.03% of those in the low HDL-C group carried nonsynonymous variants in APOA1 compared to 0% of those in the high HDL-C group (161). Biochemical characterization of the effects of these variants, combined with the prior knowledge that mutations in them cause Mendelian forms of the phenotype was crucial in identification of the contribution of rare variants at these loci. This furthers the idea that susceptibility loci for complex traits may contain both common and rare disease alleles.

Similar findings are reported for obesity. Rare, nonsynonymous variants have been observed in genes related to monogenic causes of obesity and body weight-related pathways such as lipid metabolism, adipogenesis, energy expenditure and food intake (162). In a dataset of approximately 380 obese and 380 lean individuals, the candidate genes were enriched for rare (here defined as < 1% frequency in the population), nonsynonymous variants distributed across all of the genes so that the excess was not due to clustering of variants in any single gene. An examination of the variants that were seen exclusively in one group or the other revealed a significantly larger proportion of variants in the obese group that were annotated as deleterious by a damage prediction algorithm, PolyPhen2 (163). This difference was observed exclusively for the genes associated with monogenic forms of obesity. Evidence in previously un-associated genes in body-weight related pathways was suggestive, but the low frequency of the nonsynonymous variants prohibited their confident association with obesity. The two best examples were SIM1, a
gene that carried 6 nonsynonymous variants in the obese and 0 in the lean group, and PRKAG3 with 10 variants in the obese and 4 in the lean group.

Another 4 rare variants that lower risk for type 1 diabetes (T1D) were identified in a study that sequenced 10 candidate genes – 6 of which contained common T1D variants – in 480 patients and 480 controls. Three of these four variants in the IFIHI gene are intronic, all four have a frequency < 3% in the general population and were confirmed in a sample of 30,000 individuals (164).

Plasma triglyceride levels are significantly associated to rare, nonsynonymous variants in four angiopoietin-like genes (ANGPTLs) in the Dallas Heart Study cohort (1,870 African Americans, 1,045 individuals of mixed European descent, and 601 Hispanics). One percent of the samples in this dataset had a rare variant in one of the candidate ANGPTL genes compared to four percent of those in the lowest quartile for plasma triglyceride levels. Functional studies proposed a potential mechanism for the effect of these rare variants – strengthening the evidence for their involvement in variability of plasma triglyceride levels. Despite the limited search space, this study provides another example of how rare variants are common: 86% of the 255 variants found in the candidate genes had a frequency < 0.01, and more than half of them were singletons (seen in only one individual) (165).

The general hypothesis that an excess of rare variants are present in disease-associated genes was also tested in hypertriglyceridemia. In this study, Johansen et al. sequenced the exons of four genes identified through GWAS in 438 hypertriglyceridemia patients, and found significantly more rare, missense or nonsense variants in patients than
in 327 controls. Cumulatively, 21.8% of hypertriglyceridemia patients carried one of these variants, compared to 15.3% of controls \((166)\).

In another example, Won et al. sequenced the exomes of 4,000 individuals with early onset myocardial infarction (MI) and 3,800 controls to test the hypothesis that rare (here defined as frequency < 0.01) variants contribute to risk for MI. A number of genes went forward for validation in a replication dataset of 13,432 individuals using targeted resequencing. However, only two genes previously detected to carry common variants associated with MI reached the stringent significance threshold that accounted for multiple testing correction \((2E-06)\) for the association of MI with rare, nonsynonymous variants in gene burden tests. One of them was the low-density lipoprotein receptor \((LDLR)\) with rare, nonsynonymous mutations in 4% of cases compared to 1.3% of controls \((OR = 3.1)\), and apolipoprotein A-V \((APOA5)\) where nonsynonymous variants were detected in 1.3% of cases compared to 0.6% of controls \((OR = 2.2)\). The low frequencies of these variants complicate the confident detection of genuine associations, rendering the established role of these genes in MI as major supporting evidence.

On the other hand, a search for rare variants in candidate genes of 1,000 patients with type 2 diabetes (T2D) revealed no genes with significant disease associations. The authors used gene-based tests that increase statistical power by virtue of testing the total frequency of many rare alleles and further confirmed that they had sufficient power to detect at least one causal gene using extensive simulations. The foundational assumption of this study was that some of the heritability of T2D is explained by rare variants in the coding regions of a small \((\sim 20)\) number of genes. Therefore, it seems that rare, coding
variants have different roles in different complex phenotypes and the unraveling of the precise roles will likely proceed on a disease-by-disease basis.

There are at least three main points that emerge from the examination of these examples. One is the collection multiple pieces of evidence each of which may not have been sufficient on its own to make any conclusions about the involvement of those variants in the phenotype, but which together make a stronger case. These include the biological plausibility of the genes, functional characterization of the effects of the variants, and the replication in large datasets of population samples. As a side note here, the sample sizes used in the replication studies were powerful enough only because a limited number of variants were being examined. Markedly larger population samples would be required if any type of multiple testing correction were to be applied to the association tests. The second point is the observation that studies have focused on rare, nonsynonymous variation, ignoring possible effects exerted by regulatory intergenic, intronic, or promoter region variants. Nonsynonymous variants are the focus of current studies due to their higher probability of having a damaging effect than other variant classes. The third point is that the effect sizes of the rare variants that have been discovered are in the range of OR ≈ 2. This is larger than the effect sizes of most common variants implicated in MS, but smaller than the largest effects exerted by HLA.

An increase in power can be achieved if multiple variants found in a gene are combined into a burden score. Multiple versions of gene-based tests have been developed for association studies in unrelated population samples. Each of them tackles a different challenge (reviewed by (167, 168)). However, empirical and theoretical studies are pointing out that large sample sizes are required for the identification of these effects by
association studies, even when the focus is on nonsynonymous, damaging variants in reasonable biological candidates that are considered in gene burden tests. Even for gene-based tests, power calculations indicate that datasets in the order of tens of thousands of individuals are required for an agnostic approach to have adequate power (160).

These observations highlight the main challenge posed by rare variant studies: variation arises continuously at a rate of approximately 1E-5 per gene per generation (for nonsynonymous variants) and it does not necessarily have pronounced effects on fitness since most novel nonsynonymous variants result in <1% of fitness loss. The large numbers of rare variants that are continuously discovered are is a reflection of this statement. Therefore, support from multiple pieces of evidence is required to implicate rare variants in disease. In facilitating the search for rare variants that influence complex phenotypes, sequencing the exome can serve as one way to reduce the search space, with the added advantage of relatively straightforward annotation of coding variation.

4.1.3 Family study designs in the search for rare variants

Extended families are extremely valuable in sequencing studies to identify potential disease-predisposing variants. One of the reasons for this is that the shared regions that carry rare alleles occur in multiple individuals in an extended family, reducing the challenge posed by the identification of singletons (i.e. variants identified in only one individual) which are difficult to distinguish from sequencing errors. A related advantage is that families provide a sample of genetically homogeneous individuals for the identification of disease-predisposing variants (169). This is especially true for families from isolated populations such as Sardinian families. The structure of relatedness in the dataset inform study design and the choice of samples to sequence (discussed in (170)).
Imputation using information from the sequenced relatives may be used to infer variation carried by relatives who are not sequenced, based on the haplotypes of shared segments (171). The caveat is the need to sequence both affected and unaffected relatives for imputation to be reliable since, in case-only designs the inference of the imputed variants would be based on potentially disease-associated haplotypes.

A successful example of a rare coding variant that was identified with the use of multiplex pedigrees from an outbred population is that of the phospholipase D3 gene (PLD3) in Alzheimer’s disease (172). In the initial analysis, the authors performed exome sequencing of a single pair of affected individuals and one unaffected individual from 14 families with European American ancestry. They filtered the variants to keep the non-synonymous variants that were present in both affecteds, absent from the unaffected individual from each family, and with minor allele frequency (MAF) < 0.05. Subsequent genotyping of 75 variants in the remaining family members narrowed down the list to those that segregated with the disease status. Of those, a single variant was present in more than one family and was successfully replicated in a population cohort of 11,000 samples. This variant is present in 1.3 to 2.6% of cases, and 0.26 to 0.79% of controls and has an OR = 2.10 (95% CI 1.47, 2.99). Further resequencing of the PLD3 gene in over 4,300 individuals identified additional variants that increased risk for late onset Alzheimer’s disease. Functional studies provided additional evidence of a possible mechanism for PLD3 in disease pathogenesis.

Isolated populations are likely to offer increased power to detect disease-predisposing variants which are rare in outbred populations but have risen in frequency in the population isolate. The reduced effective population size during the founding of the
Sardinian population may have attenuated natural selection against alleles that are deleterious with respect to phenotypes and allowed them to reach higher frequencies in this population isolate (0.005 to 0.05), enough to be detected in large scale population studies and increasing the probability that they are enriched in high risk families.

One example of a rare variant identified in an isolated population is the R19X variant in APOC3, which is associated with increased HDL and decreased triglyceride levels and was previously thought to be private to the isolated population of the Amish. This APOC3 variant has risen to a frequency of 0.038 in Greek population isolates compared to < 0.005 in outbred European populations (173), highlighting the practical consequences of the theoretical advantages mentioned above.

4.2 Methodological approach

4.2.1 Characteristics of the 5 sequenced pedigrees

Exome sequencing data was available on of 32 individuals from 5 multiplex pedigrees of multiple sclerosis from the Sardinian population isolate (see Chapter 2 for dataset description). The families in this dataset with the largest numbers of affected individuals were targeted for sequencing (range 4 to 11), and to minimize the kinship coefficient between the affected relative pairs. The minimum kinship coefficient between any pair in a family ranged from 0.000122 to 0.003906, thereby limiting the search space to the coding regions that are shared between affected relative pairs. The non-parametric LOD* scores calculated across all 24 sub-pedigrees in the dataset were used to define the regions of sharing (see Chapter 3). Specifically, a threshold of LOD* ≥ 1 was used to identify shared regions, as described below.
Table 4.1 Overview of sequenced pedigrees.

<table>
<thead>
<tr>
<th>Pedigree</th>
<th>Individuals (N)</th>
<th>Unaffected with DNA (N)</th>
<th>Affected with DNA (N)</th>
<th>Number of lineages</th>
</tr>
</thead>
<tbody>
<tr>
<td>23_60</td>
<td>151</td>
<td>28</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>61</td>
<td>49</td>
<td>20</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>45</td>
<td>42</td>
<td>18</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>42</td>
<td>24</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>87</td>
<td>22</td>
<td>7</td>
<td>3</td>
</tr>
</tbody>
</table>

In addition, the SardiNIA Medical Sequencing Discovery Project dataset was downloaded from the database of Genotypes and Phenotypes (dbGaP) housed at the National Center for Biotechnology Information (NCBI) (dbGaP study accession: phs000313.v3.p2, (174)). Whole-genome sequencing data on 505 unrelated Sardinian individuals was downloaded at 2x coverage. After variant calling with GATK (175, 176), I compiled a summary file with the allele counts at each position along the genome. This was necessary because the low coverage prohibited reliable calls on individual-level data. The resulting allele frequencies were used as a representative source of the allele frequencies in the Sardinian population isolate.

Given that HLA haplotypes represent a significant amount of MS risk, it is important to assess the status of the sequenced families for the HLA risk haplotypes. Rare alleles of relatively large effects may be more likely to segregate in families that are negative for the HLA risk haplotypes. In Sardinia, the ancestral HLA haplotypes that are associated with MS are A30Cw5B18DR3DQ2, which confers MS risk (OR 1.78, 95% CI 1.26-2.50), and A2Cw7B58DR2DQ1, which is protective (OR = 0.27, 95% CI 0.13–0.57) (98). HLA serotyping provided the number of affected individuals in each family with either the entire HLA haplotypes or the DR2 (protective), and DR3 (risk) alleles of DRB1 (Table 4.2). The DR2 allele represents the broad specificity of the original
serological DR16 allele for DRB1 according to official serological HLA nomenclature (177); it does not represent the DRB1*1501 allele that is known to be associated with MS in Northern European populations. These data represent the broad specificities of the serological HLA alleles obtained through serotyping. Note that in some cases, the same affected individual carries both the DR3 and DR2 alleles reported for that family.

<table>
<thead>
<tr>
<th>Pedigree</th>
<th>Serotyped affecteds (N)</th>
<th>At least one DRB1 protective allele</th>
<th>Protective haplotype*</th>
<th>At least one DRB1 risk allele</th>
<th>Risk haplotype*</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>45</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>61</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>23_60</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>

* The entire Sardinian HLA A2Cw7B58DR2DQ1 or A30Cw5B18DR3DQ2 haplotypes.

4.2.2 Exome sequencing, alignment and processing

Exome capture and sequencing were performed at the Center for Genome Technology (CGT) housed within the John P. Hussman Institute for Human Genomics (HIHG) at the University of Miami Miller School of Medicine. The exome was targeted using the SureSelect Human All Exon and Human All Exon Plus target enrichment kits manufactured by Agilent. Target enrichment ensures high coverage of the regions of the genome that are of interest during the sequencing process. Briefly, a genomic DNA library was generated from 3μg of DNA and quantified using the Qubit dsDNA assay. Sample library preparation and sequence capture for NGS is automated using the Caliper Sciclone G3 and Caliper Zephyr liquid handling workstations. The DNA library was
processed according to the Illumina Paired-End Sample Preparation Guide, with the modifications listed in the Agilent Sure Select protocol (v2.0.1). The resulting 150-200 nucleotide fragments are captured by 120-mer cRNA, biotin labelled probes that are designed for the coding regions. Multiple samples can be run in parallel, using oligonucleotide barcodes to identify the sample that each fragment originates from. The samples are barcoded post-capture. The probes hybridize to the regions that they were designed to capture, and are then separated using streptavidin-coated magnetic beads, amplified, purified, and quantified using the Qubit dsDNA BR Assay and the Caliper LabChip® GX (automation) (178). The library is then loaded on the Illumina cBot for cluster generation according to manufacturer recommendations. Massively parallel sequencing occurred on the Illumina HiSeq 2000 using the reagents provided in the Illumina TruSeq PE Cluster Kit v3 and the TruSeq SBS Kit-HS (200 cycle) kit.

Sequencing by synthesis proceeds by the detection of a fluorescently labelled nucleotide that is read as it is incorporated into the DNA. The fluorescent label is then cleaved for the next base to be incorporated during the following cycle. Both the capture technology and the sequencing technology are designed to minimize bias: the SureSelect system uses conditions that allow for mismatches during hybridization so that single nucleotide variants as well as copy number variants can be captured, while the HiSeq technology keeps all four nucleotides in the solution so that competition between the bases determines which one will be incorporated.

Processing of the sequencing data is a two-step procedure that starts after image analysis has been completed by the HiSeq’s Real Time Analysis (RTA) capability implemented by the Illumina-provided software, Casava. If the RTA fails then off-
sequencer base calling is completed by the Illumina-provided software Bustard. The raw read files are filtered by Illumina’s internal chastity filter and prepared for further processing. After filtering, the raw sequencing reads are introduced to the bioinformatics pipeline for variant calling. The HIHG bioinformatics pipeline conforms to the Broad Institute Best Practices Guidelines with the Genome Analysis Toolkit (GATK) v3 (175, 176). The FASTQ sequences are globally aligned to the human genome reference (hg19) using the Burrows-Wheeler Aligner (BWA). The aligned reads are then input into GATK for local realignment.

**Figure 4.1** Summary metrics on the depth of coverage, obtained using Picard (http://picard.sourceforge.net).

around indels, removal of PCR duplicates, and base quality recalibration. Variant calling is performed by GATK’s Unified Genotyper and it outputs Variant Call Format (vcf) files. Basic quality control metrics are also generated using the HIHG in-house pipeline. The main quality metrics that ensure the technical integrity of the experiment include fold-coverage at each of the targeted coding regions across all sequenced samples, as well as coverage statistics per sample. These are summarized in Figure 4.1.

### 4.2.3 Quality control and annotation of exome sequencing data

Starting with 77,604,263 to 90,294,155 calls in the coding regions of each of the 32 individuals, I kept those calls which were polymorphic, had genotype quality score (GQ)
> 30, and a depth of coverage (DP) > 4 in at least one individual. Following this initial filter, each sample retained 9-12% of the calls, generating a list of 70K - 113K variants per individual. I subsequently filtered out variants in each sample that had the Phred-scaled Genotype Likelihood (PL) of the homozygous reference genotype call > 100, indicating high confidence in the homozygous reference genotype. I then identified the unique positions that were kept in any one sample and compiled a list of 361,519 variants that pass these initial quality control steps across all individuals. These initial filters were applied using VCFtools v0.1.9.0 (179). Additionally, I calculated the Variant Quality Score Recalibration Lodscore (VQSLOD) after joining the 32 Sardinian samples with another 132 samples with 50MB exome data generated using the same Agilent capture kit at the HIHG. The VQSLOD represents an estimate of the probability that a variant call is a true genetic variant and not a technical artifact; it is generated adaptively using true, validated calls (truth set) to cluster the observed variants (training set) and assign scores to all variants. An in-house Perl script was used to calculate the VQSLOD. Filtering out all variants with VQSLOD < -4 retained 314, 702 variants across all individuals. This variant set constitutes the clean dataset for subsequent analyses.

The variants in the clean dataset were annotated as follows: allele frequencies in the European American, African American and combined populations were retrieved from the Exome Variant Server (EVS), part of the NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/). Additionally, the SeattleSeq Annotation 138 (180) server was used to retrieve functional annotation on all of the variants in the clean dataset, including the function of the variant as annotated in the Genome Variation Server (GVS) (intronic, exonic, missense, synonymous, near splice
site, in 5’ untranslated region (UTR), in 3’ UTR, intergenic, downstream of the nearest gene, upstream of the nearest gene). Furthermore, I retrieved the Genomic Evolutionary Rate Profiling (GERP) score (181), and the Phylogenetic Analysis with Space/Time score (PhastCons) (182), both of which are measures of conservation. PolyPhen2 scores, representing the predicted impact of nonsynonymous variants on protein structure and function were also added to the list of annotations (163). These were merged with the frequency in the SardiNIA population dataset to generate a masterfile containing the annotations shown in Table 4.3.
4.3 Coding variants in regions shared between affecteds

I examined two hypotheses with respect to variants that are shared in the exomes of MS-affected individuals in 5 Sardinian multiplex pedigrees. The first hypothesis is that rare variants in genes that are implicated by the established MS-associated loci are influencing disease susceptibility in these Sardinian families. To this end, I defined rare as having a frequency of < 0.05 in outbred Europeans. Their relevance to disease was assessed by virtue of identity-by-descent (IBD) sharing in pairs of distant relatives from 5 Sardinian multiplex pedigrees. The second hypothesis explores variants that fall inside dataset-wide regions of sharing as defined by the linkage peaks (see Chapter 3) as potentially relevant to disease. Using the variants under the linkage peaks, I examined the evidence for a founder effect increasing the frequencies of possible disease-relevant variants, identified the genes implicated by rare, nonsynonymous variants, and finally I characterized the exome variants in the genes identified through the fine-mapping analysis of the linkage peaks as described in Chapter 2. An overview of the analytical approaches taken to explore the exome data for variants that may be relevant to MS susceptibility is shown in Figure 4.2.

4.3.1 Rare variants shared identical-by-descent in MS-associated regions

The clean dataset of 321,402 variants was restricted to those that are annotated as falling in one of the 149 genes that are physically closest to one of the 110 established MS-associated SNPs according to their Genome Variant Sever Gene List annotation (33). Out of the 110 loci, 47 are annotated as intergenic; the remainder is annotated as intronic, exonic or in the 3’ or 5’ UTR. For the intergenic loci the median distance to the closest gene is 6,626 base pairs, although this measure is of little importance if these associations
**Figure 4.2** Diagrammatic overview of the various filtering approaches taken to explore the exome data for variants related to MS susceptibility. IBD = identical-by-descent, both populations = SardiNIA and Exome Variant Server European American datasets.
represent mutations in regulatory regions. For the intergenic regions, both genes flanking the associated marker were included, explaining the larger number of genes examined here compared to the number of MS-associated loci. A total of 2,395 variants are retained after this filter is applied.

IBD estimates were calculated separately for each family using ImmunoChip data on the 11,856 markers that comprise the linkage map (see Chapter 2). Use of the relatively sparse Immunochip data was essential because sequencing data was available only for the subset of individuals in each pedigree who are MS-affected. The MERLIN haplotyping option --best was used to estimate the haplotypes in the 9 sub-pedigrees of the sequenced families (see Chapter 3 for details on how the pedigrees were subdivided). The allele frequencies used in haplotype estimation represented the Sardinian population controls and the founders of all 19 pedigrees.

<table>
<thead>
<tr>
<th>Subpedigree</th>
<th># Affected</th>
<th># Exomes</th>
<th>Number of affecteds sharing regions IBD</th>
<th>Number of variants in shared regions</th>
<th>Number of variants in shared regions in 149 genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>23_601</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>241</td>
<td>21</td>
</tr>
<tr>
<td>23_602*</td>
<td>4</td>
<td>3</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>23_603</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>12290</td>
<td>43</td>
</tr>
<tr>
<td>41</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>244</td>
<td>0</td>
</tr>
<tr>
<td>51</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>2052</td>
<td>57</td>
</tr>
<tr>
<td>52</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>451</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>111</td>
<td>0</td>
</tr>
<tr>
<td>611</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>532</td>
<td>8</td>
</tr>
<tr>
<td>612</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>7425</td>
<td>57</td>
</tr>
</tbody>
</table>

*Pedigree 23_602 is not included in the nearest gene analysis due to memory limitations with Olorin when loading this pedigree.
The information on the patterns of allele flow and all clean exome variants were combined using Olorin (183). The regions of IBD sharing between N out of T affecteds were filtered to maximize N while still allowing for > 0 shared regions (Figure 4.3). A second filter was applied to ensure that at least 1 alternative allele (i.e. not reference) was carried by all affecteds. Table 4.4 shows the number of variants that remain in each family when these filters are applied.

Next, we merged the variants that were identified as IBD in affected individuals across all families. A total of 1793 variants out of the 20,833 that are in shared regions in any of the families were shared IBD between affecteds of at least two families. An additional 14 out of the 20,833 were shared IBD between affecteds of at least three families (Table 4.5). These 14 variants were identified in an analysis where IBD information was calculated on a per-family basis, in contrast to the analysis using the dataset-wide LOD* scores as described in the following section.

Filtering on the 2,395 variants across the 149 genes in or around MS-associated loci retains 144 variants from the 20,833 variants that are in regions shared IBD between affecteds in any of the families. The mean MAF in SardiNIA of the 144 variants is 0.28 (standard deviation (SD) = 0.19), and the mean GERP score is -1.82 (SD = 4.13) (Figure 4.4). In terms of the functional impact of these 144 variants, 17 of them are annotated as missense. The mean GERP score of these 17 missense variants is -0.55. A total of 12 out of the 144 variants are shared between affecteds in at least two different families. All 12 have similar MAFs in Sardinians and Europeans (median MAF = 0.39), and one of them is rare (MAF < 0.05) in both populations (rs6462722 in an intron of ELMO1, GERP score = 2.76).
Table 4 Variants in regions shared IBD between affected individuals in three different families. The variants on chromosome 16 are relatively more interesting because they are shared between sub-pedigrees that originate from unrelated families.

<table>
<thead>
<tr>
<th>Chr</th>
<th>Position</th>
<th>rs ID</th>
<th>Family 23_603</th>
<th>Family 51</th>
<th>Family 611</th>
<th>Family 612</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>66463251</td>
<td>rs4856924</td>
<td>.</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>3</td>
<td>66465259</td>
<td>rs4856928</td>
<td>.</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>16</td>
<td>24741686</td>
<td>rs7202252</td>
<td>Y</td>
<td>Y</td>
<td>.</td>
<td>Y</td>
</tr>
<tr>
<td>16</td>
<td>24807165</td>
<td>rs1030211</td>
<td>Y</td>
<td>Y</td>
<td>.</td>
<td>Y</td>
</tr>
<tr>
<td>16</td>
<td>24902194</td>
<td>rs274081</td>
<td>Y</td>
<td>Y</td>
<td>.</td>
<td>Y</td>
</tr>
<tr>
<td>16</td>
<td>25137449</td>
<td>rs7188975</td>
<td>Y</td>
<td>Y</td>
<td>.</td>
<td>Y</td>
</tr>
<tr>
<td>16</td>
<td>46565019</td>
<td>rs1436436</td>
<td>Y</td>
<td>Y</td>
<td>.</td>
<td>Y</td>
</tr>
<tr>
<td>16</td>
<td>46602415</td>
<td>rs7191748</td>
<td>Y</td>
<td>Y</td>
<td>.</td>
<td>Y</td>
</tr>
<tr>
<td>16</td>
<td>46642456</td>
<td>rs8052092</td>
<td>Y</td>
<td>Y</td>
<td>.</td>
<td>Y</td>
</tr>
<tr>
<td>16</td>
<td>46655210</td>
<td>rs6598679</td>
<td>Y</td>
<td>Y</td>
<td>.</td>
<td>Y</td>
</tr>
<tr>
<td>16</td>
<td>46655516</td>
<td>rs1865837</td>
<td>Y</td>
<td>Y</td>
<td>.</td>
<td>Y</td>
</tr>
<tr>
<td>16</td>
<td>46696284</td>
<td>rs168745</td>
<td>Y</td>
<td>Y</td>
<td>.</td>
<td>Y</td>
</tr>
<tr>
<td>16</td>
<td>46715191</td>
<td>rs700581</td>
<td>Y</td>
<td>Y</td>
<td>.</td>
<td>Y</td>
</tr>
<tr>
<td>16</td>
<td>46764482</td>
<td>rs36460</td>
<td>Y</td>
<td>Y</td>
<td>.</td>
<td>Y</td>
</tr>
</tbody>
</table>

Figure 4.3 Visualization of the segments shared identical-by-descent (IBD) in each subpedigree using Olorin. The height of the bars next to each chromosome represents the number of cases that share that region IBD. The colors represent regions that are shared between N affecteds, where N is a number specified by the user as indicated by Table 4.4. For example, for family 611, 4 out of 6 affecteds with exome sequencing data share any regions IBD, these regions are colored green. The height of the blue bars indicates successively lower thresholds of sharing. For example, for family 611, any 3 out of the 6 affecteds share the regions indicated by the tall blue bars, and any 2 of the 6 share the regions indicated by the short blue bars. Note: Subpedigree 52 showed no shared regions between the two affected individuals. Subpedigree 23_602 is excluded from this figure.
Figure 4.4 The (a) alternative allele frequency and (b) conservation score (GERP) profiles of the 144 variants that are shared IBD in affected individuals in any of the Sardinian families, located in genes that are implicated by the MS-associated loci.

Variants that are rare in Europeans but with a higher frequency in Sardinia are potentially interesting as they may represent the effects that are responsible for the increased MS prevalence observed in Sardinia. To this end, I filtered the set of 144 variants to retain only those that were rare in Europeans according to both the HapMap and EVS datasets; 36 variants are retained. Only one of the 36 variants is novel (defined here as absent from the EVS, HapMap and 1000G datasets), and is located in an intron of CDH3 with a SardiNIA MAF = 0.23. This variant (chromosome 16, hg19 position 68713622) is seen in the heterozygous state in 28 out of 31 affecteds in all subpedigrees, as well as in the unaffected father in family 61. The remaining 35 variants have rsIDs, indicating their presence in dbSNP. The CEU population of phase 1 of the 1000 Genomes Project, the CEPH families, and the TSC panel of 42 Caucasian individuals provided MAFs for 26 out of the 35 variants with rsIDs; 20 of these 26 have MAFs between 0.06
and 0.49. A variant in the *MANBA* gene is 16X more frequent in Sardinia, affording a glimpse into the effects of population genetic forces that have acted on this population isolate.

The 16 variants out of the 36 that are also rare or absent from the 1000 Genomes Project and other European datasets are shown in Table 4.5. Of the variants shown in Table 4.5 two of them may be of interest: a probably damaging, missense variant in the *MANBA* gene that is 5X more frequent in Sardinia, and another probably damaging, missense variant in *IKZF1* gene that is 10X more frequent in Sardinia; both of them are highly conserved. The *MANBA* variant is present in 6 out of 7 affecteds from family 61, and absent from the unaffected father. The *IKZF1* variant is present in all affecteds and the single unaffected father from family 61. Neither variant is present in any of the other families.
**Table 4.6** Rare (in outbred Europeans) variants in genes near MS-associated loci that reside in regions shared IBD in the affected individuals of one or more Sardinian families

<table>
<thead>
<tr>
<th>Chr</th>
<th>Pos (hg19)</th>
<th>rsID</th>
<th>Ref.</th>
<th>Alt.</th>
<th>Gene</th>
<th>Function</th>
<th>GERP score</th>
<th>MAF in SardiNIA*</th>
<th>Families</th>
<th>MAF in dbSNP†</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>85787114</td>
<td>rs233116</td>
<td>A</td>
<td>T</td>
<td>DDAH1</td>
<td>3-prime-UTR</td>
<td>-7.3</td>
<td>0.001</td>
<td>51</td>
<td>0.008</td>
</tr>
<tr>
<td>4</td>
<td>103590196</td>
<td>rs202100372</td>
<td>T</td>
<td>C</td>
<td>MANBA</td>
<td>missense</td>
<td>4.82</td>
<td>0.092</td>
<td>612</td>
<td>0.018</td>
</tr>
<tr>
<td>4</td>
<td>103675108</td>
<td>rs223492</td>
<td>G</td>
<td>C</td>
<td>MANBA</td>
<td>intron</td>
<td>-0.802</td>
<td>0.476</td>
<td>51</td>
<td>0.0283</td>
</tr>
<tr>
<td>5</td>
<td>55398269</td>
<td>rs318818</td>
<td>G</td>
<td>C</td>
<td>ANKRD55</td>
<td>intron</td>
<td>3.74</td>
<td>1</td>
<td>51</td>
<td>0.017</td>
</tr>
<tr>
<td>6</td>
<td>36368512</td>
<td>rs7755006</td>
<td>G</td>
<td>A</td>
<td>PXT1</td>
<td>intron</td>
<td>-0.451</td>
<td>0.494</td>
<td>23,603</td>
<td>NA</td>
</tr>
<tr>
<td>7</td>
<td>36934448</td>
<td>rs6462722</td>
<td>A</td>
<td>G</td>
<td>ELMO1</td>
<td>intron</td>
<td>2.76</td>
<td>0.02</td>
<td>51,23,603</td>
<td>0.033</td>
</tr>
<tr>
<td>7</td>
<td>50173979</td>
<td>rs998929</td>
<td>T</td>
<td>C</td>
<td>C7orf72</td>
<td>intron</td>
<td>-0.451</td>
<td>0.4</td>
<td>612,23,603</td>
<td>NA</td>
</tr>
<tr>
<td>7</td>
<td>50173981</td>
<td>rs1122346</td>
<td>T</td>
<td>C</td>
<td>C7orf72</td>
<td>intron</td>
<td>0.207</td>
<td>0.171</td>
<td>612</td>
<td>NA</td>
</tr>
<tr>
<td>11</td>
<td>118529069</td>
<td>rs6589670</td>
<td>G</td>
<td>C</td>
<td>Treh</td>
<td>missense</td>
<td>1.33</td>
<td>1</td>
<td>611</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>118529078</td>
<td>rs6589671</td>
<td>G</td>
<td>C</td>
<td>Treh</td>
<td>missense</td>
<td>-9.7</td>
<td>0.003</td>
<td>611</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>118529017</td>
<td>rs6589669</td>
<td>T</td>
<td>G</td>
<td>Treh</td>
<td>synonymous</td>
<td>2.25</td>
<td>1</td>
<td>611</td>
<td>NA</td>
</tr>
<tr>
<td>12</td>
<td>6439255</td>
<td>rs10849443</td>
<td>A</td>
<td>G</td>
<td>TNFRSF1A</td>
<td>intron</td>
<td>2.96</td>
<td>1</td>
<td>612</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>123471402</td>
<td>rs2049506</td>
<td>T</td>
<td>C</td>
<td>PITPNM2</td>
<td>intron</td>
<td>0.157</td>
<td>0.002</td>
<td>612</td>
<td>0.017</td>
</tr>
</tbody>
</table>

* Variants that are monomorphic in the SardiNIA dataset have the alternative allele in the Sardinian families.
†The MAF retrieved from dbSNP represent one of the three following populations: Europeans from phase 1 of the 1000 Genomes Project, individuals of European descent from Agilent's ClinSeq project, or CEU individuals from phase 3 of the HapMap Project. Variants that have an rsID but no MAF reported in dbSNP may have an rsID if they were observed as singletons.
4.3.2 Common and rare variants under linkage peaks

The experiments described in this section represent a relatively more agnostic approach than restricting the exome to MS-associated loci. In contrast to the experiments described in the previous section, the linkage peaks used here were obtained by a dataset-wide analysis so that all families have the potential to contribute to the LOD* score.

I defined the regions of sharing using the support interval around the LOD* score (see Chapter 3), calculated using the 1-LOD*-down method. Briefly, I dropped one LOD* score unit on either side of the maximum LOD* score and used this interval to define the boundaries of the shared region (152). A total of 5 regions fall within the support intervals of the LOD* scores (Table 3.4). I further limited the search space by considering rare, nonsynonymous variants in the genes closest to the associated regions, as determined by the fine-mapping results of the family-based association test (see Chapter 3).

4.3.2.1 A closer look at functional classes

In the sequenced families, 12,865 variants are contained in the 5 regions bound by the support intervals of the linkage peaks; 3,503 of them are novel. The functions of the 11,444 variants out of the 12,865 that are annotated with the GVS annotations shown in Figure 4.5 reveal that most of them are intronic (note that this figure is not normalized for the representation of each functional category in the exome). Approximately 12.7% of the 12,865 variants are missense mutations, and 544 of these missense variants are annotated as novel. The PolyPhen2 score distribution of 1,855 variants that are annotated as nonsynonymous is shown in Figure 4.5. A count of the genes that are represented by the 12,865 variants inside the support interval of the linkage peaks reveals that there are 848
genes/open reading frames in these 5 regions. The gene names (Human Genome Organization (HUGO) gene symbols) for these 848 were converted to Ensembl Gene IDs using the Hyperlink Management System (184), and submitted to the Database for Annotation, Visualization and Integrated Discovery (DAVID) tools developed by the National Institute of Allergy and Infectious Diseases (NIAID) (185).

Figure 4.5 (a) The GVS functions of 11,444 variants out of the 12,865. Note that this is presented for descriptive purposes only; the frequencies of variants in each functional class are not corrected for the representation of that functional class in the exome, and (b) the distribution of the probability of damaging effects of 1,855 nonsynonymous variants out of the 23,865 according to PolyPhen; 348 of these 1,855 variants are annotated as probably damaging. These are distributed among 223 genes/open reading frames.

I then examined the biological functions of the genes with exome sequencing variants under the linkage peaks by asking two related questions. First, I examined the gene list to identify the major gene families that are implicated by the variants underneath the linkage peaks. Second, I examined the functional annotations of the genes to identify functional annotation terms that are enriched in the genes under the linkage peaks. 701 of the 848 genes were successfully mapped to DAVID gene IDs. An enrichment score,
calculated as the number of genes or annotations that map to a particular gene family, pathway or function, was used to quantify the relative representation of genes or annotations in this gene list. DAVID implements a modified Fisher’s Exact test to compare the enrichment score to the expectation given the total number of genes or annotations in that group, genome-wide. Due to the fact that the gene list does not represent independent effects, the enrichment results were used as a qualitative way to reduce the complexity of the list of 848 genes, not to quantify the enrichment of functional gene classes in MS susceptibility. Two gene clusters generate the highest enrichment scores in functional annotation terms: calcium-dependent (C-type) lectin-like receptors (CLECs) on chromosome 12p13.31-32 (enrichment score = 10.23, Bonferroni corrected \( P \)-value = 6.1E-11), and serine peptidase inhibitors (serpins) on chromosome 18q21 (enrichment score = 4.48, Bonferroni corrected \( P \)-value = 6.9E-4). Enrichment for specific gene families, using the highest stringency setting in classification, also identifies the CLEC and the serpin clusters, as expected. Importantly, SNPs in another CLEC (CLEC16A on 16p13) were some of the first to be associated with MS after HLA (for example (186)).

Finally, I counted the number of alternative alleles (in most cases these coincide with the minor allele, but not always depending on whether the reference genome carries the minor allele) in each individual across all CLECs on chromosome 12, and all serpins on chromosome 18, plus \( SERPINE1 \) under the linkage peak of chromosome 7 (Figure 4.6). This serves a descriptive purpose only since an assessment of whether this burden is higher or lower than that expected due to chance is not possible given the current dataset.
The natural killer cell gene complex resides on 12p13, the end of the small arm of chromosome 12 that overlaps the linkage peak. Many cellular players encoded here are reasonable biological candidates in autoimmunity. The natural killer cell gene complex has been well characterized in the syntenic region of the mouse (reviewed by (187)). A more detailed look into the exome variants in this region follows in Section 4.2.3.4, where the sub-region identified through family-based association tests under the linkage peaks is further described.

The serpins are components of the coagulation cascade with significant roles in immunity. Serine proteases are important in facilitating leukocyte entry and demyelination in active lesions (for example (188)). Their inhibition by serpins is therefore believed to be advantageous, while anything that interferes with serpin function is predicted to result in increased inflammation and acute relapses. Indeed, reduced serpin concentrations have been observed in the cerebrospinal fluid of MS patients with both relapsing-remitting MS and clinically isolated syndrome (the condition that often precedes a definite MS diagnosis) (189). An extracellular serpin of the A clade (serpin A5), also known as protein C inhibitor (PCI), is expressed in active MS lesions. This is not surprising given the fact that many molecules of the coagulation cascade are expressed in active regions of inflammation. Furthermore, serpin A5 has been used as a drug target in experimental autoimmune encephalomyelitis (EAE). Administration of activated protein C (a serpin A5 antagonist) reversed clinical paralysis in EAE mice, by inhibiting the production of T_{H1} and T_{H17} cytokines (190).

The serpin cluster implicated by the linkage peak belongs to the intracellular serpin B clade that is comprised of 13 members, some on chromosome 18 and some on
The roles of serpin Bs are better characterized in cancer where they have been implicated in the prevention of metastasis. Serpin Bs do have roles in immunity: \textit{SERPINB7} on 18q21.33 has been shown to be involved in the maturation of megacaryocytes (192). \textit{SERPINB9} is thought to protect cytotoxic lymphocytes from their endogenous granzyme B, as well as dendritic cells from cytolytic cell death upon their exposure to proteolytic peptides (reviewed by (193)). However, with this limited knowledge it would seem that damaging variation in \textit{SERPINB9} would have an advantageous effect by possibly limiting the presentation of autoantigens by dendritic cells to the immune system. More research in the functional links between damaging variation in serpin Bs and autoimmunity is needed to resolve these questions.

In addition, \textit{SERPINB3}, also on 18q21.33, has a role specifically in autoimmunity by regulating apoptosis of immune cells, which in turn influences the proliferation of autoreactive cells (reviewed by (194)). \textit{SERPINB3} has been studied in patients with

![Graphs showing distribution of alternative allele burden](image)

**Figure 4.6** The distribution of alternative allele burden for common and rare variants in the 31 cases from 5 multiplex Sardinian multiple sclerosis families in the gene families under the linkage peaks for (a) the CLEC cluster on chromosome 12, and (b) the serpin cluster on chromosome 18 plus \textit{SERPINE1} on chromosome 7.
systemic lupus erythematosus (SLE), where alterations in SERPINB3 expression are thought to contribute to autoimmunity through the inefficient disposal of apoptotic debris which may favor the rescue of autoreactive immune cells (195).

4.3.2.2 The minor allele frequencies of variants under the linkage peaks

I then sought to characterize the minor allele frequency, conservation score (GERP and PhastCons), and functional damage prediction (PolyPhen2) distributions of the 12,865 variants inside the 1-LOD-down support intervals of the linkage peaks. This information is important in the context of the increased sharing of these regions between affected relative pairs, since it may reveal information about MS susceptibility alleles that can be integrated with existing evidence and thus influence our understanding of the risk landscape.

One possibility is that variants in the identified regions have been influenced by random genetic drift in ways that are consistent with the differences observed in MS prevalence. In addition, the relatively small initial effective population size could have given rise to founder effects. For example, it is plausible that the founders of the Sardinian populations isolate were high-risk individuals that carried multiple autoimmune disease risk alleles that were passed on to their descendants. Therefore, I hypothesize the presence of allele frequency differences in the regions under the linkage peaks.

First, I compared the distribution of the allele frequencies of 9,311 variants out of the 12,865 that have a reported frequency in the SardiNIA dataset, to random sets of 9,311 variants from other locations of the clean exome (Figure 4.7). The variants underneath the linkage regions have a significantly higher mean MAF than expected if they were a random sample of exome variants (bootstrap $P$-value = 1E-04). Although it is
possible that this is a circular argument, since the regions with higher allele frequencies 
have the potential to generate higher LOD* scores because of an increase in the 
information content of those regions. However, this seems unlikely to be the case for two 
reasons. First, the information content is based on the Immunochip SNPs that comprise 
the linkage map (not the exome data that is tested here), and second, an examination of 
the information content along each chromosome (see Chapter 3) shows that it remains 
relatively constant across the genome. The variance of the MAF distribution was also 
tested, since it is reasonable to expect that the potential effects of genetic drift may be 
observed as differences in the variance of the MAF distributions rather than the means, 
such that a smaller population would have an increased variance of the MAF distribution. 
Second, I compared the MAFs of the 3,540 variants of the 12,865 under the linkage 
regions that had non-missing MAFs in both the SardiNIA and EVS European American 
datasets (Figure 4.8). The difference between the means or variances of these two 
distributions are not significantly different than the difference in the same moments for 
random draws of variants from the entire exome (bootstrap $P$-value = 0.78). However, the 
variants that are included in this comparison are the ones with non-missing allele 
frequencies in both the SardiNIA and EVS European American datasets. This excludes 
variants that are Sardinia specific (only present in SardiNIA dataset but not in EVS), and 
also excludes variants that are novel to this Sardinian dataset. Those two classes of 
variants can be argued to be the most interesting in the search for large effect, population 
specific variants that predispose to MS. However, this restriction was necessary due to 
the differences in the two datasets (for example, whole genome versus whole exome, and 
2X coverage versus >70X coverage); these differences can be expected to translate to
differences in the MAF distributions. Additionally, the wide support intervals of the linkage peaks that reach a LOD* > 1 lead to the inclusion of variants that are unrelated to MS susceptibility, adding noise that would dilute the signal.

Finally, I compared the conservation scores and damage predictions for the variants under the linkage regions to random draws of the same number of variants from the remaining exome. None of these characteristics show any deviation from the null hypothesis of no difference, as shown in Figure 4.7.

Importantly, the higher mean MAF of the variants under the linkage regions does not necessarily reflect the effects of random genetic drift. Even in the simplest, most naïve scenario where the linkage regions represent MS susceptibility regions, the remaining exome has also been impacted by the same population genetic forces.

**Figure 4.7** Bootstrap tests of the mean minor allele frequency of the variants inside the support intervals of the linkage peaks against random samples of the same number of variants from other regions of the clean exome. The red points mark the observed mean SardiNIA MAF for the variants under the linkage peaks = 0.176 (bootstrap P-value = 1E-04), and the observed variance = 0.022 (bootstrap P-value = 0.2025). The distributions are built using 10,000 re-samples drawn with replacement.
Figure 4.8 The difference in mean minor allele frequency between (a) SardiNIA and (b) EVS European American datasets, for the variants under the linkage peaks, was tested against random samples from regions outside the linkage peaks. The variants under the linkage peaks have similar (c) differences in MAF mean between the two populations (bootstrap $P$-value = 0.78) and (d) differences in MAF variance between the two populations (bootstrap $P$-value = 0.33) to variants outside the linkage regions. The dataset was restricted to variants that had non-missing allele frequencies reported for both populations.
However, MS susceptibility loci would be expected to have risen to a higher frequency, whereas the effects of these forces on the remaining exome would push the frequencies in random directions, potentially with a cumulative effect that manifests as lower mean MAF. In any case, although the results of this experiment may be in support of the hypothesis that genetic forces acting on the population have influenced the frequency of MS susceptibility alleles, they are by no means conclusive.
4.3.2.3 Genes implicated by rare, probably damaging variants

I then examined the rare, nonsynonymous variants that are predicted to be damaging in any of the genes within the support intervals of the linkage regions. To do this, I filtered the variants to keep those that are rare or absent from both the SardiNIA and EVS EA datasets (762 genes and 6,037 out of 12,865 variants), and those that are annotated as probably damaging by PolyPhen2 (163 genes and 232 variants out of 6,037). Only 19 of these 163 genes carry multiple variants from this list of 232 (Figure 4.10).

A number of biologically relevant genes show up in this list. I examined their biological functions in selecting candidates, grouped by the chromosomal region that they reside on, for two reasons. One is that only one gene or gene family is probably responsible for the signal of the linkage peak, and the surrounding genes may be hitching a ride possibly due to the presence of LD. Here I chose the gene based on biological plausibility and past implication in autoimmunity, The second reason is that since phase 1 of the 1000 Genomes project, the human genome is understood to be more robust to damaging variation than previously thought (69); therefore the presence of damaging variants in the remaining genes is not necessarily linked to disease. A more detailed discussion on these points follows in Chapter 5.

None of the genes with more than two rare, probably damaging missense variants are from the peak on chromosome 4. The two genes on chromosome 6, KIA1009 and MDN1 have no links to neuroinflammation or autoimmunity in the literature. On chromosome 18, the PHLPP1 gene encodes for a serine/threonine phosphatase that functions as a tumor suppressor in multiple types of cancer. Its role in innate immunity has been explored in the context of cancer where it is implicated in the dysregulation of
macrophage responses, but there is no clear evidence of a potential link to autoimmunity or neuroinflammation. Similarly, the links to MS-related mechanisms for the genes on chromosome 7 are weak. These include the zinc finger transcription factor (ZNF) cluster, two mucins (MUC3A, MUC17), ASL, POM212, ABCA13 and COBL.

![Figure 4.10](image)

**Figure 4.10** Characterization of the missense variants under the linkage peaks that are annotated as rare or absent in both the SardiNIA and EVS EA datasets. (a) 19 genes carry > 2 of these 232 variants, and (b) the GERP score distribution of all 232 variants.

An interesting result from the analysis of rare, probably damaging variants under the linkage peaks arise on chromosome 12. The von-Willebrand Factor (VWF) gene encodes a glycoprotein with a role in the coagulation cascade. Some of the molecules involved in coagulation have additional roles in innate immunity by increasing the permeability of the vascular endothelium and acting as chemotactic agents for phagocytes. In fact, the absence of VWF has been shown to increase the permeability of the blood-brain barrier and lead to more severe forms of experimental autoimmune
encephalomyelitis (EAE), a mouse model of MS (196). This provides a possible mechanism that links damaging variation in \( VWF \) to MS pathogenesis.

One of the remaining genes on chromosome 12 (\( CACNA1C, SCNN1A \) and \( CD163L1 \)) is also involved in an autoimmune disease. \( CACNA1C \) and \( SCNN1A \) encode subunits of ion channels. \( CD163L1 \) is a member of the scavenger receptor cysteine-rich superfamily found in immune system cells; it is upregulated in preclinical autoimmune arthritis, where it may promote inflammation of the arthritic joints (197).

### 4.3.2.4 Genes implicated by rare, probably damaging variants that have risen in frequency in Sardinia

Finally, I looked at the variants that are rare in an outbred European ancestry population but have a higher frequency in Sardinia. To this end, I limited the 12,865 variants inside

![Figure 4.11](image)

**Figure 4.11** Characterization of the 63 variants in 58 genes that are rare in the EVS EA dataset, have a higher frequency in Sardinia, and are annotated as missense and probably damaging by PolyPhen2. (a) Most of the variants with a higher MAF in SardiNIA do not show large allele frequency differences. The variants with the highest differences are in the \( MYC17 \) gene. (b) The distribution of the GERP scores.
the linkage regions to those that are annotated as rare (MAF \leq 0.05) in EVS EA (1,394 out of 12,865) and have a higher frequency in SardiNIA than in the EVS EA dataset (701 variants out of 1,394, in 368 genes). A total of 63 of these 701 are predicted to be probably damaging (in 58 genes). A set of 5 genes carry more than just one of these 63 variants: \textit{BCL2L14}, \textit{FOXM1}, \textit{CD163L1}, \textit{MUC17}, and \textit{MUC3A}. The mean difference in MAF between the two populations for these 5 genes is 2.5%. The distribution of the differences and conservation scores of the missense variants across the 58 genes is shown in Figure 4.11.

\textbf{4.3.2.5 Variants implicated by fine-mapping under the linkage peaks}

Fine-mapping under the linkage regions using a family-based association test produced nominally significant associations on 4 of the 5 chromosomes (6q16.1, 7p15.1, 12p13.2 and 18q22.2, see Chapter 3). On the chromosomes where the top fine-mapping hit fell outside of a gene, the closest flanking genes were examined for rare variants in the exome sequencing data. Although methods to statistically analyze rare variants in family based designs do exist (198, 199), these are based on trios and therefore cannot be used with the present study design.

On chromosome 6, there is only 1 gene within 500KB on either side of the top fine-mapping SNP. This is \textit{EPHA7} (erythropoietin-producing hepatoma (EPH) receptor A7), which encodes a receptor tyrosine kinase with a multitude of functions including cell adhesion, cell migration, and a well-studied role in axon guidance (reviewed by (200, 201)). Network-based analysis of modest association signals in MS GWA studies has identified a neuronal module that contains two ephrins and three EPH receptors (\textit{EPHA3}, \textit{EPHA4}, and \textit{EPHB2}) (202). In functional studies, there is evidence to support the
regulation of T cell migration by ephrins and EPHA receptors in response to chemotactic signals through extensive signaling cascades (203). Ephrins and EPHA receptors are specifically implicated in the immunopathogenesis of active white matter lesions in MS and the subsequent neurodegeneration. A number of cells at the lesion sites (perivascular mononuclear inflammatory cells, reactive astrocytes, macrophages) have been shown by immunohistochemistry to express ephrins/EhpAs including *EPHA7*. Axons, reactive astrocytes and oligodendrocytes in particular show increased expression of *EPHA7* in sites of active lesions and their surrounding normal-appearing white matter, in comparison to chronic lesions and control samples (204). The upregulation may negatively impact axon regeneration, and also regulate the formation of the glial scar. Therefore, targeting ephrins/EPH receptors may be useful in modulating the impact of injury in the central nervous system.

In the exome sequencing data, there are 21 variants in the *EPHA7* gene, 3 of them are nonsynonymous, and 9 of them are rare (MAF < 0.02, 2 of these 9 are novel). There is one low-frequency variant (rs62414217, SAR MAF = 0.09, 1000Genomes pilot 1 CEU MAF = 0.067) that is observed in more than 2 affected individuals. A total of 7 out of 31 affecteds carry the minor allele of rs62414217, an intronic variant with no evidence of conservation. Of the nonsynonymous variants in *EPHA7* one is also novel, however it is observed in a single affected individual and called with high confidence (GQ score > 99, depth = 93X). It is predicted to be probably damaging and also highly conserved (hg19 position 93969170, GERP score = 5.9). A second missense variant, at hg19 position 94120219 is predicted to be probably damaging and highly conserved (GERP score =
It is rare in both populations (SAR MAF = 0.016, EVS European Americans MAF = 0.03), and present in the heterozygous state in two affected individuals.

On chromosome 7, there are 30 variants in the exome sequencing data in \textit{JAZF1}. Only 1 of them is a missense variant, it is functionally predicted to be benign and has a low conservation score. This variant along with another 15 variants out of the 30 are novel, and 11 out of these 16 novel variants are singletons. An additional 8 variants out of the 30 have a SardiNIA MAF < 0.05; however, none of those are present in more than two affecteds.

On chromosome 12, the \textit{KLRC2} gene is the closest one to the fine-mapping peak (< 500MB away). \textit{DDX12P} is a similar distance from the peak on the other side, but it is a pseudogene of a transcription factor with no biological connection to MS pathogenesis. On the other hand, Killer Cell Lectin-like Receptor subfamily C (\textit{KLRC2}) encodes a protein expressed on natural killer cells with a C-type lectin domain such as the one found in the CLEC\textsubscript{S} of the natural killer cell gene complex in this region. It recognizes MHC class I HLA-E molecules. The expression of HLA-E molecules is significantly higher in white matter lesions in MS patients compared to white matter from healthy controls \cite{205}, providing functional insight into the mechanism that implicates natural killer cells in MS pathogenesis. There are 25 variants in \textit{KLRC2}, 7 of them are rare in SardiNIA (MAF < 0.05). 3 of the 25 are benign missense variants with frequencies that range from 0.001 to 0.07 in SardiNIA and similar frequencies in outbred Europeans. An additional missense variant is predicted to be probably damaging, and has a very low frequency in both SardiNIA (MAF = 0.008) and EVS datasets (MAF = 0.002).
variant (rs145844403) is observed in two affecteds from different families and it has a moderate conservation score (GERP = 2.98).

**Figure 4.12** The burden of alternative alleles for both common and rare variants in the genes implicated by fine-mapping underneath the linkage peaks for 31 affecteds from 5 multiplex Sardinian multiple sclerosis pedigrees. (a) the *EPHA7* gene on chromosome 6, (b) *JAZF1* on chromosome 7, (c) *KLRC2* on chromosome 12, and (d) *CD226* on chromosome 18.

Finally, the chromosome 18 fine-mapping signal directly implicates Cluster of Differentiation 226 (*CD226*), which is also a gene expressed on the surface of natural
killer cells. It has been previously implicated in MS through genetic associations (for example (206)) and it has multiple functions in inflammation. It is important in the control of the activation and proliferation of proinflammatory Th17 T- cells and the balance between pro-inflammatory (Th1/Th17) and anti-inflammatory (Th2) T-cell populations (207). CD226 is also important in the transendothelial migration of immune cells (208). Variants in CD226 may therefore impact the permeability of the blood-brain barrier to activated leukocytes, providing a second mechanistic link to MS pathogenesis.

In the exome data, there are 12 variants in CD226, 10 of which are rare or low frequency (< 0.08 in either SardiNIA or EVS EA). Only two are missense, and their effects are predicted to be benign. In contrast to many of the variants observed in other genes, none of them are singletons (i.e. present in a single sample). Most of the affecteds in the dataset carry 6 or 7 alternative alleles. The burden in each individual is influenced by the allele frequencies of the variants as well as the relatedness between the affecteds, making it difficult to assess the potential impact of this clustering of alternative alleles in the affected individuals.

4.2.3. Summary

The approaches taken in the analysis of the exome sequences of the affected individuals from multiplex Sardinian families have identified partially overlapping sets of genes. Multiple pieces of evidence cumulatively support their roles in MS, including their biological plausibility, prior involvement in autoimmunity, and sharing between distantly related affecteds. There are few instances of rare coding variants that may have a severe impact on gene function (for example rs145844403 in KLRC2), while most of the genes carry variants of unknown consequences. Cumulatively, these could have a net negative
effect on protein function and/or regulation. These results provide a platform for
subsequent studies to replicate these candidate effects in a way that has higher statistical
power than an agnostic association approach which would suffer greatly from multiple
testing correction penalties. Furthermore, it would be interesting to investigate the
function and potential role of these variants in autoimmunity and neuroinflammation.
5

Discussion

5.1 Summary of findings

The genetic factors that influence MS susceptibility in Sardinia were investigated from three different but related directions. Multiple hypotheses can be formulated to explain the extraordinarily high prevalence of MS in Sardinia in comparison to other outbred European populations. The relative contributions of common and rare variants to the overall risk landscape are an important parameter that defines the continuum on which most of these hypotheses lie. Previous studies have attempted to quantify the contributions of common variants, but the absence of positive results indicates that they were probably underpowered. In addition, the demographic history of this population isolate in combination with the epidemiological observations about autoimmune diseases
may lead to the hypothesis that there is a distinct set of autoimmunity risk alleles operating in this population, with effects that are larger than those observed in populations with a lower prevalence of autoimmunity. Therefore, in order to examine the contributions of common variants to MS susceptibility in Sardinia, I built a cumulative genetic burden score based on the currently available list of 110 MS-associated SNPs and tested that for association to disease. The risk allele frequencies of these 110 markers in the derivation population range from 0.09 to 0.95 (mean = 0.51, SD = 0.23). Furthermore, I used the burden score to quantify the relative involvement of the Northern-European-discovered risk alleles by testing for association to population of origin.

The results indicate that Sardinian cases are more heavily loaded for these 110 risk alleles than Sardinian controls, underlining the relevance of the underlying MS susceptibility alleles tagged by these markers in the Sardinian dataset. Sardinian cases are also more heavily loaded for the Northern-European-derived risk alleles than a dataset of US cases with European ancestry. This could point to Sardinian individuals carrying more risk alleles than US cases, and could be argued to be one of the reasons for the increased prevalence of MS in Sardinia. However, it is crucial to examine this observation in the context of previous studies which have shown that MS cases from multiplex families carry an increased genetic burden compared to sporadic MS cases. The corrections for the correlations between relatives that were employed in this study lead to an appropriate adjustment of the variance. On the other hand, the genetic burden in the cases from multiplex families will not have been adjusted for the fact that they originate from multiplex families, while the samples from the comparison population of US cases are sporadic. Therefore, the increased MS genetic burden in the Sardinian cases
compared to US cases is expected. A similar argument can be made for the observation that Sardinian controls carry heavier loads of MS risk alleles than US population controls. Indeed, unaffected parents of MS cases have been shown to carry a heavier load for MS-risk variants using previous versions of the MS genetic burden score. In this case, however, the effects of membership in multiplex families are expected to have been attenuated by the inclusion of unrelated Sardinian population controls, although perhaps not enough to eliminate the effect of the membership of a subset of these unaffected individuals in multiplex families.

If Sardinian controls do carry a heavier load for MS risk alleles, this would be in line with the hypothesis that the population is enriched for high risk individuals in general. The results of the comparison of the risk allele frequency distributions in SardiNIA samples and UK controls from the IMSGC Immunochip dataset contradict this hypothesis, since enrichment for high risk individuals may be observed in the overall distribution of the risk alleles. A possible model that is consistent with the results of this test is that the risk alleles that are present at higher frequencies in Sardinia are sampled in place of the ones that are rarer, so that overall, the individuals carry sufficiently loaded multilocus genotypes to manifest disease, and this occurs at a frequency that is similar in the two populations. This is consistent with the reduction in the heterogeneity of the multilocus genotype that is predicted for an isolated population.

On the other hand, four of the SNPs were significantly different between Sardinian and US cases in our dataset, after correction for both relatedness and multiple testing. Three of these four have a higher frequency in Sardinia than they do in the US cases ($\Delta$MAF range = 0.17 to 0.2), and the other one has a lower frequency in Sardinia
It is tempting to formulate the hypothesis that Sardinian MS cases carry more risk alleles than the ancestral population of outbred Europeans since this would mirror observations of higher levels of deleterious variation in coding regions in European Americans compared to African Americans by Lohmueller et al. The genomes of European Americans, due to factors such as the out-of-Africa bottleneck and purifying selection, carry increased levels of deleterious variation than the older, more heterogeneous genetic background of African Americans. In the Lohmueller et al. study, the result was largely due to variants that became fixed for the derived allele in the European Americans, since considering only the variants that were polymorphic in both populations eliminated the difference. It is possible that the frequency of MS risk alleles in Sardinia has been shaped in a similar way, so that it is higher than in the ancestral population. However, the distribution of the differences in risk allele frequencies between the two populations provides a glimpse into the magnitude of the population genetic forces that have acted in the population isolate. None of the common MS-associated alleles have been driven to extinction or fixation, an observation which supports the hypothesis that Sardinian individuals do not carry increased numbers of deleterious variants for the polymorphic loci. This is in line with the known demographic history of the population, since random genetic drift is not expected to have pushed common alleles to extinction or fixation given the relatively large effective population size of Sardinia. In addition, what is known about common MS risk alleles so far indicates that

\( \Delta \text{MAF} = -0.18 \). It follows that overall, the frequencies of the multilocus genotypes that the risk alleles end up on may be higher, providing the foundation for a model where the frequency of high risk individuals may be higher.
their effects are small, so selective pressures that have been shown to have shaped the
distribution of deleterious variation in the genomes of European Americans will not have
acted on them. In other words, the results from Lohmueller et al. may not be applicable to
MS risk alleles since many of them are outside of coding regions and their effects may
not be under similar levels of purifying selection. On the other hand, the effects of drift
that are observed on the common alleles would be pronounced with respect to rare
alleles, perhaps pushing those to extinction or fixation. One possible scenario is that the
founders were enriched for rare MS-associated alleles (209). Simulation studies that
model the demographic parameters of this isolate and the epidemiological parameters of
MS could show how the behavior of rare alleles through time would impact the
epidemiology of MS. In addition, the larger effect sizes expected for the rarer risk alleles
might have translated to increased levels of purifying selection acting on them, making
the results of Lohmueller et al. directly relevant.

In any case, the variants that tag MS risk in populations of Northern European
ancestry are present in Sardinians. Although an examination of the multilocus genotypes
and a detailed description of the numbers and effect sizes of risk alleles that are sufficient
to lead to disease is an attractive endeavor, an exploration of this avenue has
demonstrated that this analysis was severely underpowered given the sheer number of
possible combinations of risk alleles. The observation that Sardinian controls carry
similar numbers of the risk alleles to Sardinian cases, in combination with the absence of
any significant differences in allele frequencies even for this small number of tests of
previously associated MS-risk alleles is most likely a consequence of their small effect
sizes. An alternative explanation could be that the tagging effects of the established
associations in samples of Northern European ancestry are lost in Sardinians due to differences in the patterns of LD between the populations. In addition, the reduction in heterogeneity in the multilocus genotypes of the Sardinian cases is most likely due to the relatedness in the sample. Alternatively, this could be a direct observation of the homogeneity expected in a population isolate. Since this expected homogeneity is due to interbreeding between individuals from a small group of founders representing a subset of the ancestral haplotypes, it also reflect relatedness albeit in a more diluted way, so perhaps the two reasons are not as different as they may initially appear.

It is noteworthy that despite the increase in the number of SNPs that are included in the MS genetic burden score, its predictive ability is not improved relative to previous assessments that used a smaller number of MS-associated risk alleles (112). This may be because in past studies, the burden scores were tested in their ability to predict disease status in the derivation population.

Support for the known MS-associations having a larger role in Northern Europeans than in Sardinians also comes from the observation that none of the MS-associated SNPs that comprise the burden score are significantly different in allele frequencies between Sardinian cases and controls. The lack of significant results in the single marker tests between Sardinian cases and controls does not contradict the observed significant difference in the MS-genetic burden score between Sardinian cases and controls, since the burden score reflects the cumulative effect of the risk alleles which independently do not show evidence of association.

Complications in the discriminatory power of the burden score using the current list of 110 non-MHC associations versus the shorter list previously used may be related to
the lower mean effect size of the current list. Given the sample sizes that consortia-sized studies are able to accumulate, the ORs of the associated SNPs that can be reliably detected have dropped down to 1.1 and in some cases even a little lower. This translates to an increase in the levels of noise that are introduced in the burden score, since unaffecteds are even more likely to carry the risk alleles than they would be for SNPs with larger effect sizes. The high frequency of some of the risk alleles in this list of 110 (some of them go up to 95% in the unaffected population) demonstrates how this increase in both affecteds and unaffecteds can be substantial. Calculations of the increase in risk associated with specific multilocus genotypes would be useful in mitigating the effects of this risk. Unfortunately, given the low frequency of each of the multilocus genotypes this may be underpowered even in consortia-sized datasets.

I took multiple approaches to search for evidence that links previously undiscovered variation to MS pathogenesis in Sardinia. I used a set of multiplex families from this homogeneous, isolated population to examine regions that are shared IBD between affected individuals from a common ancestor. Given the distant relationships and the number of affected relative pairs in this dataset, these regions could harbor variation related to the shared MS disease phenotype. I carried out analyses following two distinct courses.

In one approach, I used IBD estimates from each family separately to identify exome sequencing variants in the genes implicated by the 110 established MS-associations within the regions that are shared. Coding variants in two genes, MANBA and IKZF1, stand out as probably damaging, rare in outbred Europeans but with a higher frequency in Sardinia, and highly conserved. IKZF1 is located at 7p12, 18KB from the
intergenic Immunochip hit at rs201847125. This same locus has also shown increased levels of sharing in yet another dataset of 49 Sardinian sibling pairs. Taken together, these pieces of evidence suggest that variation in \textit{IKZF1} may be of interest for further study. If the Immunochip association at this locus is in reality implicating an intergenic regulatory variant, it is possible that variation in \textit{IKZF1} is unrelated to the phenotype. Further studies would be needed to provide additional evidence for the involvement of \textit{IKZF1} in MS.

Our second approach used dataset-wide non-parametric linkage LOD* scores to identify variants underneath the linkage peaks that may be related to the phenotype. A summary of the multiple approaches I took in searching for these variants may be aided by outlining the hypotheses that underlie each one.

First, in order to reduce the complexity in the large list of genes under the support intervals of the non-parametric linkage peaks that reached a LOD* > 1 I examined the functional classes that are over-represented in this gene list. Two clusters, one on 12p13 and one on 18q21 stood out: the CLECs and the serpins. Genes from both families are reasonable biological candidates based on the literature and past MS-associations, as described in Chapter 4.

Second, I looked for variants under the linkage peaks that are likely to have a considerable impact on the function of the protein product. To do this, I limited the variants in the support intervals of the linkage peaks to those that are predicted to be probably damaging, and rare in both an outbred European population and in unaffected Sardinians. Two genes stand out using this approach: \textit{VWF} and \textit{CD163L1}. There is a substantial body of literature that supports the biological role of \textit{VWF} in MS.
pathogenesis, and some evidence that implicates \textit{CD163L1} in systemic lupus erythematosus (SLE), another autoimmune disease.

Third, I limited the variants in the confidence intervals of the linkage peaks to those that are predicted to be probably damaging and rare in an outbred European population but have a higher frequency in unaffected Sardinian population samples. This follows the hypothesis that the increased prevalence of MS in Sardinia may be in part due to the presence of otherwise rare, damaging variants that have higher frequencies in this population, possibly due to a founder effect. Two genes stand out, \textit{BCL2L14} and \textit{CD163L1}, both of which are in the natural killer cell gene cluster on 12p13. Like the aforementioned \textit{CD163L1}, \textit{BCL2L14} has been implicated in SLE, specifically in the dysregulation of the apoptosis of T-cells. The prolonged activation of pro-inflammatory T-cells in MS could also have an adverse effect on the phenotype; conversely, the prolonged activation of anti-inflammatory T-cells could have a protective effect in neuroinflammation.

Fourth, I examined exome sequencing variants in the specific genes implicated by the fine-mapping association analysis. In the non-parametric linkage scan, the families in this dataset show increased sharing between affecteds of the 7p15 band, which overlaps with one past linkage scan in an independent dataset of 28 Sardinian sibling pairs. The 7p15 region shows homology to the rat region that harbors the EAE locus. This same region produced the top hit in the fine-mapping association analysis on chromosome 7; this signal falls in the \textit{JAZF1} gene. The \textit{JAZF1} gene carries an intronic MS-associated SNP (rs917116, OR = 1.12, 95% CI (1.07, 1.16), \textit{P}-value = 2.07E-8), established in the IMSCG dataset. Given the evidence provided by the exome sequencing data, the
increased sharing of this region in these families is probably not due to rare, coding, probably damaging non-synonymous variants that segregate in JAZF1. Multiple factors may explain this observation. Some intronic variation is captured by the exome capture kit, and some of the rare, intronic variants in JAZF1 that are observed in this dataset are highly conserved, indicating that they may have important functions. Although these do not segregate in the affecteds and they are observed as singletons or doubletons, they are all in different individuals, perhaps because carrying any one of these rare, conserved, intronic variants is what affects risk to MS. Further studies are needed to decipher the importance of these variants in MS.

Aside from the chromosome 7 peaks, the linkage scan shows little overlap with past scans from any population. This is not surprising as it mirrors all past linkage scans in MS, which produced very little overlapping results. In addition, no genes in the chromosome 4 peak are implicated by any of the subsequent analyses. This may be explained by the narrow interval of this peak, driven primarily by family 212, and the low marker density used in the fine-mapping analysis. Fewer genes fall inside this region than in the other peaks, therefore it is expected that fewer genes in this region may be carrying rare, damaging variants, or be biologically reasonable candidates in MS pathogenesis.

The genes implicated by the fine-mapping hits on chromosomes 6, 12 and 18 (EPHA7, KLRC2 and CD226 respectively) also carry no rare, probably damaging variants that segregate in all affecteds. This supports the idea that the overall burden of mildly deleterious variants may be the primary driver of susceptibility. Although it was possible to describe this burden in the affecteds, the lack of exome sequencing data on the parents or other unaffected individuals prohibits their comparison to the burden in
controls and any inference with regards to the significance of the observed burden in these genes.

It is of interest to note that the HLA region at 6p21 achieved a LOD* score of 0.5. This mirrors past linkage scans in Sardinia that have not reproduced the result in the HLA region. In order to further examine what underlies this result, I could have stratified the linkage analysis by using just the HLA negative families or used the HLA risk haplotypes in an ordered subset analysis (OSA). However, these approaches were unlikely to have helped given the HLA profiles of the Sardinian families.

Lastly, with respect to yet unidentified variation that may be related to MS, I broadly examined the support intervals of the non-parametric linkage peaks that reached a LOD* > 1 for genes that are reasonable biological candidates but did not pop up in any of the other analyses. One of the genes that may be notable in this list is PRMT8. It is also part of the natural killer cell gene cluster and its coding sequence is highly conserved. Following the trends observed in the aforementioned genes, there are no non-synonymous, rare variants in PRMT8 within this dataset.

I also examined the evidence for a founder effect in the variants under the linkage peaks that could manifest as allele frequency differences. To this end, I looked at the distribution of allele frequencies in the linkage regions compared to the rest of the exome, to examine the hypothesis that Sardinians carry higher frequency of minor alleles in the MS regions relative to other regions of their exome. The mean MAF for the variants under the linkage peaks is significantly higher than for variants in other regions of the exome. This is in line with an overrepresentation of minor alleles in the MS regions,
which could be because of an overrepresentation of the minor alleles in the linkage regions in the founders.

An implicit assumption of this analysis is that the minor allele is the risk allele; this assumption is violated for many of the common SNPs that are associated with MS, as evidenced by the frequencies of the risk alleles in the UK controls of the Immunochip experiment. However, this comparison includes mostly rare alleles that are predicted to have slightly larger effects, and for which the minor allele may indeed be the risk allele. The observation that the linkage regions carry a higher mean frequency of minor alleles may reflect either the presence of fewer rare variants, or the presence of more, common variants. If a representative subset of these is indeed related to MS pathogenesis, figuring out which of these alternatives generated this result could be important in shaping the susceptibility landscape of MS in the Sardinian population.

Furthermore, these analyses were completed using unaffected population samples, which dampen the signal from true MS susceptibility alleles. The observation of a difference in the unaffecteds, mirrors observations from Chapter 2, where the Sardinian controls were seen to carry a relatively high burden of the known MS-associated risk alleles, comparable to that of the outbred, sporadic US cases. On the other hand, the Sardinian controls in this dataset may be different than the unaffected population samples from the SardiNIA datasets, because some of them are related to cases in multiplex families. Finally, it is somewhat surprising to see a significant difference in the mean of the MAF distributions between regions of linkage and regions of no linkage given the large numbers of alleles that are sampled for this analysis, since it wouldn’t be unreasonable to expect the few true susceptibility alleles to be drowned out by the alleles
that are irrelevant to the phenotype. Perhaps the observation that the susceptibility alleles are providing enough of a signal to push the mean allele frequency higher is a reflection of their large numbers. This hypothesis is in line with the en masse analyses of MS that estimates the number of susceptibility alleles to be several hundreds.

In order to assess whether more of the founders carried variants in the linkage peaks compared to the European population, I examined the difference in the mean minor allele frequency of the variants under the linkage peaks between Sardinians and outbred Europeans for the variants under the linkage peaks compared to the variants outside the linkage peaks. Based on this analysis, there is no difference in the differences in minor allele frequencies between the two populations, indicating that the variants under the linkage peaks have had no overall influences that would be detected using this approach. This analysis does not capture finer level detail, including information that may point to the MS-related subset of this variation having different frequencies in Sardinia. One hypothesis is that the risk variants are at an increased frequency and the protective ones at a lower frequency, generating a net effect that shows no difference. A second alternative may be that the variants under the linkage peaks are be unrelated to the higher prevalence of MS in Sardinia. The use of unaffected individuals provides an opportunity to comment on the results of this analysis only with respect to the frequencies of possible MS susceptibility alleles in the Sardinian population compared to outbred populations, and not with respect to the frequencies of those variants in cases.

Studies suggest that exomes are a good place search for disease associated variation. One reason for this is the observation that an enrichment of association signals for the seven diseases covered in the Wellcome Trust Case Control Consortium
(WTCCC) is found in exons, genes in general, and also their flanking regions (40kb in either direction) \( (210) \). The lack of overwhelming evidence for variants of large effects segregating through these pedigrees does not mean that the exome variation that is observed in this study is not relevant. The observational study design used in the present study enables the broad assessment of variation in regions that are implicated in disease by virtue of being shared between affected relatives. Although the rare variation in known MS genes being shared in all affecteds within a family is a promising result (for example the \( MANBA \) and \( IKZF1 \) variants in family 61), it is still challenging to make definitive statements about their involvement in MS susceptibility.

5.2 Synthesis

Overall, the data supports the prediction of extensive allelic heterogeneity as a major feature of the genetic architecture of MS in Sardinia. Demographic reasons provide a foundation for exploring the hypothesis that rare variants of large effects are present at a higher frequency in Sardinia, explaining the epidemiological observation of extraordinarily high disease prevalence. However, this study has generated limited evidence to support this idea. Instead, it seems that the general MS risk profile in Sardinia is not as different from outbred European populations as perhaps previously thought. It has been suggested that the generally negative findings of family-based linkage studies are not necessarily incompatible with rare variants models at least partially underlying disease susceptibility, although it is unlikely that these rare variant models are oligogenic \( (70) \). In either case, the results of these analyses are consistent with a primary role for cumulative, small and moderate effects of genetic variation driving MS susceptibility in the Sardinian population.
Rare, damaging variants in genes that are reasonable biological candidates are present in these Sardinian families but most of these are observed in just a few affected individuals. The magnitude of their effects and therefore their relevance in MS requires further investigation. Functional studies into the mechanism by which these variants may affect MS susceptibility will be useful. However, functional studies would by necessity focus on the effects of one or at best a few of the risk alleles; therefore potentially overlooking net effects of their combinations into multilocus risk genotypes.

There appear to be two unifying themes in the immune-related genes that are implicated by the analyses presented in this dissertation. One of them is the involvement of natural killer cells. Aside from the natural killer cell gene cluster on 12p13, CD226 from 18q22 is also expressed on the plasma membrane of natural killer cells, platelets, monocytes and some T-cells. The second theme is made up of genes that are involved in the transmigration of leukocytes across the blood-brain barrier. The VWF gene on 12p13, as well as CD226 have suggestive evidence supporting roles in this process.

Furthermore, the present study highlights the importance of examining multiple sides of the overall hypothesis. Each approach identified a different set of reasonable biological candidates. Replication in an independent dataset with adequate statistical power is needed to assess the strength of these approaches to detect MS-relevant variants. In addition to a first-pass assessment of the association of individual variants with disease status in a potential replication dataset, it would be interesting to calculate a poly-variant burden score across all implicated loci and assess its association to disease status. Further exploration of yet undiscovered rare and common variants in the regions implicated by the linkage peaks and a comparison of the burden in these candidate regions between
cases and controls would also generate evidence to further validate or refute their involvement in MS.

Overall, the data underlines the importance of the net effects of multiple variants acting cumulatively. Examining this in a larger more well powered dataset will be more informative than an interrogation of individual variants out of that context. Furthermore, despite the fine-mapping peaks producing reasonable biological candidates, additional genes under the linkage peaks may play roles in MS pathogenesis. This possibility is routinely addressed by conditional association analyses that search for secondary independent signals in regions that have established primary associations, and could be true for the genes under the linkage peaks as well. Burden tests in population studies of rare variants also address the net effects of variants in the gene that is being tested. In any case, these results are in line with the evidence that has accumulated in the field so far, which indicates that the overall balance of events taking place in multiple cell types will determine the outcome with respect to demyelination, inflammation and axonal degeneration.
References


10. K. Cunnusamy et al., Disease exacerbation of multiple sclerosis is characterized by loss of terminally differentiated autoregulatory CD8 T cells. Clinical Immunology (2014).


74. I. A. Eaves et al., The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. Nat. Genet. 25, 320-323 (2000).


116. R. Lampis et al., The inter-regional distribution of HLA class II haplotypes indicates the suitability of the Sardinian population for case-control association studies in complex diseases. Hum. Mol. Genet. 9, 2959-2965 (2000).


143. M. Pitzalis et al., Genetic loci linked to type 1 diabetes and multiple sclerosis families in Sardinia. BMC Medical Genetics. 9, 3 (2008).


146. S. Purcell et al., PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559-575 (2007).


175. A. McKenna et al., The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297-1303 (2010).


