ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

INVESTIGATING LOW-FREQUENCY AND RARE GENETIC VARIANT CONTRIBUTION TO MULTIPLE SCLEROSIS SUSCEPTIBILITY IN MULTIPLEX TURKISH MS FAMILIES

FURKAN BÜYÜKGÖL

M.Sc. THESIS

DEPARTMENT OF BIOSTATISTICS AND BIOINFORMATICS

SUPERVISOR
Prof. Osman Uğur Sezerman

SECONDARY SUPERVISOR
Prof. Eda Tahir Turanlı

ISTANBUL-2025

ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

INVESTIGATING LOW-FREQUENCY AND RARE GENETIC VARIANT CONTRIBUTION TO MULTIPLE SCLEROSIS SUSCEPTIBILITY IN MULTIPLEX TURKISH MS FAMILIES

FURKAN BÜYÜKGÖL

M.Sc. THESIS

DEPARTMENT OF BIOSTATISTICS AND BIOINFORMATICS

SUPERVISOR
Prof. Osman Uğur Sezerman

SECONDARY SUPERVISOR
Prof. Eda Tahir Turanlı

ISTANBUL-2025

DECLARATION

I declare that this thesis work is my own work, I had no unethical behavior at any stages from the planning to the writing of the thesis, I obtained all the information in this thesis in accordance with academic and ethical rules, I cited all the information and comments that were not obtained with this thesis work, and I provided resources in the list of references. I also declare that there was no violation of any patents and copyrights during the study and writing of this thesis.

25.06.2025

Furkan Büyükgöl

PREFACE AND ACKNOWLEDGEMENT

This thesis is submitted in partial fulfillment of the requirements for the degree of Master of Biostatistics and Bioinformatics at Acıbadem Mehmet Ali Aydinlar University Institute of Health Sciences. The research presented in this work was done during the period from 2022 to 2025 under the supervision of Prof. Osman Uğur Sezerman and Prof. Eda Tahir Turanlı.

The motivation for this research stems from my passion and interest in genetics and using computational and statistical tools to entangle the complicated web of causal chains in complex traits and disorders. Throughout this journey, I have gained invaluable experience and insights, both academically and personally. This work reflects my dedication and passion to exploring and discovering underlying complex genetic architecture of human traits that can potentially be translational for the well-being of patients.

I would like to express my deepest gratitude to everyone who has supported me throughout this journey. First and foremost, I am sincerely thankful to my supervisors - Prof. Osman Uğur Sezerman and Prof. Eda Tahir Turanlı - for their guidance, expertise, and encouragement. Their generous support, both academically and personally, has been invaluable to my personal and academic growth during my master's studies.

I am especially grateful to Berk Gürdamar and Alper Bulbul for their dedicated collaboration on this project and their insightful contributions as co-authors of the resulting paper. Working with them has been an inspiring and rewarding experience.

I also want to acknowledge the members of the Sezerman lab and the Eda Tahir Turanli lab for their warm welcome and the enriching academic environment they provided. Their willingness to share knowledge and experiences significantly enhanced my learning process. In particular, I would like to thank Berkay Aslan, my fellow master's student in the Sezerman lab, for his invaluable friendship and support.

Our movie days played a crucial role in maintaining my well-being throughout this journey.

I am grateful to my close friends from the "Bordeaux guys" group and the "Serbo watch party" for their unwavering friendship and the joy they brought into my life during challenging times.

I would also like to express my heartfelt appreciation to Dr. Jason Flannick for offering me the opportunity to intern during my bachelor's studies, guiding me through topics in statistical genetics, and continuing to mentor me as I advance in my career. His mentorship has been instrumental in shaping my skills as a researcher.

I would like to thank TÜBİTAK for their financial support through the 2210-A MSc program, which made it possible for me to pursue my research and studies.

Lastly, I want to thank my family for supporting me at every step of my academic journey. I am especially grateful to my sister Feyza and her family, particularly my lovely niece Sara, for bringing so much joy into our lives. I am deeply thankful to my elder sister, Büşra, for welcoming me into her home during my master's studies, making it possible for me to live in Istanbul when I couldn't afford housing on my own. Her generosity, support, and encouragement have been invaluable. Most importantly, I want to thank my parents for being the best parents in the world. Their unwavering support and belief in me, no matter how ambitious or unconventional my plans, mean everything to me. I love them dearly.

This thesis would not have been possible without their guidance, encouragement, and unwavering support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Investigating Low-frequency and Rare Genetic Variant Contribution to Multiple Sclerosis Susceptibility in Multiplex Turkish MS Families

Multiple sclerosis (MS) is an immune-mediated disorder of the central nervous system, characterized by chronic inflammation, demyelination, and neurodegeneration. While genetic risk factors for MS have been extensively studied through both family-based and large-scale analyses, the contribution of low-frequency and rare variants (LFRVs) remains less understood, partly due to the overrepresentation of European samples in most MS studies. This thesis investigates the role of LFRVs in MS susceptibility using one of the largest family-based MS cohorts in Turkey, comprising 215 individuals from 59 families with multiple affected members. Whole exome sequencing was performed on all participants, including both affected and unaffected relatives. We first examined the human leukocyte antigen (HLA) loci due to their established role in MS. Next, we conducted segregation analysis to identify variants co-segregating with MS under various inheritance and dominance models. A family-based burden analysis was also performed to complement the segregation approach. Pathway enrichment analysis was employed to interpret the functional significance of candidate variants and genes. Our results underscore the potential involvement of the extracellular matrix and blood–brain barrier in MS pathogenesis, particularly through laminin-related genes (LAMA5, LAMB1) and structural genes (DST, PLEC, COL16A1). This study represents one of the first and most comprehensive family-based genetic investigations of MS in the Turkish population, offering valuable insights into LFRVs and their contribution to disease risk. The discovery of novel variants and pathways may help expand our understanding of MS biology and guide future therapeutic strategies.

Keywords: Multiple sclerosis, Genetic burden, Segregation, Exome, HLA

ÖZET

Türk MS Ailelerinde Aile Tabanlı Yaklaşımla Multiple Skleroz Riskiyle İlişkili Düşük Frekanslı ve Nadir Genetik Varyantların Araştırılması

Multipl skleroz (MS), merkezi sinir sistemini etkileyen, kronik inflamasyon, demiyelinizasyon ve nörodejenerasyon ile karakterize bağışıklık aracılı bir hastalıktır. MS'in genetik risk faktörleri aile temelli ve büyük ölçekli çalışmalarla kapsamlı şekilde araştırılmış olsa da, düşük frekanslı ve nadir varyantların (LFRV'ler) hastalık duyarlılığı üzerindeki etkisi hâlâ net değildir. Bu durum, MS genetik çalışmalarında Avrupa kökenli örneklerin fazlalığı nedeniyle oluşan örneklem yanlılığıyla ilişkilidir. Bu tez, Türkiye'de MS tanılı bireylerin bulunduğu en geniş aile temelli kohortlardan birini kullanarak LFRV'lerin MS riskine katkısını incelemektedir. Çalışmada, 59 çoklu MS vakası içeren aileden toplam 215 bireyin tamamına, hem hasta hem de sağlıklı akrabaları kapsayacak şekilde tüm ekzom dizileme uygulanmıştır. İlk olarak, MS ile ilişkisi iyi bilinen insan lökosit antijeni (HLA) bölgeleri incelenmiştir. Ardından, farklı kalıtım ve dominantlık modelleri dikkate alınarak MS ile birlikte segregasyon gösteren varyantlar tespit edilmiştir. Bu segregasyon analizine ek olarak, aile temelli yük analizleri gerçekleştirilmiştir. Tespit edilen varyant ve genlerin işlevsel önemini anlamak amacıyla yolak zenginleştirme analizleri yapılmıştır. Bulgularımız, lamininle ilişkili genler (LAMA5, LAMB1) ve yapısal öneme sahip genler (DST, PLEC, COL16A1) aracılığıyla ekstraselüler matriksin ve kan-beyin bariyerinin MS patogenezindeki rolünü vurgulamaktadır. Bu çalışma, Türkiye'deki MS genetiğine yönelik ilk ve en kapsamlı aile temelli araştırmalardan biri olup, LFRV'lerin hastalık riskine etkisine dair önemli bilgiler sunmaktadır. Tanımlanan yeni varyantlar ve yolaklar, MS'in moleküler temeline dair bilgimizi genişleterek gelecekteki tedavi hedeflerinin belirlenmesine katkı sağlayabilir.

Anahtar Sözcükler: Multiple skleroz, Genetik yük, Segregasyon, Ekzom, HLA

1 INTRODUCTION AND AIM

Multiple sclerosis (MS) is an autoimmune disorder of the central nervous system (CNS) characterized by chronic inflammation, demyelination, and progressive neurodegeneration in the majority of affected individuals. It is a prevalent cause of neurological disability in young adults and more than 2.8 million individuals are affected worldwide (1,2). MS significantly reduces the quality of life for those affected, particularly because it often manifests in early adult life, leading to profound impacts on their personal and social well-being. Although the exact disease etiology of MS still remains unclear, the current evidence from a plethora of studies and extensive research efforts over the years suggests that it is a complex and multifactorial disease where genetic and epigenetic factors in combination with environmental risk factors such as Epstein–Barr virus (EBV) infection, certain vitamin deficiencies and smoking play an important role in MS susceptibility (3,4).

Over the years, numerous genes and genetic loci contributing to MS susceptibility have been identified through various studies, most notably the human leukocyte antigen (HLA) region. This region has already been associated with most autoimmune disorders and HLA-DRB1*15:01 allele in this region is one of the strongest MS risk variants where carriers of this allele are three times more likely to develop MS compared to non-carriers according to the recent estimations (4). Recent studies focusing on this region have identified additional protective and MS susceptibility variants. Current estimates suggest that the HLA region accounts for 20-30% of the genetic susceptibility in MS (5).

With the advancement in sequencing technologies, genome-wide association studies (GWAS) utilizing large case-control cohorts have allowed the identification of genetic loci and variants with small effect sizes that contribute to MS susceptibility. Multiple GWAS over the years as well as meta-analysis approaches to combine the signals across different studies in order to improve statistical power have revealed many more common variants(minor allele frequency(MAF) > 5%) with minor effects sizes and confirmed the long-held hypothesis of MS susceptibility is driven by the

combination of many common variants with small effect sizes across the genome (1). Over the years, GWAS sample sizes continued to improve and consortium efforts such as International Multiple Sclerosis Genetics Consortium (IMSGC) enabled it to reach very large sample sizes via meta-analysis of association studies conducted over the years. A recent meta-analysis conducted by the IMSGC has revealed over 200 genetic associations, providing significant insights into the genetic architecture of MS (5). This study underscored the critical role of the immune system in the pathogenesis of the disease.

However, the genome-wide effects associated with MS over the years were found to explain 18.3% of the overall disease heritability (IMSGC citation). With the discovery of new variants alongside the HLA variants in the major histocompatibility (MHC) locus, the heritability explained by the MHC locus variants has reached to 20.2%. Combining the common variant effects discovered through genotyping arrays with the MS associated variants in MHC locus, approximately 39% of disease heritability now can be attributed to the identified MS susceptibility alleles (4,5). Although the identified common MS variants account for a substantial portion of the disease's heritability, a significant portion of the genetic predisposition to MS remains unexplained, a phenomenon referred to as the "missing heritability" problem (6).

There are several explanations hypothesized over the years to pinpoint the source of remaining disease heritability such as non-additive epistatic effects where some variants contribute to disease risk more when combined, gene-environment interactions, epigenetic effects and the contribution of low-frequency and rare variants (LFRV). Different studies over the years have found no significant evidence of epistatic effects between common MS variants contributing to disease heritability thus it was concluded that epistatic interactions between identified common MS variants are not the major source of the missing heritability (7).

Several studies focusing on this missing heritability problem for MS concluded that the low-frequency and rare variants can be a major source of MS risk. However, due to weak linkage of rare variants with other common alleles, capturing disease-

associated rare variants through commonly used genotyping arrays and linkage disequilibrium (LD) based imputation methods is highly challenging. Furthermore, the low MAFs of these rare variants make it even more challenging to identify rare-variant associations through regression-based testing between case and control samples resulting in low statistical power. It requires very large sample sizes to overcome this reduced statistical power and recently a study led by IMSGC utilizing 68,379 cases and controls identified four novel genes and a set of low frequency codings variants that explains up to 5% of MS risk (6,7).

To address the issue of low statistical power, family-based studies have been adopted as a complementary approach alongside large-scale ascertained association studies. These studies aim to identify novel MS risk variants and contribute to a better understanding of the missing heritability in the disease. Family-based studies leverage the fact that rare disease-associated variants, which are difficult to detect in population-based case-control studies, can theoretically be found in multiple instances within families that have several affected individuals. To address the low power for detecting rare risk variants, which is further exacerbated by the smaller sample sizes typically seen in family-based studies compared to population-based studies, gene-based rare-variant association tests have been developed and widely utilized in many rare variant studies. Gene-based association tests, or variant aggregation tests, combine variants into genetic units which are typically genes, and then use these aggregated units to perform the appropriate statistical tests (8).

Another issue is that many GWAS and large-scale population-based association studies have predominantly focused on samples from individuals of European ancestry. Studying samples from diverse populations is particularly important for uncovering novel low-frequency and rare risk variants, as these variants can be more population-specific due to their relatively recent origin compared to common variants. Thus, studying multi-ancestry and admixed cohorts for rare-variant discovery are essential towards having a more complete and broader understanding of genetic architecture of MS (9).

In this study, one of the largest family-based cohorts to date, comprising 60 multiplex Turkish families and a total of 215 samples, was subjected to whole-exome sequencing using next-generation sequencing (NGS) technologies, aiming to identify novel low-frequency and rare MS risk variants, as well as disease-associated genes, to uncover new aspects of disease biology. To achieve this, a multi-method approach was employed, incorporating both segregation-based variant and gene selection across the families and rare-variant discovery analysis through gene aggregation and association testing. Combining results from different gene prioritization strategies is crucial to overcome sample size and power limitations and to increase our confidence in the discovered associations. Additionally, a literature driven approach was taken to analyze the risk variants and genes previously associated with MS from large GWAS and other population-based association studies as well as other family-based studies to compare the known MS loci with our results and further scrutinize the signals that were re-discovered by our analysis. Furthermore, the impact of known HLA variants was investigated within our families to determine whether these variants could account for increased MS susceptibility in any of the families analyzed in this study. As a complementary analysis to understand the influence of consanguinity on MS susceptibility, which has yielded conflicting results in previous studies, we examined consanguinity rates in our families and their association with MS prevalence. This analysis aims to contribute to the literature on this topic by leveraging the sample size and unique genetic background of our cohort.

The primary goal of this study is to identify novel risk variants and genes associated with MS, which may help to explain the increased disease prevalence observed in the families of this cohort. By doing so, the study aims to enhance and contribute to the current understanding of genetic factors involved in MS heritability and to uncover new biological mechanisms underlying the disease etiology and pathogenesis. These findings could later be examined and validated through in vitro and in vivo experiments, with the ultimate goal of translating the results into clinical therapies to benefit the millions of people affected by MS.

2 BACKGROUND

2.1 From Immune Response to Pathogenesis of Multiple Sclerosis

Multiple sclerosis (MS) is an autoimmune mediated demyelinating disorder of the central nervous system (CNS) causing severe physical and cognitive impairments in those affected. Demyelination primarily arises from inflammation driven by the infiltration of T-lymphocytes and macrophages, followed by the death of oligodendrocytes. As a result, CNS plaques formed composed of inflammatory cells, demyelinated axons, and abnormal astrocyte activity in both white and gray matter. These lesions cause neuronal dysfunction, leading to visual disturbances, ataxia, sensorimotor impairments, weakness and even emotional issues (10).

Inflammation in the white and gray matter of CNS tissues is caused by the infiltration of immune cells and their associated cytokines. To understand the immune reaction that initiates MS progression, it is crucial to comprehend how immune system response function. There are two main types of immune response: innate and adaptive.

The innate immune response is initiated by microbial products activating specific receptors, primarily toll-like receptors (TLRs), in an antigen-nonspecific manner. Microbial products bind to specific subsets of TLRs causing the cytokine production that regulates the adaptive immune response. The innate immune response plays a role in both the initiation and progression of MS by modulating T and B cells. For instance, when dendritic cells (DCs) are activated via TLRs, they become semi-mature and prompt regulatory T cells to secrete inhibitory cytokines like IL-10 or TGF-β. As these dendritic cells further mature, they guide CD4+ T cells to differentiate into Th1, Th2, or Th17 phenotypes. Differentiation into the Th1 phenotype specifically leads to the promotion of inflammation (2,10).

The adaptive immune response is modulated by the specific antigen presentation to T lymphocytes via antigen presenting cells(APCs). The interaction between T cells including CD4+ and CD8+ phenotypes and APCs is the central mechanism that

initiates and regulates the adaptive immune response. Upon the exposure of specific interleukins, effector cells such as Th1, Th2 and Th17 are polarized as a response. When these effector T cells are polarized, they secrete certain cytokines that have proinflammatory effects such as IL-4 and IL-13. These effector CD4+ T cell subsets were found to be promoting inflammation in MS. After primed in the periphery region, Th1 and Th17 were found to be migrating to CNS that cause demyelination and axonal loss observed in MS(2,11).

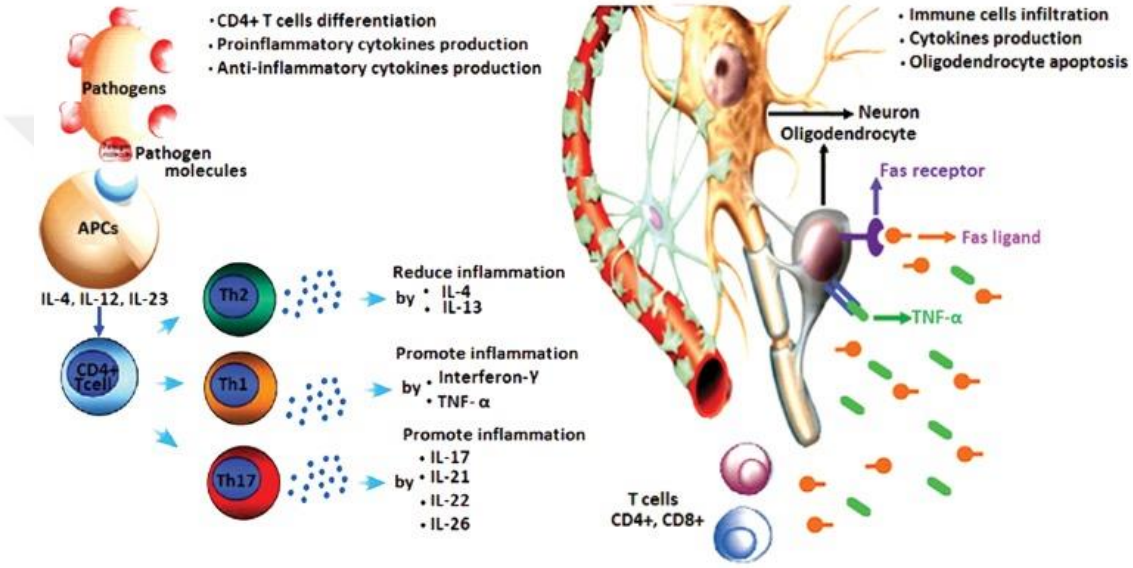


Figure 1. Immune cells and cytokines produced by those immune cells that are associated with the multiple sclerosis pathogenesis (2).

In addition to CD4+ T cells, such as Th1 and Th17, B lymphocytes and the cytokines they secrete also play a role in the pathogenesis of MS. Transforming growth factor beta (TGF-β) and TNF-α are the cytokines produced by B lymphocytes that can initiate inflammation. Conversely, IL-10 cytokines produced by B lymphocytes have an anti-inflammatory effect thus B lymphocytes have both negative and positive influence in MS development.

Additionally, cytotoxic T cells or CD8+T can also be detected in MS lesions and play an important role in the MS pathogenesis by increasing vascular permeability and initiation of oligodendrocyte death that impairs the process of myelin repair. Another process that causes the disruption of myelin synthesis and contribute to the disease

pathogenesis is binding of fas ligand (FasL) that is produced by lymphocyte cells to Fas, a TNF superfamily receptor, on oligodendrocyte cells causing these cells to initiate apoptosis(2) .

2.2 Disease Etiology

The exact cause of MS and underlying disease etiology still is not fully understood however it's been known that both genetic, epigenetic, environmental factors as well as interactions of those contribute to the etiology and progression of MS (12).

2.2.1 Environmental factors associated with MS

There are multiple infectious agent has been associated with increased MS risk such as human herpesvirus type 6 and mycoplasma pneumonia but exposure to Epstein-Barr virus, even though the causal relationship with EBV is still debated, has been persistently suggested by many studies to play a significant role in MS across multiple ancestry groups (2). A meta-analysis conducted by Handel et al. demonstrated that individuals with a history of EBV infection have a twofold increased risk of developing MS (13). Later studies have demonstrated the significance of the temporal relationship between EBV infection and the MS risk. Specifically, infection during adulthood has been associated with an increased risk of MS, a pattern not observed in cases of EBV infection during childhood. Another nuanced point, and a potential counterargument to the relationship between EBV infection and MS, is the high prevalence of EBV seropositivity in the general population. This observation suggests that EBV infection alone may not directly increase the risk of MS. Instead, this effect might be mediated by other risk factors, such as gene-environment (GxE) interactions, where certain genetic predispositions are necessary for the increased MS risk to manifest in individuals exposed to EBV (3,12).

Another extensively studied environmental factor for MS is vitamin D levels and sun exposure. These two factors are highly intertwined and correlated as ultraviolet radiation from the sun can also trigger vitamin D synthesis endogenously. Both

exposure to the sun and high vitamin D levels are associated with decreased MS risk especially before the age of 20. This association between vitamin D levels and MS risk is further supported by the genetic evidence where the variants that can alter the genes and proteins related to the vitamin D metabolism such as *CYP27B1* were associated with MS. There are also few studies that explored gene-environment interactions in the context of vitamin D and *HLA DRB1*15:01* which is the strongest known genetic MS risk factor but in vitro results have not been reproduced in humans (3).

There are other environmental factors involved in increased MS risk that have been studied such as smoking due to the toxic effects of nitric oxide (NO) and carbon monoxide (CO) on neurons and oligodendrocytes and diet where there is a positive correlation between MS risk and obesity.

2.2.2 Genetic architecture of MS

2.2.2.1 Estimation of disease heritability

Heritability is one of the key metrics in genetics that shows the proportion of the variation in a given phenotype that is caused by genetic factors. This parameter is important to understand the relative contribution of genetic and environmental factors to the variation in a particular trait in a population. There are two types of heritability parameter: narrow-sense heritability(h^2) which only considers additive genetic effects and broad-sense heritability(H^2) where total genetic factors are included into the equation including dominance and other non-additive genetic effects. Narrow-sense heritability is the mainly used heritability parameter since correlation between relatives is mainly dependent on h^2. This term is calculated by dividing the variation caused by additive genetics effects in the population by total observed variation controlled by the known covariates such as sex and age. There are numerous traditional designs for estimating heritability such as twin studies where genetically identical monozygotic(MZ) and dizygotic(DZ) twins' phenotypic differences are examined and if genetic variation affects phenotypes, we expect to observe MZ twins to be phenotypically more similar compare to DZ twins (14,15).

Multiple sclerosis, similar to other complex disorders, exhibits significant heritability via genes and variants. Studying familial cases and siblings are a crucial step to understand the relative contribution of these genetics and environmental factors to the disease etiology. A national registry based study from Sweden that includes 15 million individuals with 28,396 MS patients reported estimated sibling recurrence risk of 7.1 and heritability estimation 0.64 using 348 proband twins with MS (16).

This significant evidence of heritable component in MS attracted many population studies to narrow down on genetic components that contribute to disease heritability.

2.2.2.2 Early association studies and HLA region

Until the first large-scale MS GWAS was published, the only genetic signals that had been detected by multiple studies were from the HLA region. There had been a handful of genes associated with increased MS risk with the HLA-DR15 allele being the lead signal in terms of effect size (17).

Major histocompatibility complex (MHC) is a dense region that includes HLA genes which play a pivotal role in histocompatibility and adaptive immune response. Due to its role in displaying antigenic peptides to T lymphocytes, this region has been extensively mapped and associated with a plethora of autoimmune and inflammatory disorders as well as with some cancer types. This region is the most polymorphic loci in the genome and over the years, more than 10.000 alleles have been identified. There are two main classes for HLA genes, HLA class I and class II. HLA class I mainly play a role in displaying peptides derived from viruses and tumors to CD8 + T cells meanwhile the HLA class II molecules such as HLA-DR and -DQ which creates a groove shape when assembled where peptides recognized by CD4+ T lymphocytes bind to (18).

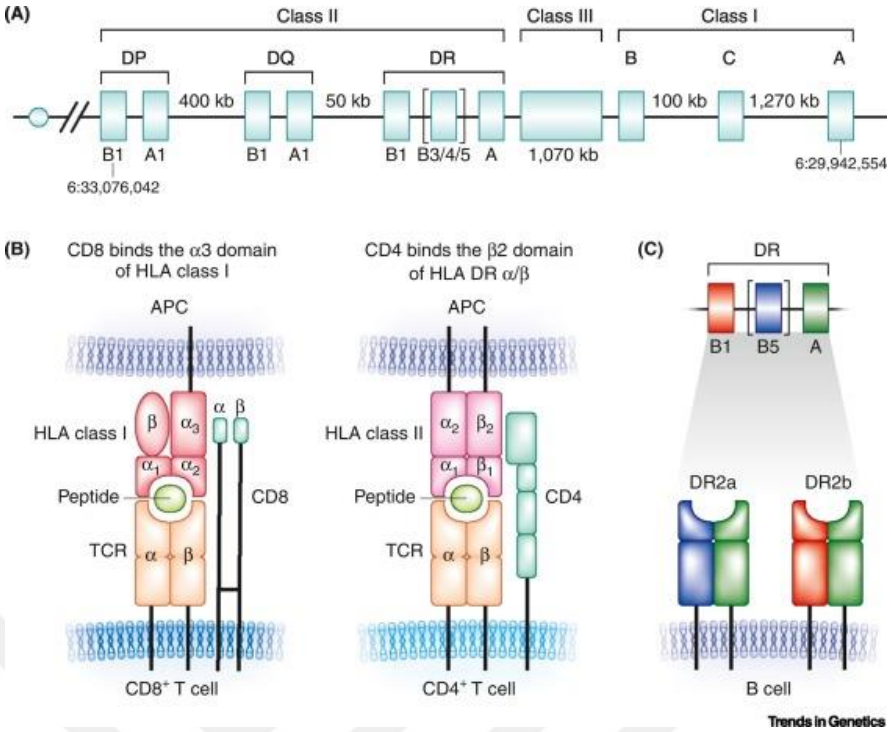


Figure 2. A schematic and structural view of the HLA region on chromosome 6 and different HLA classes (18).

The effect of HLA in MS risk was first recognized even before the molecular definition of the class II HLA molecules via serological association methods. The extensive research of the polymorphisms in the HLA region continued over the years and with the advancements in genotyping methods, the genotyping resolution was improved and other genes and alleles were associated with MS risk such as DQA1*01:02 and DQB1*06:02. The main association for HLA-DR15 was further refined to DRB1*15:01 allele which is the strongest MS risk factor to this day. This association was later confirmed by multiple GWASs and one of the latest and most extensive GWAS in terms of sample size conducted by IMSGC with 47 000 patients and 68 000 controls reported an association between MS and HLA-DRB1*15:01 with p-value 1e-1900 with 3.9 average odds ratio. The HLA-DRB1 gene in the class II region of the MHC alone was found explaining up to 10% of genetic variants underlying MS risk (5,17).

2.2.2.3 The era of GWAS and other large scale association studies for MS

Mapping genetic variants and estimating disease heritability have proven to be highly effective strategies for elucidating the underlying molecular and biological mechanisms of disease. Studying families and tracking disease status across multiple generations and its correlation with inherited genomic segments have identified multiple genetic loci particularly for the Mendelian diseases. However, this approach has been less successful in identifying risk loci for MS, primarily due to the complex and adult-onset nature of the disease.

Advancements in sequencing technologies with a significant reduction in costs, have allowed the mapping of genetic variants across a large number of samples. With these advancements, GWAS, which analyzes hundreds of thousands of genetic variants and their frequencies across large cohorts of independent samples to identify loci associated with specific diseases or phenotypes, have become the standard for large-scale association studies. This progress has led to a surge in GWAS across hundreds of diseases, resulting in the identification of numerous significant genetic loci. These findings have, in turn, advanced our understanding of disease biology and facilitated drug discovery efforts. Large-scale, disease-specific consortia have emerged to consolidate multiple independent studies, enabling joint analyses that leverage the increased statistical power provided by the aggregated sample sizes.

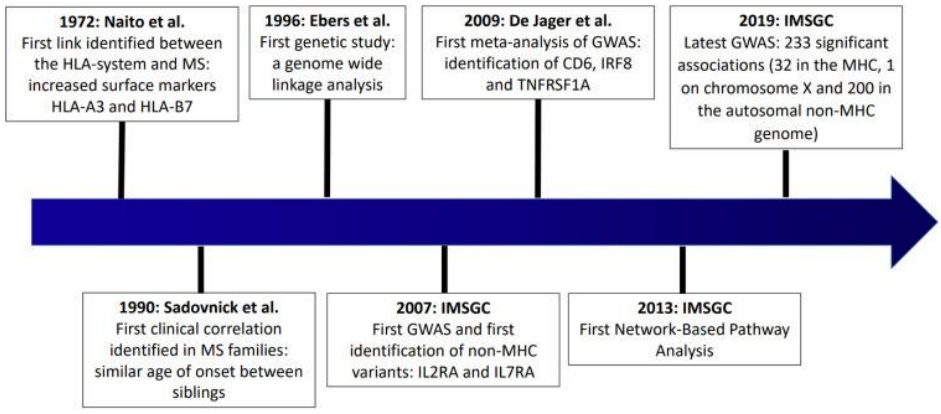


Figure 3. A timeline of key milestones in genetic studies of multiple sclerosis.

The most comprehensive of these efforts for MS is the formation of the IMSGC, which has been conducting large-scale association studies and meta-analyses across diverse cohorts to identify MS risk loci and develop a comprehensive genetic map of the disease. One of the early GWAS conducted by IMSGC in 2007, using over 900 family trios and a substantial replication cohort for its time, led to the identification of the first non-MHC variants in the IL2RA and IL7RA genes. In 2019, the IMSGC published the most comprehensive GWAS for MS to date, analyzing allele frequencies in 68,248 control samples and 47,351 MS cases. This extensive study identified 233 MS risk variants, including 200 autosomal signals located outside the MHC region, 32 independent signals within the broader MHC region, and one signal on the X chromosome. Altogether, approximately 39% of MS disease risk can now be attributed to validated genome-wide risk alleles. When accounting for disease prevalence and environmental factors, these risk variants explain about 20% of MS risk in the general population.

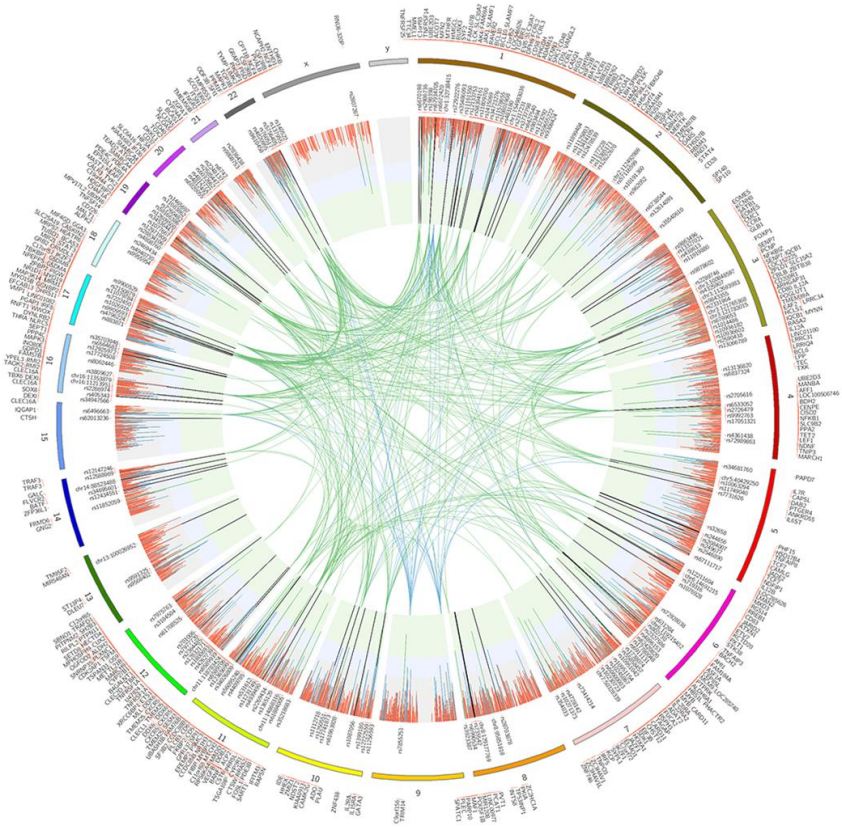


Figure 4. The genetic map of MS from the milestone IMSGC GWAS paper.

While the large sample size and statistical power of this GWAS have enabled the discovery of many new genome-wide signals, translating these findings into underlying molecular pathways and advancing our understanding of MS biology remain challenging. This difficulty arises largely from the correlation between variants caused by linkage disequilibrium and fine-mapping strategies have been applied to narrow down the signals to fewer variants and detecting the "causal" genes in order to gain insight regarding disease biology. Although fine-mapping efforts have resolved some regions, most loci identified by GWAS remain unresolved. Addressing these regions will require larger sample sizes to enable more effective fine-mapping.

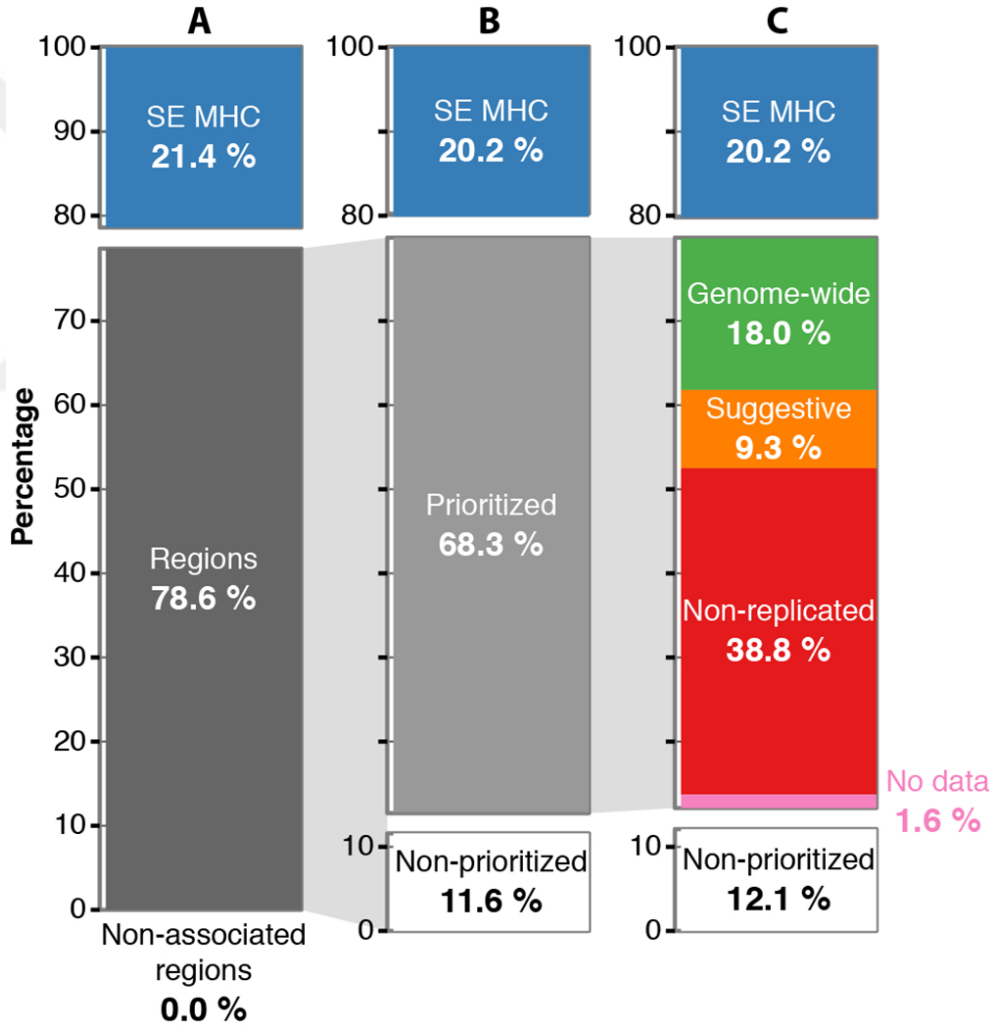


Figure 5. Overall narrow-sense heritability partitioned into different genetic components.

The MHC region appears to mediate a substantial portion of MS risk (Figure X). The 32 independent signals within this region, which include both previously known and novel associations, may serve as a gateway to uncovering the underlying biological mechanisms that drive autoimmune reaction and the interaction between EBV infection and MS development.

Outside of the MHC region, the discovered signals are hypothesized to act through gene enhancers and promoters that are involved in the immune system and its components such as macrophages, microglia, T-cells, B-cells and natural killer cells. Many of these MS risk variants are also associated with other autoimmune and inflammatory disorders, such as rheumatoid arthritis and type 1 diabetes, suggesting that their effects are not CNS-specific but instead disrupt immune system function more broadly. Given their effect sizes, these variants likely influence immune cell function near the CNS through gene regulation. Over time, these effects can accumulate and contribute to the development of the MS phenotype.

Similar to other complex traits, even though common variants (MAF > 5%) and loci discovered through GWAS explain most of the disease heritability, they cannot explain it entirely. Most of the variants identified by GWAS account for a relatively small portion of heritability with each variant having small effect sizes together adding up to impact a trait or disease risk in a meaningful way. Even though thousands of new loci are common diseases and, in this context, more than 200 loci identified for MS, still a huge chunk of disease heritability remains unexplained. There are many explanations have been suggested over the years for sources of missing heritability including unidentified common low effect size signals that would close the heritability gap, low frequency and rare variants with moderate to high effect sizes that are hard to detect using common genotype arrays due to their rarity, structural variants and gene-gene and gene-environment interactions (19). To understand the role of low frequency and rare variants in MS risk and see the missing heritability can be explained by such variants that might even be exclusive to individual families which can potentially lead to novel biological insights, a very extensive exome-wide rare variant study was conducted by IMSGC in 2018 (7).

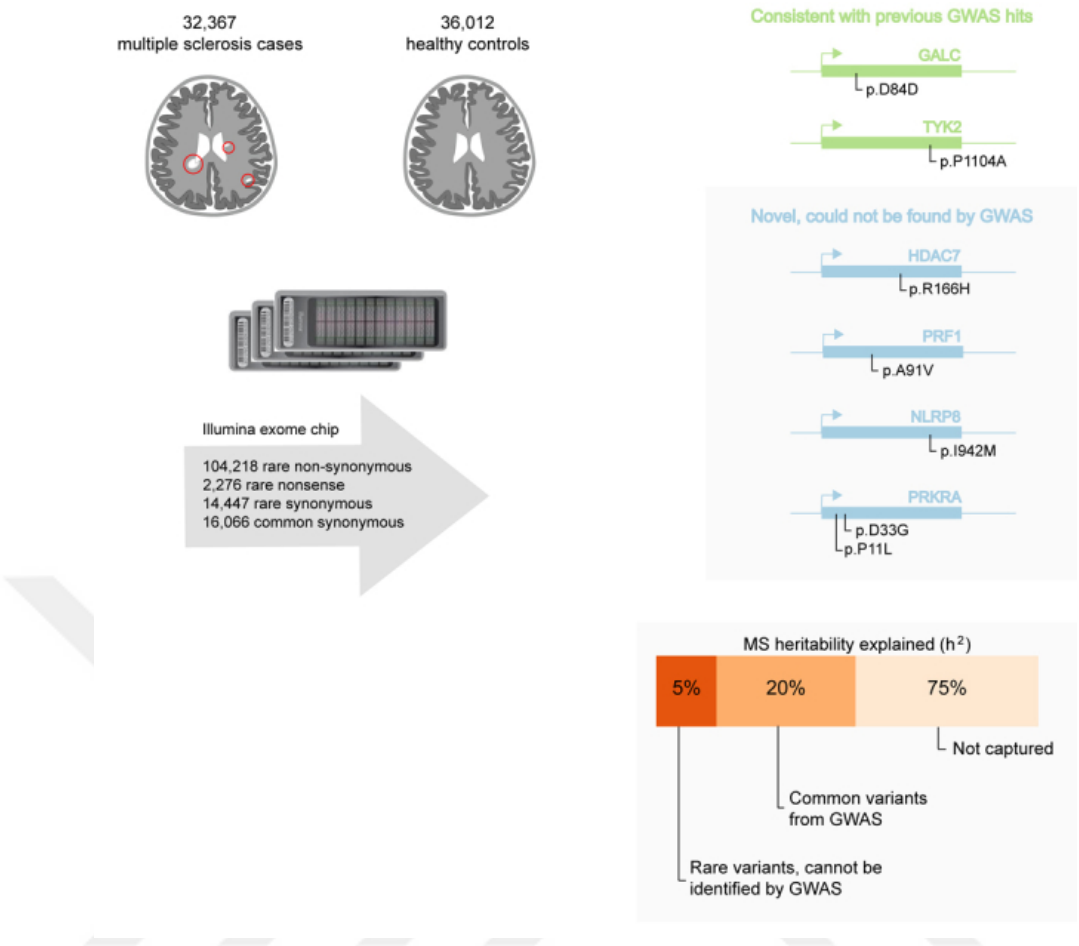


Figure 6. Overview of the low frequency and rare variant study conducted by IMSGC 2018.

To investigate this, a total of 120,991 variants across all autosomes were tested using a custom exome-chip array in a cohort comprising 68,379 cases and controls. The analysis revealed that low-frequency and rare variants contribute to disease heritability, which could not be detected through common variant association studies. The findings suggest that increasing sample sizes has the potential to identify additional low-frequency and rare variants associated with MS risk. Collectively, the MS-associated signals identified in this study account for up to 5% of the previously unexplained disease heritability. Notably, some of these signals are consistent with prior GWAS findings, and the study identified four novel genes associated with MS.

Two out of those four genes -*PRKRA* and *NLRP8*- were found to be associated with the innate immune system directly and *HDAC7* is known to be playing a critical role in T cell development in thymus. The study concluded that due to their direct and

relatively high effect sizes, studying the effect of these variants and genes can potentially reveal novel pathways and molecular processes that contribute to MS pathogenesis.

Although many common and low-frequency variants have been identified through large-scale population studies, these variants are insufficient to fully account for the increased risk of MS, particularly in familial cases. Another caveat is that most large-scale association studies have primarily been conducted using samples from individuals with European ancestry background. Consequently, methods designed to estimate the cumulative burden of common variants, such as polygenic risk scores (PRS), are known to underperform when applied to relatively understudied populations (20). Everest *et al.* investigated the contribution of common and rare variants in increased disease risk using familial MS cases from a Turkish cohort. For this, they have calculated the PRS score using a set of high confidence independent signals from IMSGC meta-analysis that consists of 175 SNPs.

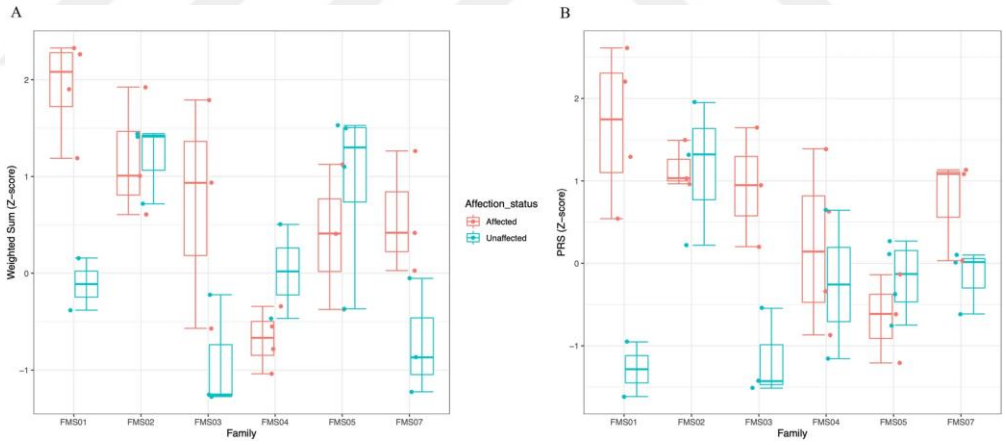


Figure 7. PRS and weighted sum score comparison across the familial MS cases

Even though they observed a higher PRS and weighted sum score in the affected individuals from families compared to control cases and families, no significant PRS difference was observed between familial and sporadic MS cases. Furthermore, a comparison of PRS and sum scores among individuals within families revealed that affected family members exhibited higher values compared to their unaffected relatives in three out of six families. This suggests that, while polygenic inheritance

plays a critical role in MS risk, the contribution of possibly family-specific, fully penetrant or near-fully penetrant low-frequency and rare variants, in combination with other risk factors, may also play a significant role in the inheritance and development of MS (21).

Although challenging due to the large sample sizes required to achieve adequate statistical power, investigating highly penetrant low-frequency and rare variants in familial MS cases, particularly within understudied populations, holds significant value. Such efforts have the potential to uncover novel variants and genes associated with MS, providing insights into previously unrecognized disease pathways and biological mechanisms.

2.3 General Background on the Statistical and Computational Methods Used

2.3.1 Transmission disequilibrium test

Transmission disequilibrium test (TDT), first proposed by Spielman et al. in 1993 is a family-based association test that compares the allele transmission rate from heterozygous parents to their affected offsprings (22). It is a powerful and robust analysis even in the presence of population stratification but its statistical power depends on the number of informative parents with heterozygous genotypes. TDT test was originally developed to analyze family trios with two parents and their affected offspring. TDT analysis formula is quite simple for family trios and it takes advantage of calculating the ratio transmitted and non-transmitted alleles using heterozygous parents to find out if allele transmission rates divert significantly from the expected range i.e causing a “disequilibrium” (23).

We can then tabulate the allele transmissions using a 2×2 table, as shown below:

Table 1. 2x2 allele transmission table for TDT test.

	Non-transmitted allele		
Transmitted allele	**M_1**	**M_2**	**Total**
M_1	*a*	*b*	*a + b*
M_2	*c*	*d*	*c + d*
Total	*a + c*	*b + d*	*2n*

Where, *a* and *d* are the cases where both transmitted and non-transmitted alleles are the same whereas *b* and *c* represent heterozygous parents. Here only heterozygous parents (*b + c*) contribute to the test and the null hypothesis states that likelihood of the transmission of both alleles are equal and 0.5. Then we can test whether this ratio follows our prior using binomial distribution with one degree of freedom:

$$\chi 2 = \frac{(b-c)2}{b+c}$$

If there are no transmission disequilibrium, $b \approx c$ and our test statistics $\chi 2$ will be close to 0 otherwise we will observe a high $\chi 2$ that would suggest a transmission bias. This test later on expanded to include two affected offsprings, quantitative traits as well as extended pedigrees via generalized relationship matrices (GRMs) as an alternative analysis method to family-based burden and SKAT tests we are going to discuss in the next section.

2.3.2 Family-wise gene burden test for rare variants

The core statistical method used by GWASs is running regression analysis for each common marker on case-control dataset and generating effect size and p-value test statistics after multiple testing correction to detect top associated variants for a given phenotype. Even though such variant-by-variant analysis has been quite successful to discover common variants contributing to a phenotype, it has been shown to underperform for rare variants due to their low MAF and sampling across study samples. In order to increase statistical power, multiple groupwise analyses have been published, where variants are grouped into predefined genetic units, such as loci or simply into genes, to aggregate weak signals across many variants in a region and strengthen the association signal. Burden and variance components tests, such as SKAT, are the two most commonly used groupwise association tests. The burden test assumes a uniform effect size and direction for all variants in a given region, whereas variance components tests make no such assumptions regarding effect size and direction. However, variance components tests are less powerful when effect sizes and directions are consistent across all variants (24).

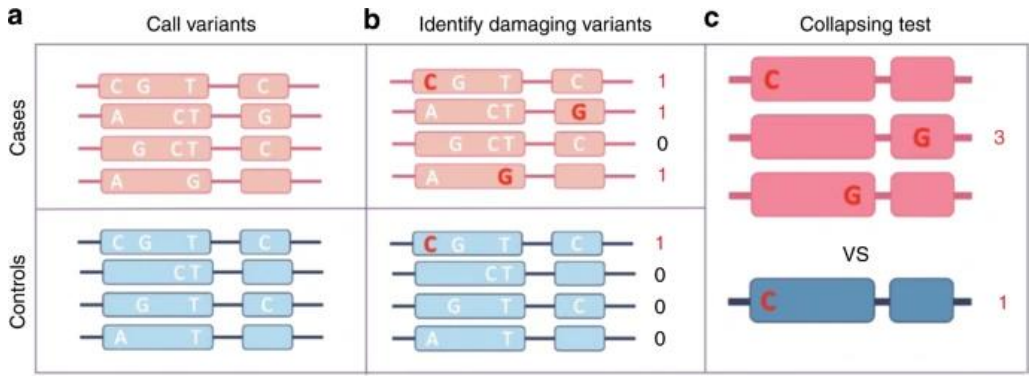


Figure 8. Visualization of region-based collapsing analysis for case-control data.

Even though such collapsing strategies have been quite successful at discovering novel disease associated genes and regions, these methods have mainly been developed for case-control studies in which the samples are assumed to be independent. This assumption of independence would not hold for analyzing samples from families where the samples are related. To utilize grouping and collapsing

strategies for a better statistical power for rare variants and at the same time, taking advantage of enriched presence of rare variants from families with multiple affected subjects, several family-based region-based association tests have been developed such as FamBAC, FBAT, pedgene. Even though their approaches differ in some ways, these methods take advantage of linear mixed models (LMMs) and relationship matrices either calculated from known pedigrees or estimated using large-scale genomic data (25–27).

2.3.3 Pathway enrichment analysis

Pathway enrichment analysis has been a widely used method to gain biological insights from a list of genes generated from a wide range of omics experiments. Although there are different approaches to achieve this, the main idea is to identify biological pathways that are relevant and enriched for our set of genes to a greater extent than would occur by chance. However, this requires curating a collection of gene groups with biological and/or functional annotations that would reflect our prior knowledge for these genes and pathways that we can use to compare with our list of genes and quantify this enrichment. For this purpose, The Gene Ontology (GO) was created in 2001 to associate genes with functional biological terms in a systematic way which enabled gaining biological insights into a given gene list by pathway enrichment (28).

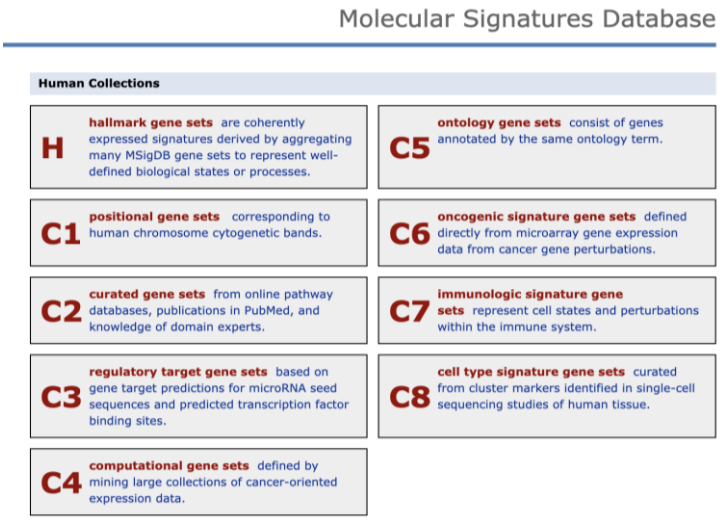


Figure 9. Gene-set categories presented in MSigDB database for human data.

There is a plethora of pathway databases available on top of GO such as Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, WikiPathways, Molecular Signatures Database (MSigDB) as well as a wide selection of enrichment tools like gProfiler, Enrichr, DAVID, Gene Set Enrichment Analysis (GSEA) where the first was two used in this manuscript. Different tools use various statistical approaches to calculate the relevance of a given list of genes with existing gene sets, such as Fisher's exact test, the hypergeometric test, and the Kolmogorov-Smirnov test (29).

Over-representation analysis is one of several pathway enrichment approaches that can determine the degree of overlap between pre-defined gene sets (GO, KEGG) and a gene list of our interest. For this, we can employ the Hypergeometric distribution to calculate enrichment p-values, which reflect the likelihood that a given observation, such as an enriched pathway, is due to random chance.

$$P(X \geq x) = 1 - P(X \leq x - 1) = 1 - \sum_{i=0}^{x-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

Hypergeometric formula to calculate probability of observing and overlap of *x* or more with a given gene set. *N*: the number of background genes, *n*: the number of genes we are interested in i.e from our burden analysis, *M*: the number of genes that are annotated to a predefined gene-set *S*, *x*: the number of genes of our interest found in gene set *S*.

As an example, assume we have a list of 60 genes from our association test and we found 20 of those in a gene set from GO with a total size of 100. If we use 20.000 for the number of background genes and calculate the probability of 20 or more genes annotated to genet set *S* using the formula above, we get a p-value of 4.51e-32 which suggests that the probability of randomly observing such a result is highly unlikely, indicating that the pathway or biological process associated with gene set *S* may be worth further study.

3 MATERIALS AND METHODS

3.1 Variant Calling and Annotation

Fastqc quality control program was used in order to check the read quality of fastq files coming from sequencing. Here we checked the quality of reading based on metrics such as per base sequence quality, per base sequence and qc content and duplication levels to make sure qualities are sufficient across our data for moving to next steps(30). In order to remove adapters according to the sequencing technology we used and low quality sequences from the reads, Trim-Galore was used (31). Our sequence reads were mapped to human reference genome build 38(GRCh38 or hg38) utilizing Burrows-Wheeler Aligner (BWA) tool version 0.7.17 (32). BWA was chosen based on its speed and accuracy for aligning both short read and long read sequences to large reference genomes. After aligning our reads to the hg38 reference genome, we moved to the pre-processing and variant calling step where Genome Analysis Toolkit's (GATK) best practices workflow was followed (33). Here we started the pre-processing step with identifying and correcting duplicate reads using GATK's MarkDuplicates tool. Duplicate reads are the reads coming from a single DNA fragment and can emerge while constructing libraries. MarkDuplicates run a sequence comparison algorithm to detect those duplication artifacts and mark them with a bitwise flag. Next, FixMateInformation tool was used to synchronize each read with its pair. GATK's BaseRecalibrator tools was used to calculate based quality using a wide range of information where poor base quality indicates errors and problematic reads. Finally, HaplotypeCaller was used to call SNPs and indels very accurately using a loca re-assembly strategy based on likelihoods where whenever the program detects a variation in a region, it's reassembles the reads in this region from scratch in order to do a very precise variant calls especially for the challenging genomic regions to make a call (34). Variant calling step was done per family basis to generate variant calling format(vcf) files and bcftools was used to merge vcf files from all of our families into one large vcf file (35). Bcftools was also utilized to left-align and normalize indels to have a consistent indel starting position across the dataset for downstream analysis. Lastly, multiallelic sites were split into biallelic sites using

bcftools' "norm" function to finalize our unannotated vcf file. The merged vcf file that contains all of our sample then was annotated using vcfanno annotation tool. Here, Ensembl's Variant Effect Predictor(VEP) version 109, ANNOVAR and dbNSFP version 4.2 were used as annotation databases in order to obtain a wide range of variant information such as gene names, variant consequence(missense, stop gained, frameshift etc.), existing variants and their variant identifiers(e.g rsid) as well as their minor allele frequencies(MAF) in existing databases (36–38). On top of a plethora of generic variant information, we gathered information from *in silico* prediction tools such as CADD, SIFT, PolyPhen-2, MetaRNN and REVEL in order to see how a specific variant alters and affects protein function (39–42).

3.2 Quality Control, Pedigree and Gender Check

Kinship-based INference for Gwas(KING) software was employed in order to check pedigree errors and get relationship inference in order to detect and correct any errors and unknown relatedness in our families (43). KING software's --kinship flag with default settings was used to analyze our vcf file and estimate kinship coefficients between our samples and inconsistencies in kinship estimations with pedigree kinship estimations were further investigated. In addition to that, Peddy tool was used to detect any gender discrepancies between the actual gender of a sample and its label on the pedigree by estimating heterozygosity ratio in the X chromosome (44). Peddy tool was run with the default settings and 0.60 heterozygosity ratio was used to determine the sample's estimated gender. As an additional quality control, variants with mendelian inconsistencies were detected using 34 complete trios from our families and were excluded from the analysis utilizing vcftools' "--mendel" flag. On top of that, variants diverging too far from Hardy-Weinberg Equilibrium (HWE) at $p < 1e\text{-}4$ were excluded. Hardy-Weinberg p-values were calculated using vcftools' "--hardy" flag for each site using only founder samples from families. If there is no "founder" sample available, one sample from each of those families was used according to the examples from literature (45).

3.3 HLA Typing

For typing(determining pre-defined allele groups) HLA alleles for our samples, HLA typing from High-quality Dictionary(HLA-HD) tool was used in order to determine HLA alleles with up to 6-digits using NGS data in fastq format (46). HLA-HD works by mapping reads from fastq files to the HLA alleles pre-defined in a dictionary and calculating a score to determine HLA alleles for a given sample with high precision and speed. Utilizing this tool, 29 different HLA alleles were typed across our samples using mostly the default parameters with only exception of changing read length parameter from default 100 to 90 due to encountering an error for some of our samples when we run them with read length parameter 100. HLA typing was performed per sample basis, then a custom python script was used to compile the individual results by the HLA allele type.

3.4 Variant Filtering

Variants were filtered according to their minor allele frequency as well as their predicted functionally using in silico predictors in order to filter out common and low impact variants and get a set of low frequency and rare high impact variants for the downstream segregation and rare variant association analysis. For variant filtering, depending on the downstream method, different masks with different strictness were compiled and applied to the variant data. The main mask that was used in this analysis were based on the variant MAF and CADD score. Here, variants with max MAF across 1000 genomes project, ExAC database and GnomAD database equal or less than 5% were kept to ensure selected variants are low frequency and rare across different ancestries (47). If there is no MAF information in any of the databases mentioned above for a given variant i.e variant was not captured in any of those databases, we kept the variants as they are novel variants from our cohort. To assess variant deleteriousness and keep the variants that are likely to disrupt protein function, variants with CADD score equal or more than 15 were retained. In addition to CADD score filter, variants satisfying one of the following filters; PolyPhen2 prediction == D (probably damaging), REVEL score equal or more than 0.5, MetaRNN score equal or

more than 0.5 were also included since for some special cases such as splice donor/acceptor variants and splice region variants, there were no CADD score generated. Also, frameshift indels that are annotated with HIGH impact in VEP were also included. A morestrict filter was used to further narrow down our variant selection. For this, variants with max MAF 0.01 or less across the mentioned databases above were kept. Variants that are satisfying three of the following four in silico variant prediction filters which are CADD score equal or more than 20, PolyPhen2 prediction == D (probably damaging), REVEL score equal or more than 0.5 and MetaRNN score equal or more than 0.5 were kept to have a more confidence in selected variants having a damaging effect to protein function.

3.5 Segregation Analysis

Variants segregating with MS in our families were detected using a custom script written in Python. For this analysis, both dominant and recessive inheritance modes were examined, as well as cases of complete and incomplete penetrance, to ensure a comprehensive and robust assessment of segregation. For the scenario of complete penetrance in variants exhibiting autosomal dominant inheritance, we selected variants that were present in all MS cases within a family and absent in all healthy individuals within the same family. Additionally, we utilized all available healthy samples to ensure that the variants passing the "in-family" filters were not detected in any of the healthy samples.

Even though we had a small number of families suitable for exploring recessive inheritance, we applied a recessive inheritance filter that ensures all healthy individuals within a family are heterozygous for a given position, whereas all MS samples in a family are homozygous alternative for the same position. For the incomplete penetrance filter for the dominant inheritance assumption, variants found in at least 66% of all MS samples in a family were included. Here, 33% of healthy samples within a family and 33% of healthy samples across all of our samples were allowed to carry these variants to have a more lenient segregation filter. These numbers were chosen according to common practices in the literature and because, in the majority of our families, there are between one to five healthy samples and a similar number of MS

cases. Thus, in practice, we allowed one healthy sample and one MS sample in a family to be inconsistent with complete segregation patterns.

For all inheritance modes and penetrance cases, we conducted both variant-based and gene-based assessments. For variant-based analysis, we expected a given variant to segregate in more than one family to increase our confidence in the variants passing our filters. For gene-based analysis, instead of expecting a variant to segregate in more than one family, we collapsed variants into the genes they are located in and then checked how many families carry at least one segregating variant within a gene. Finally, we sorted both variants and genes according to the number of families with observed segregation for variant and gene prioritization.

3.6 Functional Enrichment Analysis and Protein Interaction Networks

Candidate genes from segregation filters were further analyzed using gene-set enrichment analysis and protein-protein interaction networks to have a molecular and biological insight. For this, a wide range of tools were employed and their results were compared due to each tool having different strengths and using different gene ontology databases in order to capture a more complete picture. Enrichr was the main enrichment tool used for this analysis (48). The reasoning behind using Enrichr as our main gene set enrichment tool was that Enrichr uses a plethora of gene set libraries such as REACTOME, Kyoto Encyclopedia of Genes and Genomes (KEGG), Biocarta and Molecular Signatures Database (MSigDB) Hallmark. This variety is crucial as each database offers distinct perspectives on biological pathways with varying degrees of curation. Therefore, adopting a parallel and comparative approach in pathway enrichment analysis is a common practice. Additionally, Enrichr has the advantage of analyzing gene sets based on pathways, cell types, and diseases. Enrichr calculates a number of statistics mainly based on input gene set, annotation set as well as background gene set which is usually all the 20.000 human genes. For p-value calculation, it uses a hypergeometric test under the assumption of binomial distribution and independence of any gene belonging to any gene set. Output p-values are then corrected based on the number of tests conducted using Benjamini-Hochberg (BH)

procedure. This tool also calculates an odds ratio as well as a combined score that is multiplication of odds ratio with -log(p-value).

On top of Enrichr tool, gProfiler's g:GOSt (citation) over-representation tool was also used for enrichment profiling (49). Although this tool uses almost the same gene-set databases as Enrichr, it provides an informative graphical representation of the enrichment results with their adjusted p-values. Genes belonging to enriched gene sets were further analyzed using STRING, a protein-protein interaction and functional enrichment tool (50). STRING is a very useful tool to discover how proteins are interacting with each other using both physical and functional mapping as well as a wide range of sources like experimental results, co-expression and co-occurrence analysis and text mining from large databases. Here, genes from top enriching pathways were analyzed using STRING and protein-protein association networks were further grouped using k-means and (Markov Clustering) MCL clustering methods which were then followed by a cluster specific enrichment analysis.

3.7 Burden Testing of Rare Variants

Burden testing is a commonly used strategy in rare variant association studies (RVAS) in order to increase statistical power for detecting rare variants associated with diseases. Due to their extremely low frequencies in the population, single rare variants are underpowered to uncover statistical signals between case and control samples. Even the rare variants with very high effect sizes require huge sample sizes to become statistically significant from linear and logistic regression based association tests. To overcome this limitation, an aggregation based approach was proposed where variants are collapsed to genomic regions which is usually a gene and association tests were conducted based on those regions under different assumptions of variant directionality (8). Aggregation based methods were proved to increase statistical power for discovering novel rare variants and is a common practice for rare variant studies. However, these variant-set tests were originally developed to analyze unrelated cases and controls under the assumption of sample independence. These methods later have been extended to be able to analyze familial data with known pedigree structure by accounting the relatedness between samples.

Here, “pedgene” tool was used for gene-based association tests while accounting for pedigree structure in order to detect the genes that are significantly associated with MS in our family cohort (27). Pedgene is available as a R package and in this study, R version 4.2.0 was used to analyze our samples. Variant selection was done according to the filters mentioned in the variant filtering section and two different variant sets, lenient and strict, were used. However, for the aggregate unit tests, genes with at least two variants were used to calculate burden score according to previous literature. Pedgene calculates a gene burden score and a p-value based on this score where a positive burden score indicates that a given gene increases the disease risk and a gene with negative score can be interpreted as a “protective” factor. Calculated p-values were then adjusted for multiple testing in order to prevent p-value inflation and type 1 error. Here, the Bonferroni correction procedure was used to adjust p-values based on the number of genes tested overally. Finally, the top genes with lowest p-values were chosen to be reported and used in downstream analyses.

4 RESULTS

4.1 Exome Sequencing, Quality Control and Variant Filtering

Exome sequencing of 215 individuals (138 MS, 77 healthy) from 59 families in total was done achieving average 40x coverage in the exonic region. Quality control statistics for each sample fastq file was assessed via the quality report from fastqc tool. Here all samples passed the quality control check as seen in the example fastqc report (Figure X). After the quality control step, variants from all of the samples were called together producing a vcf file consisting of 589,587 variants in total. Based on the minor allele frequencies (MAFs) reported in the gnomAD database and other databases mentioned in the methods section, this variant set comprised 46% common variants ($MAF > 0.05$), 15.5% low-frequency variants ($0.01 < MAF \leq 0.05$), and 38.5% rare variants ($MAF \leq 0.01$).

Mendelian inconsistencies reported based on 34 family trios and variant sites that failed Hardy-Weinberg equilibrium (HWE) at a p-value less than 0.0001 were excluded, resulting in a variant set composed of 510,057 variants. According to our segregation and heritability assumptions, X and Y chromosome variants, as well as all common variants with $MAF > 0.05$, were removed, and a quality control filter based on variant genotype quality was applied, further excluding 220,649 variants. Subsequently, in silico prediction filters for variant deleteriousness, detailed in the methods section, were applied to the remaining high-quality, low-frequency, and rare variants, resulting in a final set comprising 38,851 variants across 14,036 genes.

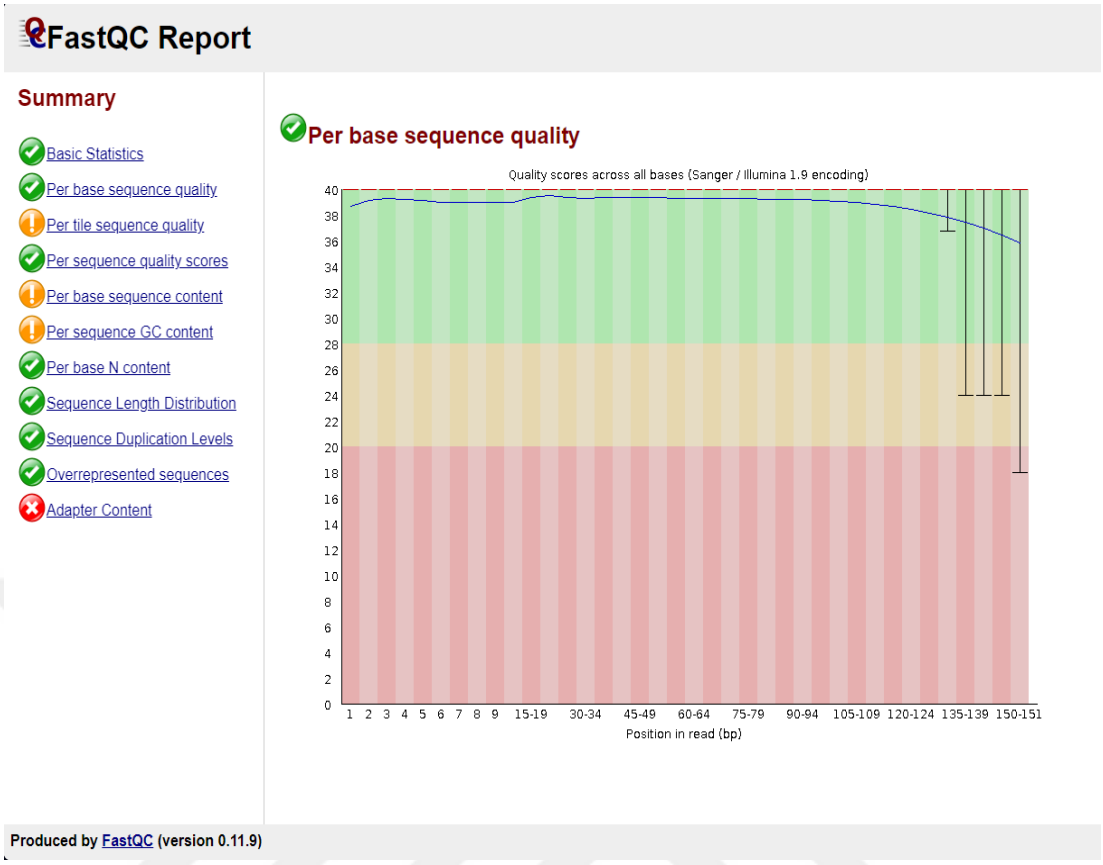


Figure 10. Example passing fastqc quality control report of one the sample analyzed.

4.2 Relatedness Analysis and Detecting Pedigree Inconsistencies

Variant-based relatedness analysis was done using KING software. Aim of this evaluation was both detecting in-family inconsistencies in reported pedigrees as well as unreported relationship between families which can affect segregation analysis. Here, variant-based kinship scores between each sample were calculated and degree of relatedness determined based on the guidelines specified in KING manual were compared to degree of relatedness assigned based on drawn pedigrees (Figure 6).

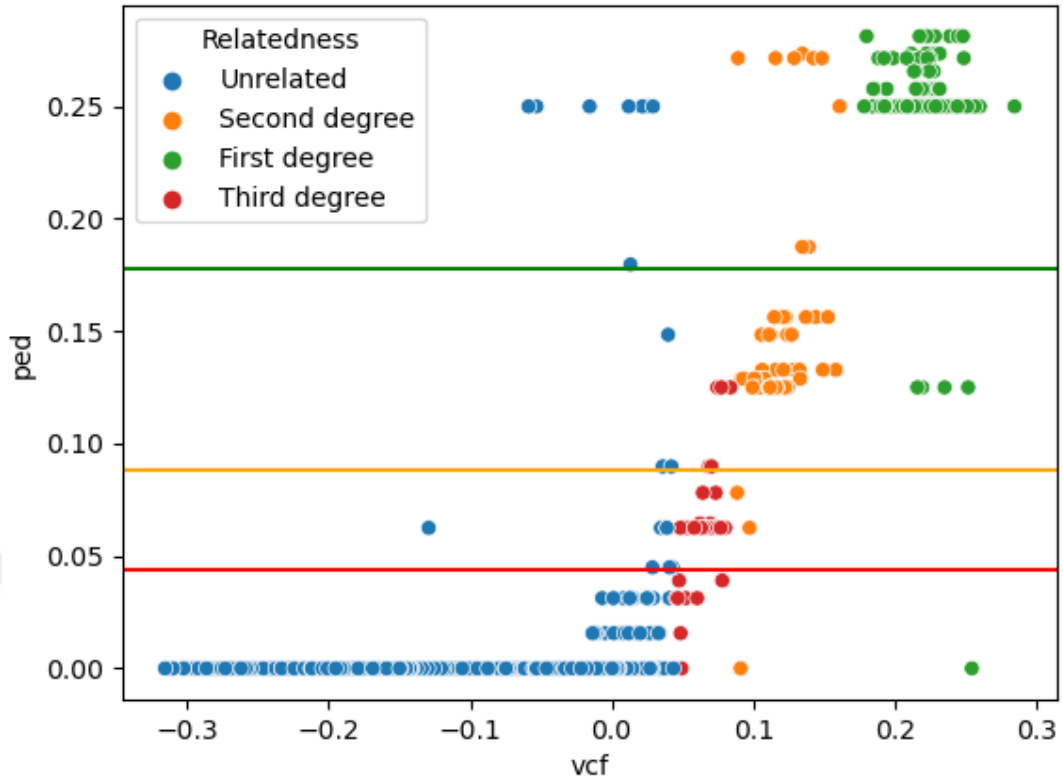


Figure 11. Comparing pedigree kinship estimation to variant-based kinship estimation from vcf file with annotated and colored thresholds based on the assigned relatedness of samples.

Here in Figure 11, a diagonal distribution of points with positive slope were expected for samples with correct relationship estimation. Majority of sample-sample estimations followed the pedigree kinship estimations. Unrelated samples with pedigree kinship estimation 0 were mostly estimated correctly with an expected variance which resulted in a line of blue dots in the bottom of the scatter plot. Threshold lines were used to visually determine the high degree of consistencies such as a cluster of blue dots in the top left section of figures that indicates samples drawn as first degree relatives in pedigrees might not be relatives due to an annotation error. To compensate for the estimation errors and detect the significant inconsistencies, samples with at least two degrees of relationship inconsistency were chosen and reported (Table 2).

Table 2. Kinship inconsistencies between known vs estimated relatedness

sample 1	sample 2	Pedigree kinship	Estimated kinship	Pedigree relatedness	Estimated relatedness
FMS0412	FMS0434	0.148438	0.0399	Second degree	Unrelated
FMS0501	FMS0513	0.179688	0.0133	First degree	Unrelated
FMS0506	FMS0513	0.0898438	0.0361	Second degree	Unrelated
FMS0511	FMS0513	0.0898438	0.0422	Second degree	Unrelated
FMS2001	FMS2007	0.25	-0.0156	First degree	Unrelated
FMS3401	FMS3414	0.25	-0.0531	First degree	Unrelated
FMS3403	FMS3414	0.25	0.0257	First degree	Unrelated
FMS3901	FMS3903	0.25	0.0216	First degree	Unrelated
FMS7401	FMS7402	0.25	0.0292	First degree	Unrelated
FMS7701	FMS7702	0	0.2545	Unrelated	First degree
FMS7701	FMS7703	0.25	0.0121	First degree	Unrelated
FMS7702	FMS7703	0.25	-0.0587	First degree	Unrelated

These inconsistencies were then verified with the clinics where the samples were collected, and the pedigree errors were either resolved or the samples were excluded to ensure a more robust analysis. In addition to intra-family inconsistencies, inter-family relationships were also scrutinized to assess if any unreported relationships existed between the families. No significant relationships between any two families were detected, which is an important factor to consider before proceeding with the segregation analysis. Finally, the reported sex of individuals checked for discrepancies using a method based on the ratio of heterozygote and homozygote alternate variants X chromosome. Only three inconsistencies were found across all of our samples and then corrected accordingly.

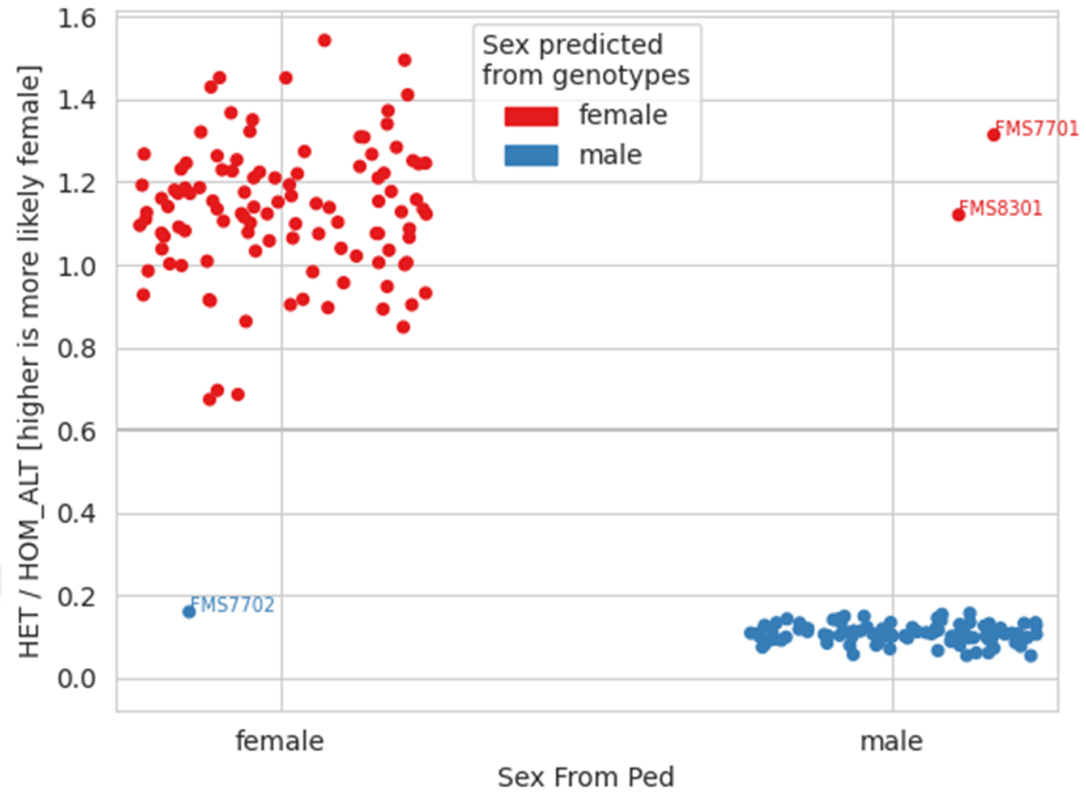


Figure 12. Comparison of reported sex with estimated sex from sample genotypes

4.3 Segregation Analysis for Variant and Gene Prioritization

Segregation analysis was run using custom Python scripts in order to filter down remaining variants from previous filters according to dominant and recessive segregation models and under different penetrance assumptions. The segregating variants were further scrutinized both in variant and in gene level to have a more robust analysis given the complex underlying genetic architecture of MS.

4.3.1 Variants reaching complete segregation with MS in at least two families

Running segregation analysis according to the autosomal dominant inheritance model under complete dominance assumptions and filtering the variants found segregating in at least two families with both MS and healthy samples resulted in 12 variants(Table 3). In addition to families with both MS and healthy samples, families with only MS samples were used as a form of validation to see if the relative segregation pattern was also observed for the given variant.

Table 3. Variants that are completely segregating with MS in families according to the autosomal dominance filter with the genes they belong to, CADD scores, segregation families and their variant identifiers(rsid).

SNP	Existing variation	Gene	CADD	Segregated families
chr12:103982922:A:C	rs61937630	*TDG*	23.8	FMS32', 'FMS77'
chr12:121949060:G:T	rs151062299	*CFAP251*	28	FMS21', 'FMS77'
chr14:77822487:G:C	rs150283628	*ADCK1*	21.8	FMS17', 'FMS77'
chr15:22882981:C:T	rs139635799	*CYFIP1*	16.29	FMS16', 'FMS77'
chr20:47651174:C:T	rs138733364	*NCOA3*	26	FMS32', 'FMS77'
chr22:21001287:C:G	rs145388226	*THAP7*	24.3	FMS32', 'FMS77'
chr2:33280135:G:A	rs61754247	*LTBP1*	21.4	FMS108', 'FMS119'
chr3:4766606:T:G	rs143093165	*ITPR1*	22.2	FMS19', 'FMS28', 'FMS44'
chr4:70839692:G:T	rs769022123	*GRSF1*	18.58	FMS28', 'FMS81'
chr5:169598495:A:T	rs143453400	*SPDL1*	24.1	FMS28', 'FMS77', 'FMS43'
chr6:73771463:C:T	rs41266745	*CD109*	24	FMS17', 'FMS18'
chr9:41174308:C:T	rs62555254	*ZNG1F*	31	FMS81', 'FMS119'

Here 10 variants were found to be segregating in two families and only two variants within *ITPR1* and *SPDL1* genes were segregated in three families as two mixed families and one additional MS only family. Out of these 12 genes, only one, *NCOA3*, were previously associated with MS in a family-based study (6). These variants and genes with *CD109* due to its potential disease relevance were further analyzed and scrutinized in the discussion section using pathway and literature based approaches.

No variants were found completely segregating in more than two families according to the recessive inheritance pattern across our samples.

4.3.2 Gene-level analysis of segregating variants

Gene-level analysis was run by collapsing variants from segregation analysis for all inheritance models as well as penetrance assumptions in order to overcome low

power for detecting MS associated signals due to small probability of finding the exact same variants segregating in multiple families because of their very low MAFs.

For this, first variants that are segregating completely for autosomal dominant inheritance mode were collapsed into the genes they are on and the number of families with at least one variant segregating completely with MS were calculated in order to evaluate and prioritize genes. 86 genes in total were found to carry at least one segregating variant for this inheritance and penetration filter for at least two families. Top genes with the most number of families with observed segregation were *DST* with 5 families then followed by *FBN3*, *AHNAK*, *TTN*, *TDG*, *SPDL1* and *CD93*.

Table 4. Top 9 genes from gene-level segregation analysis of variants from the complete autosomal dominant filter.

Gene	**Families**	**Family count**	**Segregating variants**
DST	['FMS28', 'FMS108', 'FMS18', 'FMS19', 'FMS17']	5	rs201437391', 'rs758532940', 'rs34767818', 'rs75671065', '6:56645899:T:A'
FBN3	['FMS77', 'FMS21', 'FMS119', 'FMS28']	4	rs751652110', 'rs79008349', 'rs757604392', '19:8090227:G:C, 'rs146679363'
AHNAK	['FMS81', 'FMS25', 'FMS28', 'FMS18']	4	11:62527560:A:C', 'rs147636421', 'rs143391519', 'rs114799698'
TTN	['FMS04', 'FMS77', 'FMS31', 'FMS32']	4	rs200782068', 'rs201717871', '2:178534803:A:C', 'rs34819099', 'rs778525620'
TDG	['FMS25', 'FMS77', 'FMS32']	3	rs376531574', 'rs61937630', 'rs61937630'
SPDL1	['FMS77', 'FMS28', 'FMS16']	3	rs150689580', 'rs143453400', 'rs143453400'
CD93	['FMS77', 'FMS81', 'FMS16']	3	rs1282500529', 'rs1046890888', 'chr20:23085831:T:TGGGGGCC'
DUOX2	['FMS19', 'FMS31', 'FMS28']	3	rs372159746', 'rs781680371', 'rs147772932'
KRT14	['FMS31', 'FMS17', 'FMS85']	3	rs556526711', 'rs11551760', 'rs201069984'

In addition to reporting top genes based on the number of families, a literature check was done based on a collection of genes that were associated with MS previously in either a large-scale ascertained population study such as International Multiple Sclerosis Genetics Consortium(IMSGC) or in another family-based study (5,7). Here, five genes were found to be previously associated: *'CUX2'*, *'NCOA3'*, *'PLEC'*, *'SDK1'* and *'SF3B2'*.

Table 5. Genes previously associated with MS with their relative sources and their segregating variants for listed families.

Gene	Families	Family count	Variant IDs(if exist)	Variants	Source
CUX2	['FMS81', 'FMS32']	2	['COSV55649466', nan, nan]	['chr12:111347632:CA:C', 'chr12:111347621:CA:C', 'chr12:111347624:CCAGA:C']	IMSGC loci
NCOA3	['FMS77', 'FMS32']	2	['rs138733364', 'rs138733364']	['chr20:47651174:C:T', 'chr20:47651174:C:T']	Family studies
PLEC	['FMS77', 'FMS32']	2	['rs782080936', 'rs200895043&COSV59667675', 'rs2857824']	['chr8:143917085:G:A', 'chr8:143927053:T:C', 'chr8:143925278:G:A']	IMSGC loci
SDK1	['FMS21', 'FMS10']	2	['rs150519551&COSV67391270', 'rs375950389']	['chr7:4221335:C:T', 'chr7:4208164:G:T']	IMSGC loci
SF3B2	['FMS77', 'FMS28']	2	['rs11554199&COSV99043782', 'rs781551448']	['chr11:66052460:G:T', 'chr11:66068167:C:T']	IMSGC loci

Next, variants segregating incompletely with MS according to the segregation rules explained in detail in the method section were aggregated by the genes they are in. This filtering process aimed to be more lenient and was expected to generate a list that is a superset of the gene list generated from the complete dominant segregation filter to capture the complex and polygenic genetic architecture of MS. 584 genes were captured in the end with at least one variant passing the incomplete segregation filter for the dominant inheritance mode for more than two families with both MS and healthy samples. Here, *TTN* was found to be the top gene with the relative segregation pattern observed in 12 families followed by *PLEC*(7), *KIAA1522*(7), *DST*(7), and *ABCA4*(6) genes.

Literature check for this segregation filter revealed 27 out of these 584 genes to be previously associated with MS either through: a genome or exome wide association study, a family-based study or transcription factor based literature scan detailed in the paper by Mascia et al. (6). Genes that were previously associated with MS in an IMSGC study such as *CUX2*, *CD58*, *ILF3* and *CHD6* were detected through this analysis(Table 6).

Table 6. Intersection of genes that were detected via the incomplete segregation filter and previously associated MS genes from various sources.

Gene	**Variants**	**Source**
PLEC	['chr8:143916829:G:A', 'chr8:143934821:C:T', 'chr8:143921870:G:A', 'chr8:143916574:G:A', 'chr8:143917085:G:A', 'chr8:143927053:T:C', 'chr8:143925278:G:A', 'chr8:143929203:G:T']	IMSGC loci
CHD6	['chr20:41420810:T:A', 'chr20:41421674:G:A', 'chr20:41425364:G:A']	IMSGC loci
LRP2	['chr2:169128828:C:T', 'chr2:169128828:C:T', 'chr2:169174129:C:T']	Families studies
ATF1	['chr12:50814031:C:A', 'chr12:50814339:C:G', 'chr12:50814339:C:G']	Transcription Factors
CUX2	['chr12:111347632:CA:C', 'chr12:111334483:G:C', 'chr12:111347621:CA:C', 'chr12:111347624:CCAGA:C']	IMSGC loci
SF3B2	['chr11:66052460:G:T', 'chr11:66068167:C:T', 'chr11:66052460:G:T']	IMSGC loci
SDK1	['chr7:4221335:C:T', 'chr7:4208164:G:T', 'chr7:3967428:A:G']	IMSGC loci
FCRL3	['chr1:157697857:G:A', 'chr1:157700458:C:T', 'chr1:157697903:C:G']	IMSGC loci
C1orf167	['chr1:11784463:C:T', 'chr1:11784463:C:T']	IMSGC loci
CASKIN2	['chr17:75502571:G:A', 'chr17:75502571:G:A']	IMSGC loci
ETV7	['chr6:36368987:G:A', 'chr6:36368987:G:A']	IMSGC loci, Transcription Factors
ATF5	['chr19:49932603:G:GCTCCTCCCCCCCCCCCC', 'chr19:49933012:G:A']	Transcription Factors
ESPN	['chr1:6425245:T:A', 'chr1:6456182:G:T']	IMSGC loci
CUBN	['chr10:16890491:G:T', 'chr10:16890491:G:T']	Families studies
CD58	['chr1:116544574:T:C', 'chr1:116544574:T:C']	IMSGC loci, German GWAS

Table 6. Intersection of genes that were detected via the incomplete segregation filter and previously associated MS genes from various sources. (continue)

CASKIN2	['chr17:75502571:G:A', 'chr17:75502571:G:A']	IMSGC loci
ETV7	['chr6:36368987:G:A', 'chr6:36368987:G:A']	IMSGC loci, Transcription Factors
ATF5	['chr19:49932603:G:GCTCCTCCCCCCCCCCCC', 'chr19:49933012:G:A']	Transcription Factors
ESPN	['chr1:6425245:T:A', 'chr1:6456182:G:T']	IMSGC loci
CUBN	['chr10:16890491:G:T', 'chr10:16890491:G:T']	Families studies
CD58	['chr1:116544574:T:C', 'chr1:116544574:T:C']	IMSGC loci, German GWAS
KIF21B	['chr1:201008899:G:A', 'chr1:200986865:T:C']	IMSGC loci
ILF3	['chr19:10687390:G:A', 'chr19:10681301:C:G']	IMSGC loci
KCNH8	['chr3:19518018:G:C', 'chr3:19518018:G:C']	IMSGC loci
NCOA3	['chr20:47651174:C:T', 'chr20:47651174:C:T']	Families studies
LIMK2	['chr22:31278338:C:G', 'chr22:31262219:C:T']	IMSGC loci
MYB	['chr6:135192533:A:G', 'chr6:135203350:T:A']	IMSGC loci
OSBP2	['chr22:30893490:G:C', 'chr22:30893490:G:C']	IMSGC loci
P2RX7	['chr12:121156133:C:T', 'chr12:121167570:G:A']	Families studies
RREB1	['chr6:7231563:G:A', 'chr6:7230023:C:G', 'chr6:7231128:G:T']	IMSGC loci
USP34	['chr2:61295019:T:A', 'chr2:61281137:T:C']	IMSGC loci
RNF213	['chr17:80383030:G:T', 'chr17:80369804:A:G']	Families studies
PIK3IP1	['chr22:31292302:G:A', 'chr22:31291079:C:T']	IMSGC loci

4.4 Gene Enrichment Analysis and Pathway Grouping

Genes from both complete and incomplete segregation filters were further analyzed using a range of enrichment tools as well as a plethora of gene ontology and pathway databases. The top results were reported mainly according to the false discovery rate(FDR) corrected p-values from hypergeometric or fisher's exact tests that are commonly used statistical tests for overrepresentation analysis for given genes and gene-sets. Pathway grouping was done using clustering methods in order to break

down relatively large and complex enrichment terms to have a more complete understanding of underlying gene and variant signals.

4.4.1 Enrichment analysis over the genes from complete segregation filter

To begin with, 86 genes from the complete segregation filter were analyzed with Enrichr and gProfiler pathway enrichment tools. Enrichr produced pathway enrichment for more than 25 gene-sets databases but the results from only a number of commonly used databases were scrutinized due to the repetitiveness of the enrichment results for a majority of databases.

Table 7. Top pathways enriched for the genes from complete dominant segregation filter with adjusted p-value < 0.05 from the REACTOME 2022 database in Enrichr.

Term	Overlap	Adjusted P-value	Odds Ratio	Genes
Type I Hemidesmosome Assembly R-HSA-446107	3/9	0.002188713825	118.4940476	DST;ITGB4;PLEC
Cell Junction Organization R-HSA-446728	5/86	0.005073974187	14.92923818	SDK1;DST;ITGB4;FLNC;PLEC
Collagen Formation R-HSA-1474290	5/90	0.005073974187	14.22381636	COL16A1;DST;ITGB4;COL6A5;PLEC
Assembly Of Collagen Fibrils And Other Multimeric Structures R-HSA-2022090	4/57	0.009239318065	18.05864969	DST;ITGB4;COL6A5;PLEC
Cell-Cell Communication R-HSA-1500931	5/120	0.01192031537	10.49734889	SDK1;DST;ITGB4;FLNC;PLEC

ID	Source	Term ID	Term Name	p_{adj} (query_1)
8	GO:CC	GO:0099512	supramolecular fiber	2.536×10^{-7}
12	REAC	REAC:R-HSA-44...	Type I hemidesmosome assembly	3.908×10^{-5}
9	GO:CC	GO:0030056	hemidesmosome	4.565×10^{-4}
3	GO:BP	GO:0030216	keratinocyte differentiation	6.453×10^{-4}
13	REAC	REAC:R-HSA-44...	Cell junction organization	9.415×10^{-4}
1	GO:MF	GO:0051015	actin filament binding	1.119×10^{-3}
4	GO:BP	GO:0009653	anatomical structure morphogenesis	1.565×10^{-3}
5	GO:BP	GO:0031581	hemidesmosome assembly	2.024×10^{-3}
6	GO:BP	GO:0030239	myofibril assembly	1.316×10^{-2}
7	GO:BP	GO:2000344	positive regulation of acrosome reaction	1.647×10^{-2}
10	GO:CC	GO:0070161	anchoring junction	3.015×10^{-2}
11	GO:CC	GO:0036156	inner dynein arm	3.660×10^{-2}
2	GO:MF	GO:0004687	myosin light chain kinase activity	4.972×10^{-2}

Figure 13. gProfiler enrichment results for the completely segregating genes across multiple gene ontology databases.

Top pathways enriching for the genes passing complete segregation filter were "Type I Hemidesmosome Assembly R-HSA-446107", "Cell Junction Organization R-HSA-446728" and Collagen Formation R-HSA-1474290 from the REACTOME pathway database as shown in detail with the genes overlapping with the given pathway in Table 7. In addition to that, IL-2/STAT5 signaling pathway was significantly enriched after the p-value adjustment for multiple testing in MSigDB Hallmark 2020 database(28). Enrichment analysis from gProfiler produced similar results as it includes a set of common databases with Enrichr. Here, "supramolecular fiber" was the most significantly enriched term for cellular component(CC) category while keratinocyte differentiation and actin filament binding were the most significantly enriched term for biological process(BP) and molecular function(MF) categories. Only pathway term enriching from the KEGG database was "gastric acid secretion". Overall, "hemidesmosome assembly" was the most frequently enriched pathway term for the gene hits from complete segregation filtering across different databases examined in both Enrichr and gProfiler.

4.4.2 Exploring the signals from incomplete penetrance variant filtering

Next, pathway enrichment and overrepresentation analyses were conducted using genes identified through a more relaxed incomplete dominant filter to capture a broader range of signals from our family cohort. Enrichr and gProfiler were executed with the same settings as in the previous analysis. The results were reported in various formats to better understand the pathway patterns emerging from the segregation analysis.

Table 8. Top REACTOME pathway enrichment results for the genes from the incomplete segregation filter.

Term	Overlap	Adjusted P-value	Combined Score	Genes
Extracellular Matrix Organization R-HSA-1474244	29/291	8.36E-06	71.14895	*COL18A1;FBN3;LAMA5;COLGALT2;COL15A1;PTPRS;COL14A1;ITGB4;LAMA1;COL11A1;LTBP4;NID1;LTBP1;COL19A1;TNN;ITGB7;CAPN1;ITGB6;DST;LAMB1;HSPG2;COL5A1;COL5A3;ITGA10;COL7A1;COL6A3;COL6A6;COL6A5;PLEC*
Collagen Chain Trimerization R-HSA-8948216	11/44	1.48E-05	193.6538	*COL18A1;COL15A1;COL5A1;COL14A1;COL11A1;COL5A3;COL7A1;COL6A3;COL6A6;COL6A5;COL19A1*
Collagen Formation R-HSA-1474290	15/90	1.48E-05	115.2103	*COL18A1;COLGALT2;COL15A1;DST;COL14A1;ITGB4;COL11A1;COL19A1;COL5A1;COL5A3;COL7A1;COL6A3;COL6A6;COL6A5;PLEC*
Assembly Of Collagen Fibrils And Other Multimeric Structures R-HSA-2022090	12/57	1.69E-05	149.2525	*COL18A1;COL15A1;DST;COL5A1;ITGB4;COL11A1;COL5A3;COL7A1;COL6A3;COL6A6;COL6A5;PLEC*
Collagen Biosynthesis And Modifying Enzymes R-HSA-1650814	12/67	8.78E-05	108.1738	*COL18A1;COLGALT2;COL15A1;COL5A1;COL14A1;COL11A1;COL5A3;COL7A1;COL6A3;COL6A6;COL6A5;COL19A1*
Laminin Interactions R-HSA-3000157	7/22	3.17E-04	206.7458	*COL18A1;LAMA5;ITGB4;LAMA1;LAMB1;NID1;HSPG2*

Filtering out the enriched pathways with an adjusted p-value greater than 0.05 resulted in a set of six pathways. The most significant term, with the lowest p-value, was "Extracellular Matrix Organization R-HSA-1474244," where 29 genes from our segregation set overlapped with this pathway. The next four terms were primarily related to collagen formation and biosynthesis, driven mainly by the "*COL*" collagen family genes. The term "Laminin Interactions R-HSA-3000157" was also significantly enriched, with the highest combined score calculated by multiplying the adjusted p-value with the estimated odds ratio. Seven genes from our gene list were found in this specific pathway, with laminin-related genes such as *LAMA5*, *LAMA1*, and *LAMB1* being the main enrichment drivers.

Table 9. Significant KEGG pathway enrichment results for the genes from the incomplete segregation filter.

Term	**Overlap**	**Adjusted P-value**	**Combined Score**	**Genes**
ECM-receptor interaction	17/88	1.32E-07	174.8004	*LAMA5;VWF;ITGB4;LAMA1;LAMB2;TNC;LAMB1;THBS2;HSPG2;FRAS1;TNN;ITGA10;COL6A3;ITGB7;COL6A6;ITGB6;COL6A5*
Focal adhesion	22/201	1.28E-05	67.7926	*MYLK2;LAMA5;SHC3;VWF;ITGB4;LAMA1;LAMB2;TNC;VEGFC;LAMB1;THBS2;MYLK4;TNN;ITGA10;COL6A3;FLNB;ITGB7;COL6A6;PIP5K1C;FLNC;ITGB6;COL6A5*
Protein digestion and absorption	12/103	0.00361	44.7323	*SLC9A3;COL18A1;COL15A1;COL5A1;COL14A1;COL5A3;COL11A1;COL7A1;COL6A3;COL6A6;COL6A5;COL19A1*
Human papillomavirus infection	22/331	0.01795	19.7134	*LAMA5;JAG1;VWF;ITGB4;LAMA1;LAMB2;FZD9;TNC;UBR4;LAMB1;THBS2;RBL2;PPP2R1A;TNN;ITGA10;COL6A3;ATM;ITGB7;COL6A6;ITGB6;COL6A5;ATR*

“ECM-receptor interaction” term from the KEGG pathway database was found to be the top enriching term with 17 overlapping genes in total given the total of 88 genes in this pathway. Clustergrams were created using Enrichr in order to visualize the

common and different overlapping input genes for the enriched terms for both REACTOME and KEGG pathways analyses (Figure 13).

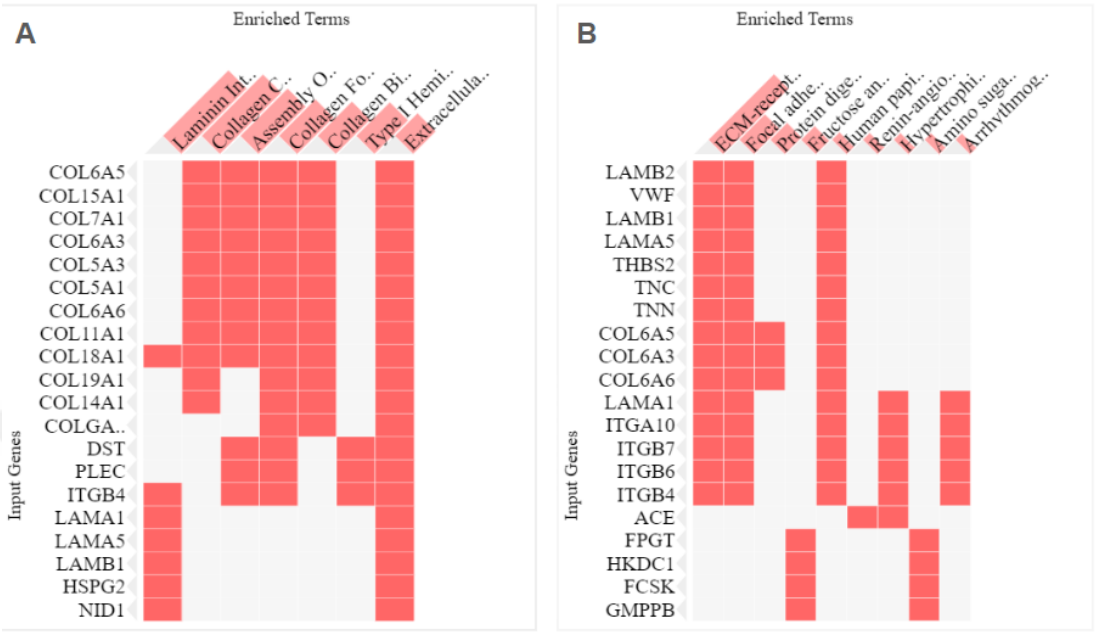
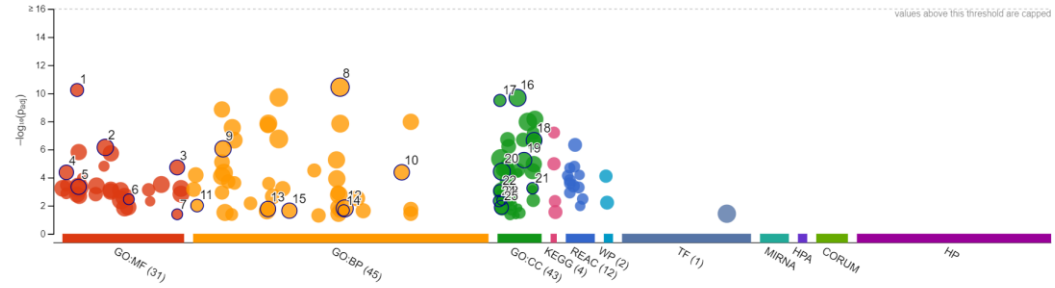


Figure 14. Clustergrams of the top enriching pathways and their relative input genes for REACTOME(A) and KEGG(B).



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0005201	extracellular matrix structural constituent	5.875×10^{-11}
2	GO:MF	GO:0030554	adenyl nucleotide binding	7.134×10^{-7}
3	GO:MF	GO:0140657	ATP-dependent activity	1.901×10^{-5}
4	GO:MF	GO:0003779	actin binding	4.261×10^{-5}
5	GO:MF	GO:0005509	calcium ion binding	4.593×10^{-4}
6	GO:MF	GO:0045505	dynein intermediate chain binding	3.530×10^{-3}
7	GO:MF	GO:0140658	ATP-dependent chromatin remodeler activity	4.124×10^{-2}
8	GO:BP	GO:0048856	anatomical structure development	3.670×10^{-11}
9	GO:BP	GO:0007155	cell adhesion	8.975×10^{-7}
10	GO:BP	GO:0097435	supramolecular fiber organization	4.214×10^{-5}
11	GO:BP	GO:0001578	microtubule bundle formation	9.894×10^{-3}
12	GO:BP	GO:0051128	regulation of cellular component organization	1.575×10^{-2}
13	GO:BP	GO:0030029	actin filament-based process	1.729×10^{-2}
14	GO:BP	GO:0050982	detection of mechanical stimulus	2.183×10^{-2}
15	GO:BP	GO:0034330	cell junction organization	2.281×10^{-2}
16	GO:CC	GO:0042995	cell projection	2.114×10^{-10}
17	GO:CC	GO:0005604	basement membrane	3.246×10^{-10}
18	GO:CC	GO:0099512	supramolecular fiber	2.254×10^{-7}
19	GO:CC	GO:0070161	anchoring junction	5.745×10^{-6}
20	GO:CC	GO:0005856	cytoskeleton	3.778×10^{-5}
21	GO:CC	GO:0098636	protein complex involved in cell adhesion	5.760×10^{-4}
22	GO:CC	GO:0005581	collagen trimer	8.577×10^{-4}
23	GO:CC	GO:0005858	axonemal dynein complex	4.081×10^{-3}
24	GO:CC	GO:0001533	cornified envelope	4.674×10^{-3}
25	GO:CC	GO:0005788	endoplasmic reticulum lumen	1.338×10^{-2}

version	e111_eg58_p18_f463989d
date	8/8/2024, 12:49:08 AM
organism	hsapiens

g:Profiler

Figure 15. gProfiler pathway enrichment results for incompletely segregating genes across different pathway databases and a list of significantly enriched pathways.

Compiling the results from different enrichment tools and a wide range of gene-set databases revealed notable enrichment of extracellular matrix (ECM) structure and interaction-related pathways. To assess the interactions between the genes from these pathways and gain a deeper understanding of the underlying biology of the gene hits from segregation analysis, pathways with the most significant enrichment were analyzed using the STRING protein interaction and functional enrichment tool.

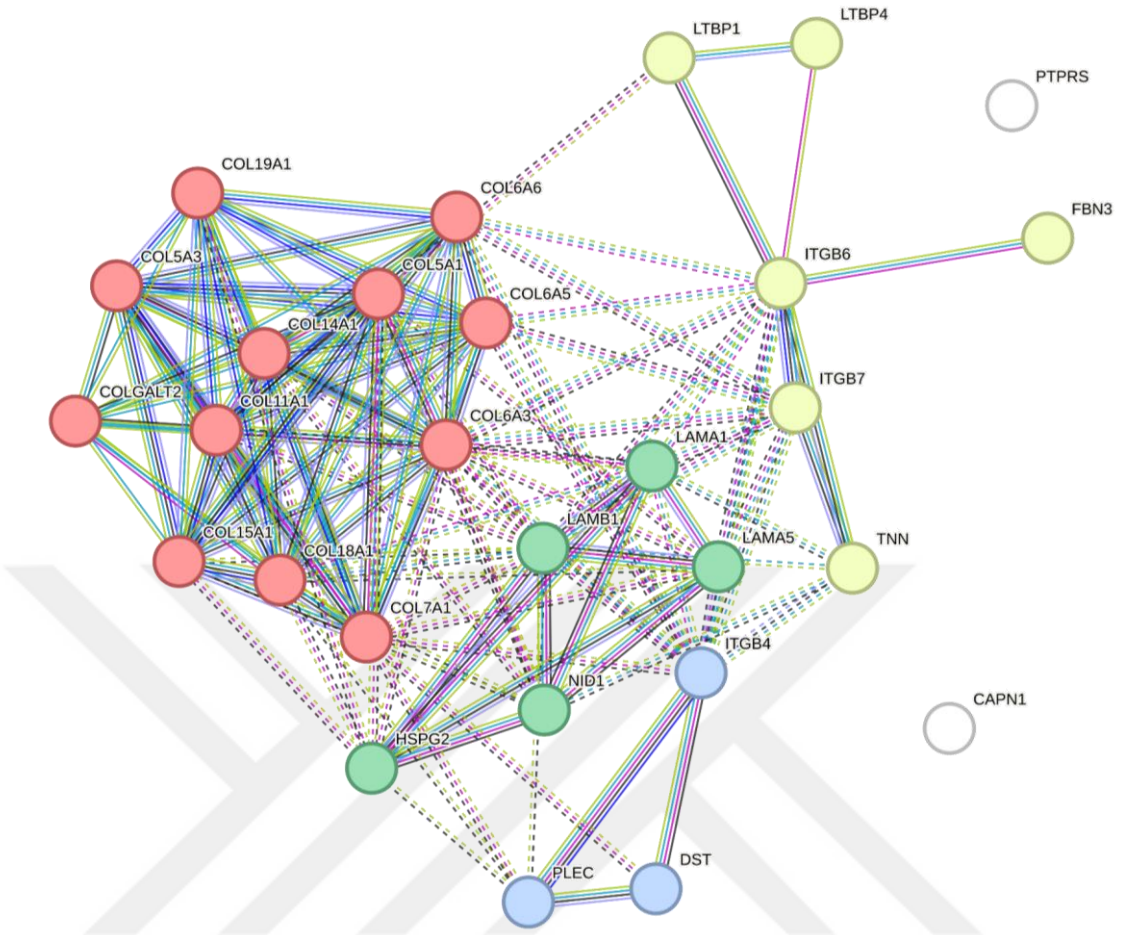


Figure 16. STRING functional protein interaction network for ECM organization pathway and clustering with Markov Cluster(MCL) algorithm.

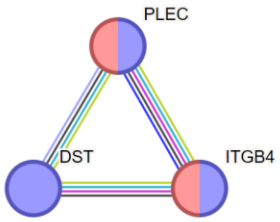
A compiled list of genes from ECM-related pathways was then analyzed using STRING's web-based analysis tools. Interaction maps were created based on both functional and physical protein associations. Furthermore, the generated networks were broken down using the MCL clustering algorithm with an inflation parameter of 4, which resulted in four distinct sub-networks, each color-coded accordingly. Notably, only two genes were not part of any cluster or sub-cluster. Detailed MCL clustering results for this pathway can be found in Table X.

Table 10. Clusters, cluster descriptions, and the genes constituting these clusters for the analysis of the genes from ECM related pathways

cluster color	**gene count**	**primary description**	**secondary description**	**tertiary description**	**protein names**
Red	12	Collagen chain trimerization	-	-	*COL7A1, COL6A5, COL6A6, COL6A3, COL11A1, COL5A1, COL14A1, COL15A1, COL18A1, COL5A3, COL19A1, COLGALT2*
Yellow	6	Molecules associated with elastic fibres	Microfibril	-	*TNN, ITGB7, ITGB6, FBN3, LTBP4, LTBP1*
Green	5	Laminin interactions	Basement membrane	laminin-1 complex,	*LAMA1, LAMB1, LAMA5, HSPG2, NID1*
Blue	3	Hemidesmosome assembly	Myelination in peripheral nervous system	-	*ITGB4, DST, PLEC*

The largest cluster from MCL clustering analysis which was highlighted with red color in Figure 15. consists of 12 genes and this sub-network was mainly enriched for the "Collagen chain trimerization" pathway term. This cluster was followed by the clusters primarily enriched for "Molecules associated with elastic fibres"(yellow) and "Laminin interactions"(green) which consist of 6 and 5 genes relatively. The smallest cluster which consists of 3 genes shown in blue color was primarily enriched for the term "Hemidesmosome assembly" and notably enriched for "Myelination in peripheral nervous system" term secondarily. The fourth cluster was further analyzed to identify the genes driving its primary and secondary functional descriptions (Figure 16).

A



B

	Biological Process (Gene Ontology)			
GO-term	*description*	*count in network*	*strength*	*false discovery rate*
GO:0031581	Hemidesmosome assembly	3 of 6	3.52	1.03e-06
GO:0022011	Myelination in peripheral nervous system	2 of 31	2.63	0.0256
GO:0009611	Response to wounding	3 of 444	1.65	0.0256

Figure 17. Color coded sub-network of the 4th cluster described in Table 10 (A) and the enriched gene ontology terms with specific colors highlighting the genes driving the enrichment results (B).

4.5 Burden Analysis of Rare Variants

Burden analysis of rare variants was conducted using the *pedgene* R package. This analysis utilized the variant set previously generated for the segregation analysis, which will be referred to as the "*damaging*" set in this section. In addition to the damaging variant set, which encompassed low frequency and rare variants(LFRV) with a high probability of causing disruptive impacts on proteins, a "*control*" set of LFRVs were chosen. For the *control* set, synonymous variants that are annotated as "LOW" impact from VEP variant consequences were chosen. Full pedigree information was integrated into this analysis in order to correctly calculate and correct for the sample relatedness.

In this analysis, genes were used as the units for variant aggregation. Consistent with common practices in the literature, only genes containing at least two variants were included in the burden testing. The *damaging* variant set originally comprised 38,851 variants across 14,036 genes. However, after filtering out genes with only a single variant, these numbers were reduced to 34,347 variants across 8,360 genes. A *control* variant set, consisting of 30,744 synonymous variants, was similarly processed. Before running the burden analysis, genes with only one variant were excluded, and only those genes that overlapped with the *damaging* gene set were retained for the final analysis of the *control* set.

Table 11. Top 20 associated genes from the burden analysis of the *damaging* set sorted by the burden p-value.

Gene	Chr	n.variants	Burden statistic	Burden p-value
CLIC6	21	3	-3.810494466	0.0001386891172
CHIC2	4	2	-3.628830979	0.0002847075622
FBN1	15	4	-3.563410046	0.0003660681769
STAG3	7	7	-3.397472609	0.0006801139625
HSD17B7	1	3	-3.360256861	0.000778700381
GRM2	3	3	-3.33801824	0.0008437821019
CORO6	17	3	-3.286731917	0.001013572747
FGF6	12	2	-3.283276901	0.001026078252
AQP5	12	2	-3.283276901	0.001026078252
P2RX5	17	4	-3.218477063	0.001288732774
ADGRB1	8	12	3.217339184	0.001293855384
PCDH1	5	6	-3.156346036	0.001597592051
SLC45A1	1	15	-3.13602897	0.001712522864
ACADS	12	3	-3.126736785	0.001767581015
LTN1	21	3	-3.114273907	0.001843982194
C1QTNF12	1	2	-3.089348665	0.002005958729
BRD1	22	3	-3.004808692	0.002657479408
PNPLA5	22	2	-3.001564811	0.002685958564
HGD	3	3	-3.00099167	0.002691019262
NR4A1	12	2	-2.997824932	0.002719138229

After applying p-value correction for multiple testing across the 8,360 genes (with a significance threshold of $\alpha = 0.05/8360 = 5.9e\text{-}06$), no gene was found to be statistically significant. Among the top 20 genes reported in Table 11, only one gene, *ADGRB1,* exhibited a positive burden statistic, indicating that the variant burden for this gene was higher in the MS samples compared to control samples. However, this result was not statistically significant.

Table 12. Top 20 genes with a positive burden statistic.

Gene	Chr	n.variants	Burden statistic	Burden p-value
ADGRB1	8	12	3.217339184	0.001293855384
GFPT2	5	3	2.622491076	0.00872895538
B4GALNT4	11	4	2.539976458	0.01108599306
GRM1	6	4	2.539976458	0.01108599306
PARP3	3	3	2.453377098	0.01415218909
NUCB1	19	2	2.441837937	0.01461270479
SIGLEC10	19	4	2.423596742	0.01536765952
ANKRD44	2	4	2.39758327	0.01650362932
NBAS	2	7	2.355809343	0.01848240107
ZNF875	19	5	2.353717091	0.01858675111
SPINK5	5	6	2.345935782	0.01897937723
LAMB1	7	5	2.337965604	0.01938903252
MSLN	16	3	2.307598332	0.0210214874
MAPK4	18	5	2.22712234	0.02593909853
KRT14	17	4	2.222886944	0.02622342439
SKA3	13	6	2.222007626	0.02628279026
PPP1R12C	19	5	2.221762309	0.02629937318
HLA-DRB1	6	2	2.192735386	0.02832644969
RIN2	20	13	2.191836206	0.02839133575
PDPR	16	11	2.158322208	0.03090278929

Subsequently, a comparative analysis was conducted to evaluate the most significant genes with positive burden scores from the damaging variant set. This involved comparing their burden scores and burden p-values to those of the same genes within the control variant set. Additionally, genes that exhibited marginally higher burden scores and lower burden p-values in comparison to the control burden analysis were further assessed. This assessment was based on prior literature associations as well as findings from our segregation analysis.

Table 13. Genes showed increased variant burden for the *damaging* filter compared to the *control* filter.

Gene	Damaging b.stat	Damaging p-val	Control b.stat	Control p-val	Literature	Segregation
B4GALNT4	2.5400	0.0111	0.9837	0.3253		
GRM1	2.5400	0.0111	-0.4288	0.6680		
PARP3	2.4534	0.0142	-0.8885	0.3743		
NUCB1	2.4418	0.0146	0.8532	0.3936		
SIGLEC10	2.4236	0.0154	1.3315	0.1830		
ANKRD44	2.3976	0.0165	-1.0735	0.2831		Yes
ZNF875	2.3537	0.0186	1.3326	0.1827		
SPINK5	2.3459	0.0190	NA	NA		
LAMB1	2.3380	0.0194	-0.6353	0.5252		Yes
MSLN	2.3076	0.0210	0.5827	0.5601		
MAPK4	2.2271	0.0259	0.8772	0.3804		
KRT14	2.2229	0.0262	0.0949	0.9244		Yes
SKA3	2.2220	0.0263	NA	NA		
PPP1R12C	2.2218	0.0263	1.1948	0.2322		
HLA-DRB1	2.1927	0.0283	NA	NA	IMSGC loci	
RIN2	2.1918	0.0284	-0.9371	0.3487		
PDPR	2.1583	0.0309	0.0273	0.9783		
ARFGAP3	2.1416	0.0322	0.6823	0.4950		
MST1	2.1380	0.0325	0.6941	0.4876		
PCDH9	2.1174	0.0342	-0.8352	0.4036		
AKAP12	2.1073	0.0351	1.0013	0.3167		
PALLD	2.0979	0.0359	NA	NA		Yes
RLF	2.0974	0.0360	0.7334	0.4633		Yes
PAK6	2.0971	0.0360	0.4071	0.6840		
HIPK1	2.0967	0.0360	NA	NA		
ARHGEF40	2.0966	0.0360	-1.1689	0.2425		
ASPSCR1	2.0943	0.0362	-0.7238	0.4692		
SPI1	2.0943	0.0362	NA	NA	IMSGC loci	

Table 13. Genes showed increased variant burden for the *damaging* filter compared to the *control* filter. (continue)

ALPK3	2.0910	0.0365	-0.0959	0.9236		
URGCP	2.0834	0.0372	-0.5478	0.5838		
IGF2BP2	2.0810	0.0374	NA	NA		
GPR146	2.0787	0.0376	-0.3778	0.7056		
FRMD4A	2.0739	0.0381	0.4860	0.6270		
ATP5MC2	2.0731	0.0382	NA	NA		
KIFBP	2.0689	0.0386	NA	NA		
NLRC5	2.0609	0.0393	0.4299	0.6673	IMSGC loci	
MICAL2	2.0464	0.0407	0.7459	0.4557		Yes
FZD9	2.0316	0.0422	0.0949	0.9244		Yes
ADGRD1	2.0302	0.0423	-1.3505	0.1769		
KRT6A	2.0281	0.0426	0.6148	0.5387		
PMF1	2.0196	0.0434	NA	NA		
ALDH1L2	2.0031	0.0452	0.0949	0.9244		
SEZ6L	2.0008	0.0454	0.6478	0.5171		Yes
PLD2	1.9947	0.0461	-0.6086	0.5428		
SCAF8	1.9852	0.0471	-1.3669	0.1717		

Comparing the burden results between the damaging and control filters, and retaining only those genes with marginal burden differences, led to the identification of 45 genes. Among these, three genes—*HLA-DRB1*, *SPI1*, and *NLRC5*—had been previously associated with multiple sclerosis (MS) in studies conducted by the IMSGC or GWAS. Additionally, eight genes—*ANKRD44*, *LAMB1*, *KRT14*, *PALLD*, *RLF*, *MICAL2*, *FZD9*, and *SEZ6L*—were detected in our segregation analysis. The top genes with a variant burden difference and lowest burden p-values for the *damaging* set were *B4GALNT4*, *GRM1*, *PARP3*, *NUCB1* and *SIGLEC10*.

4.6 HLA Typing and Examining *HLA-DRB1*15:01* Variant in Our Families

HLA typing was done using the HLA-HD program for each sample separately using sample fastq files. In total, 29 HLA haplotypes were typed up to 6 digits using this program for all of 215 samples. The final results were merged using a custom script written in Python. Here, *HLA-DRB1*15:01* was our main focus as it's one of the strongest MS risk variants (51).

Table 14. Families with complete or acceptable segregation of *HLA-DRB1*15:01* variant with MS and the relative inheritance modes.

Family	Number of samples	Number of samples following/not following the relative segregation pattern	Notes on inheritance modes
FMS78	4	4/0	Recessive segregation pattern observed where 2 healthy parents carrying one copy of this variant and MS children carrying two copies
FMS01	7	6/1	Autosomal dominant
FMS08	7	6/1	Autosomal dominant
FMS09	6	5/1	Autosomal dominant
FMS10	5	4/1	Autosomal dominant
FMS115	5	4/1	Autosomal dominant

The *HLA-DRB1*15:01* variant was identified as fully segregating in a single family, FMS78, following a recessive inheritance pattern, as illustrated in Figure 17. Furthermore, five additional families—FMS01, FMS08, FMS09, FMS10, and FMS115—demonstrated a pattern consistent with autosomal dominant inheritance. In these cases, only one individual deviated from the expected segregation pattern, either by presenting as a healthy individual carrying a copy of the *HLA-DRB1*15:01* variant or by being an MS patient without any copies of the variant.

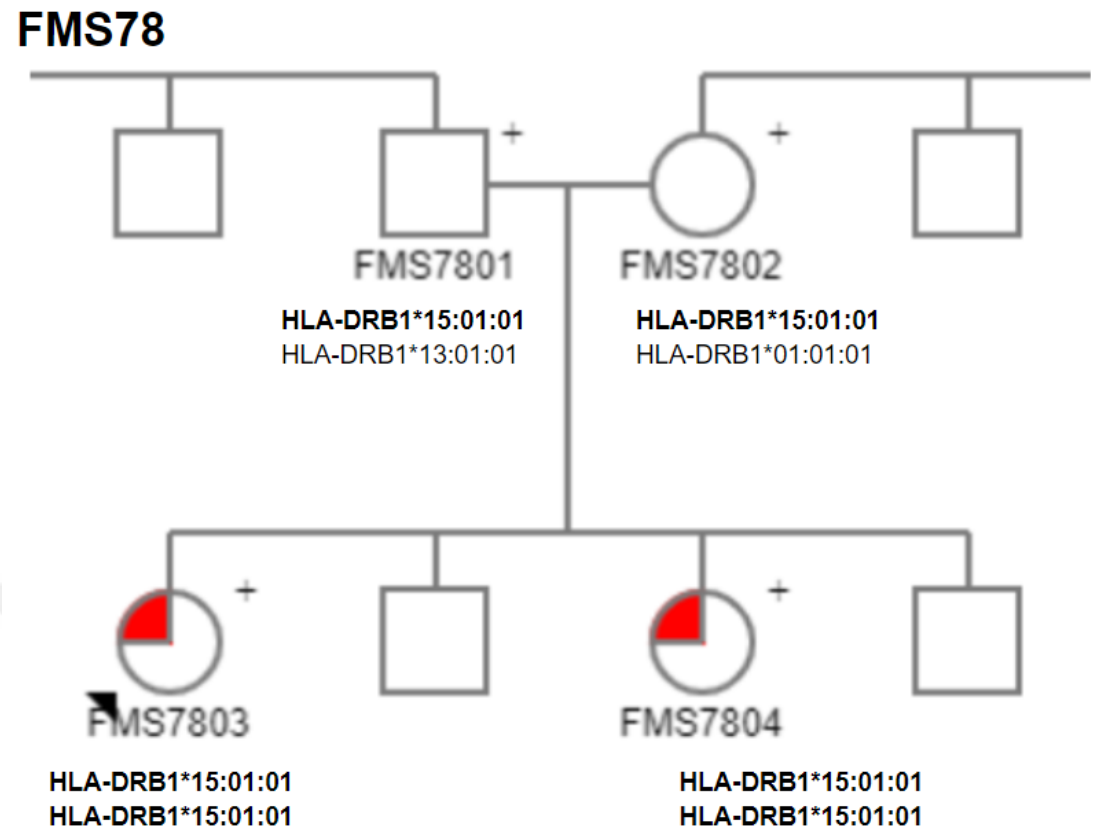


Figure 18. Recessive segregation of *HLA-DRB1*15:01* in FMS78.

Furthermore, the distribution of the *HLA-DRB1*15:01* variant was analyzed by categorizing individuals as homozygous carriers, heterozygous carriers, or non-carriers. This analysis was conducted to compare the presence of this variant between healthy individuals and those diagnosed with MS. In this analysis, the distribution of other MS-associated HLA variants, including *HLA-DRB1*03:01*, *HLA-DRB1*04:05*, *HLA-DRB1*13:03* and *HLA-DRB1*08:01*, was also examined alongside *HLA-DRB1*15:01*. Furthermore, two protective HLA variants reported in the literature, *HLA-DRB1*11 and HLA-DRB1*01:01*, were analyzed for assessing a more broad assessment of HLA variants.

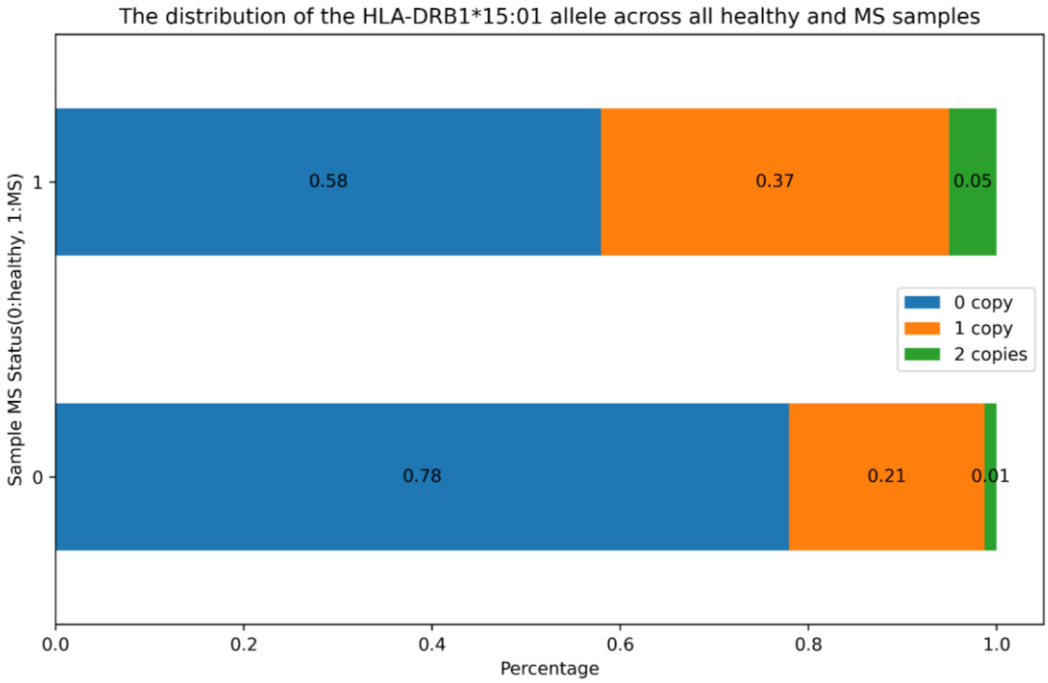


Figure 19. Bar plot illustrating the distribution of the *HLA-DRB1*15:01* variant across all samples.

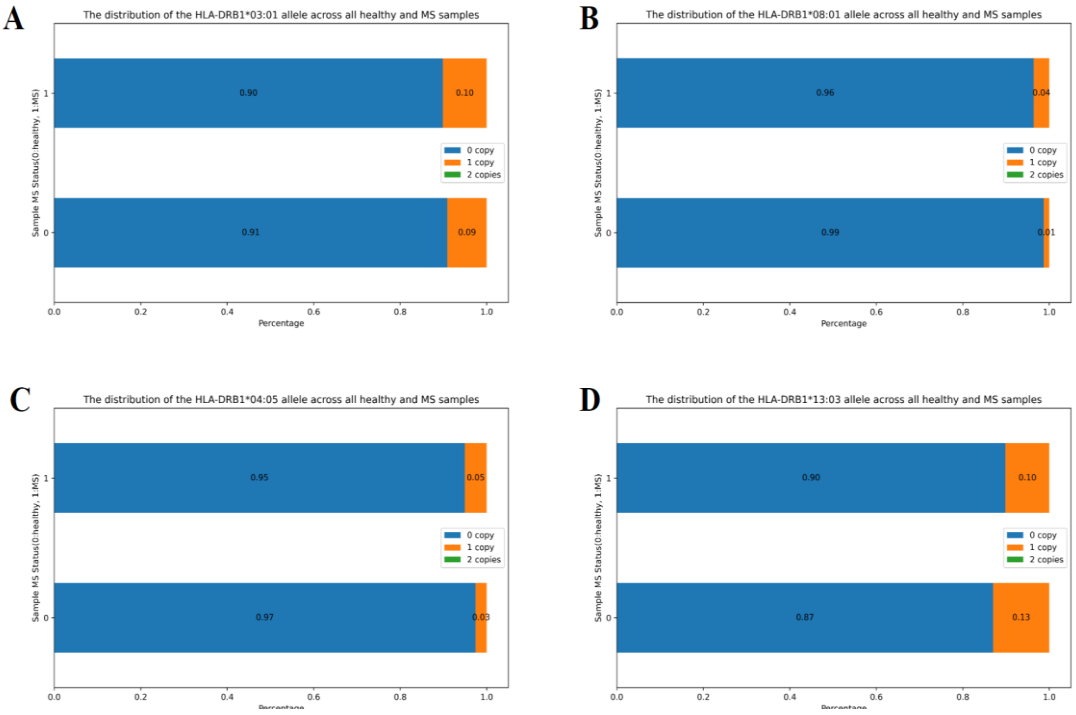


Figure 20. Bar plots illustrating the distribution *HLA-DRB1*03:01*(A), *HLA-DRB1*08:01*(B), *HLA-DRB1*04:05*(C) and *HLA-DRB1*13:03*(D) MS-susceptibility variants.

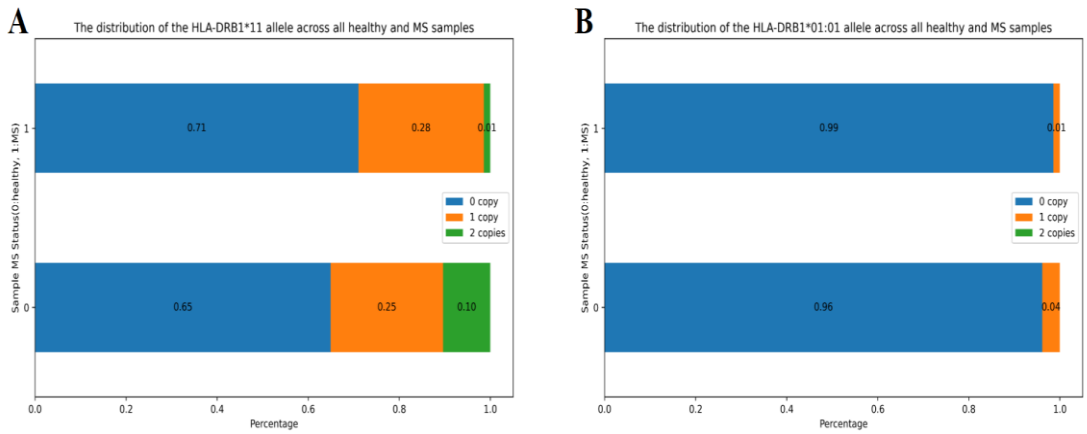


Figure 21. Bar plots illustrating the distribution *HLA-DRB1*11*(A) and *HLA-DRB1*01:01*(B) protective HLA variants for MS.

The distribution analysis revealed that more than half of the MS samples did not carry the *HLA-DRB1*15:01* variant. Approximately 37% of the MS samples were found to carry the variant in a heterozygous form, while the remaining 5% were homozygous for it. For the healthy samples, the distribution was as follows: 78% did not carry the *HLA-DRB1*15:01* variant, 21% were heterozygous, and 1% were homozygous for this variant.

For the other HLA variants associated with MS susceptibility in the literature (Figure 20), no significant segregation patterns in the families or notable distribution differences were observed between the MS and healthy sample groups.

Among the two protective MS variants analyzed, only *HLA-DRB1*11* exhibited a notable distribution. Specifically, 10% of the healthy samples were homozygous for this variant, while 25% were heterozygous. Meanwhile across all MS samples, 28% carried the variant in a heterozygous form, while 1% were homozygous.

5 DISCUSSION

5.1 QC and Relatedness Analysis

Strict quality filtering was a key part of our pipeline to make sure the variants we kept were reliable. Even with good coverage (~40x), raw variant calls can include false positives due to alignment issues, sequencing errors, or sample contamination. Checking for Mendelian inconsistencies (taking advantage of having familial sequencing data) in the 34 trios helped us catch variants that didn't follow expected inheritance patterns—often a sign of bad genotypes. Filtering out variants that failed Hardy-Weinberg equilibrium (HWE) at $p < 1e\text{-}4$ also helped clean up the data, as extreme deviations usually point to technical problems rather than biology.

We also removed common variants and those on sex chromosomes to focus on rare coding changes on autosomes, which are more likely to segregate with disease in families. This step cut down noise from background variation and kept the analysis focused on potentially high-impact changes.

After all filters were applied—including functional predictions—we were left with around 39,000 variants across 14,000 genes. This is a manageable set for downstream analysis and much more likely to contain real disease-related signals compared to the original 589,000. It shows how much filtering is needed to get from raw sequencing data to something biologically meaningful.

Making sure the pedigree information is accurate was a crucial part of preparing for our segregation analysis. Since our goal is to find rare variants that segregate with MS across different families, we rely heavily on the assumption that these families are truly independent. If two families are actually related but we treat them as separate, a variant shared between them could look like it's segregating in multiple families—when in reality, it might just be inherited from a shared ancestor. That kind of mistake could give us false confidence in a variant and lead us to prioritize something that's not actually important for the disease.

To avoid this, we used the KING software to estimate genetic relatedness directly from the variant data. This gave us kinship scores between every pair of individuals, which we compared to the expected relationships in the pedigrees. Most of the time, the genetic data matched what was reported, but there were some inconsistencies. In a few cases, samples listed as unrelated showed kinship scores suggesting they were actually first- or second-degree relatives. In others, supposed close relatives didn't show enough genetic similarity. We focused on pairs where the difference was two degrees or more, followed up with the clinics, and either corrected the errors or removed the affected samples. These fixes helped us avoid errors that could throw off our family-based results.

Just as important, we looked for any hidden relatedness *between* families. If, for example, two families turned out to be distantly related, we might mistakenly count a shared variant as two independent segregation events. This would inflate our confidence in that variant. Fortunately, we didn’t find any unexpected inter-family relationships, which supports our assumption that the families in our dataset are genetically distinct. We also ran a sex check using X chromosome data to catch possible sample mix-ups or labeling errors. By looking at the ratio of heterozygous to homozygous alternate genotypes on the X chromosome, we could infer whether a sample was genetically male or female. Out of all samples, only three had mismatches between reported and inferred sex. These were corrected before continuing. While this might seem like a minor detail, sex mismatches can lead to problems in trio analysis or throw off interpretation when analyzing X-linked variants.

In the end, these checks gave us more confidence that our data is clean and well-structured, which is especially important for segregation analysis where pedigree structure and sample identity really matter. Without this step, we’d risk basing our findings on incorrect assumptions about family structure or sample identity, and that could seriously affect the conclusions we draw from the data.

5.2 HLA Analysis

HLA typing was a key part of our analysis because of the well-known role of certain HLA alleles especially *HLA-DRB1*15:01* in MS risk. Using the HLA-HD tool, we were able to type 29 different HLA haplotypes at high resolution (up to 6 digits) for all 215 individuals. This level of detail is important, since subtle differences at the 4- or 6-digit level can mean very different functional impacts when it comes to HLA biology and disease association. The typing process worked well across samples, and we merged the results with a custom script for easier comparison across individuals and families.

As expected, *HLA-DRB1*15:01* stood out in our analysis. In one family (FMS78), the variant showed full segregation with MS in a way that matched recessive inheritance—meaning affected individuals were homozygous for the allele. This kind of clean segregation is rare and offers strong support for the role of this allele in that family. In five other families (FMS01, FMS08, FMS09, FMS10, and FMS115), we saw segregation patterns consistent with dominant inheritance. Each of those families had just one person who didn't follow the expected pattern, either a healthy carrier or an MS patient without the allele. This kind of "almost complete" segregation is still biologically interesting and suggests the allele may contribute to disease risk in those families, but it also shows that the picture is more complex—likely involving other genetic or environmental modifiers.

At the cohort level, we looked at the distribution of *HLA-DRB1*15:01* among MS cases and healthy controls. About 37% of MS patients were heterozygous, and only 5% were homozygous for the variant. Interestingly, more than half of MS patients didn't carry the allele at all, which agrees with other studies showing that while this allele increases risk, its effect size is still moderate . Among healthy individuals, the allele was much less common—only 1% were homozygous and 21% were heterozygous which supports its role as a risk allele.

We also checked the other HLA variants that have been associated with MS risk or protection in previous studies. Variants like *HLA-DRB1*03:01*, *04:05*, *13:03*, and *08:01* didn't show any clear segregation patterns in our families, nor did they stand

out in their distribution between MS and healthy individuals. It's possible that the effect sizes of these alleles are smaller or that their impact depends on interactions with other genes or environmental exposures that we weren't able to capture here due to the scope of study.

Among the two HLA variants reported as protective—*HLA-DRB1*11* and *HLA-DRB1* 01:01*—only *DRB1*11* showed some interesting results. It was more common among healthy individuals, particularly in homozygous form (10% in healthy vs. 1% in MS), which matches earlier reports. But even here, the effect wasn't dramatic, and many healthy individuals didn't carry it at all.

Overall, while *HLA-DRB1*15:01* clearly plays a role in some families, it's also clear that many MS cases in our dataset are not explained by this allele alone. This highlights the importance of looking beyond well-known risk loci and exploring rare, possibly family-specific variants elsewhere in the genome.

5.3 Segregation Analysis

Segregation analysis helped us narrow down the variant list from earlier filters by focusing variant co-segregation with MS within families. Since MS likely involves multiple rare variants with variable penetrance, we approached this step using both strict and more relaxed inheritance assumptions. The main goal here was to find variants—and ultimately genes—that show consistent segregation with the disease across multiple families, which could point to real genetic contributors worth investigating further.

Under a strict autosomal dominant model, we found 12 variants that fully segregated with MS in at least two families. This is a strong signal, especially given how rare these variants are and how unlikely it is for the exact same variant to segregate by chance in multiple unrelated families. Two of these variants, located in *ITPR1* and *SPDL1*, showed segregation in three families (two with mixed MS/healthy individuals and one with only MS individuals), which adds extra weight to their potential role. Out of the 12 genes with segregating variants, only *NCOA3* has been

previously linked to MS in a family-based study, which makes it a promising hit. The other genes, including *CD109*, were kept for follow-up based on biological relevance and potential links to immune or neurological pathways.

On the recessive side, no variants passed the full segregation filter in more than one family. That's not too surprising given the lower chance of recessive inheritance patterns showing up in multiple small pedigrees, especially with rare variants.

To get around the low power issue from focusing only on variant-level segregation, we also looked at the gene level by collapsing segregating variants across families. This helped highlight genes that might not carry the exact same variant across different families but still show a consistent pattern of having disease-linked variants. Using this approach, 86 genes had at least one segregating variant in two or more families under the dominant model. Some of the top genes—*DST*, *FBN3*, *AHNAK*, and TTN—showed segregation in several families, making them strong candidates for follow-up. Getting *TTN* as a top gene is not surprising given its very large size and hits we have observed in this analysis for this gene is likely false positive and not real associations with MS. We also checked these genes against prior MS studies, and five (*CUX2*, *NCOA3*, *PLEC*, *SDK1*, *SF3B2*) came up as previously associated, which helps validate our approach.

Finally, we ran a more lenient filter to catch variants that mostly—but not perfectly—segregate with MS, to reflect the reality of complex traits like this. This expanded the list to 584 genes showing relative segregation in at least two mixed families. *TTN* stood out again, with segregation in 12 families, followed by others like *PLEC*, *KIAA1522*, and *DST*. Importantly, 27 of these genes have been mentioned in earlier MS studies or large-scale datasets like the IMSGC, including *CD58*, *CUX2*, *ILF3*, and *CHD6*. This overlap with known MS genes supports the idea that our filtering approach captures biologically meaningful signals, even in a limited family-based dataset.

This analysis gave us a focused set of variants and genes showing real patterns in MS families. These findings lay the groundwork for the next step—digging into the biology and known pathways behind the strongest hits.

5.4 Rare Variant Burden Analysis

Burden analysis provides an orthogonal approach to segregation analysis we conducted, offering a different perspective on identifying genetic signals in MS by aggregation and performing a regression analysis using generalized linear models to take genetic relatedness into account in the form of genetic relatedness matrix (GRM) between each of our samples (8). While segregation analysis focuses on family-specific inheritance patterns, burden analysis aggregates rare variants across all samples to detect whether the combined effect of these variants in specific genes is significantly higher in MS cases compared to controls. This method helps validate findings from segregation analysis by looking for broader patterns of genetic risk, and it's especially useful for capturing commonalities in complex diseases like MS, where multiple rare variants with modest effects may contribute to disease risk.

For our burden analysis, we utilized a "damaging" set of rare and low-frequency variants (LFRVs) with high likelihoods of disrupting protein function, which were previously filtered in the segregation analysis. These were compared to a "control" set consisting of synonymous variants, which are expected to have almost no functional impact. Importantly, we integrated full pedigree information with GRMs generated from both pedigree information as well as from the variants to ensure the analysis accounted for familial relationships which is critical for correcting sample relatedness and avoiding confounding factors.

The burden analysis was conducted on genes that contained at least two variants, as this is the standard approach in the literature to avoid power issues from analyzing genes with too few variants (6,26,27). After applying several filters, including excluding genes with a single variant and adjusting for multiple testing, no gene in the damaging set reached statistical significance at the threshold of $\alpha = 5.9e\text{-}06$. However, in the top 20 genes, we did observe one gene, *ADGRB1*, with a higher burden in MS samples, although this result was not statistically significant.

To further refine our results, we compared the burden scores for genes in the damaging set with those in the control set. By focusing on genes that showed marginally higher burden scores and lower p-values in the damaging set, we identified

45 notable genes . Interestingly, three genes (*HLA-DRB1*, *SPI1*, and *NLRC5*) were previously associated with MS in large-scale studies like the IMSGC. In addition to that, eight genes that had been detected by our segregation analysis (*ANKRD44*, *LAMB1*, *KRT14*, *PALLD*, *RLF*, *MICAL2*, *FZD9*, and *SEZ6L*) also showed higher burdens, strengthening the biological relevance of these genes in the context of MS risk.

The top genes with the most promising burden differences and the lowest p-values included *B4GALNT4*, *GRM1*, *PARP3*, *NUCB1*, and *SIGLEC10*. These findings provide novel candidates for future investigation and could be potential targets for further studies exploring the genetic background of MS susceptibility.

Concluding this section while no single gene reached statistical significance in our burden analysis, the identification of several genes with increased burden in MS cases, particularly those previously linked to MS in the literature and those found by our segregation analysis, suggests that this approach is a valuable orthogonal complement to segregation analysis. It helps uncover genetic patterns that may not be immediately obvious from family-based inheritance studies alone, offering a more holistic view of the genetic architecture of MS.

5.5 Pathway Enrichment Analysis

The gene enrichment analysis for the genes from our segregation analysis have provided very interesting insights into the genetic and molecular mechanisms underlying MS risk and progression. We have identified a number of biologically relevant pathways, particularly related with ECM structure and its components by using pathway enrichment tools such as Enrichr and gProfiler. The pathways we have found here are especially important for the structural integrity of tissues and play a critical role in tissue remodeling during disease progression. The top pathways enriching for the genes from both complete and incomplete segregation filters were collagen formation, ECM organization, and cell-matrix interactions, suggesting that disruptions in these pathways could have significant impact for MS susceptibility and progression.

Pathways related with laminins and their cellular interactions were one of the top pathways we have found in our enrichment analysis with key genes like *LAMA5*, *LAMA1*, and *LAMB1* from the segregation analysis as well as our burden comparison. Laminins are structural components of the ECM and they regulate interactions with other matrix proteins and cell surface receptors. Potential disruptions in laminins and the pathways they are involved in could have an important role in the disease risk by potentially altering the structure of blood-brain barrier and the movement of immune cells. Laminins were not the only structural components enriched from our pathway analysis. Collagen formation and organization pathways were also found enriching and these enrichments were driven by a number of genes encoding collagens from the segregation analysis. Collagens play an integral role in the structural integrity in a wide variety of tissues including the nervous system, and their dysfunction could contribute to MS risk.

Another pathway we have found enriching for our top genes from the segregation analysis was the hemidesmosome assembly. The components of this pathway plays an important role in maintaining cell adhesion to the ECM and disruptions in this pathway can lead to problems in cell attachment and signaling which both can be a factor in MS disease progression. This further supports the previous findings on the important role of alterations in cellular adhesion and cytoskeletal dynamics in MS.

Overall, the results from enrichment analysis highlights the genetic complexity and multifactorial nature of MS. In the next section, we will deep dive into these findings and explore the top pathways and genes in the context of their known and potential relevance with MS which could allow ust to draw a more broad picture of how LFRVs we have detected in our families can potentially increase MS risk.

5.6 Deep Dive into Pathways and Genes Found Associated with MS

Understanding the biological mechanisms underlying MS is very important for identifying potential drug targets and extending our knowledge of disease etiology. Our segregation and burden analysis revealed several pathways and gene signals that may play significant roles in MS pathogenesis. In this section, we are going to explore

the biological relevance of the top pathways from our analysis based on the published literature.

5.6.1 Relevance of laminin-related genes in multiple sclerosis

Based on our pathway enrichment analysis of segregation and burden analysis results, we focused on three laminin-related genes: *LAMA5*, *LAMA1*, and *LAMB1*. We found that rare and low-frequency variants in these genes are linked to a higher risk of MS. These variants could impact the structure or function of laminin subunits, possibly affecting disease onset.

LAMA5 and *LAMB1* seem to play a role in maintaining the integrity of the blood-brain barrier (BBB). Variants in *LAMA5* and *LAMB1* can affect downstream pathways that might eventually alter BBB function, potentially making it more permeable and allowing immune cells to infiltrate, which is a known alteration in MS. One study found out that the expression of *LAMA5* significantly increased in the cerebral cortex of experimental autoimmune encephalomyelitis (EAE) mice compared with healthy controls. EAE is one of the most commonly used models to study MS. This increase in the gene expression was more noticeable in female EAE mice, with *LAMA5* expression levels increasing 3.85-fold compared with a 3.04-fold increase in males. These findings suggest that laminins play an important role in the MS pathophysiology by causing structural changes in the brain associated with neuroinflammation (52). Laminins also play a role in oligodendrocyte function and myelination process thus alterations in these genes could impact the integrity of CNS (53).

Laminins also interact with endothelial integrins to contribute structural maintenance of BBB. Alteration in these interactions caused by the effects of genetic variants could result in a more permeable barrier. Previous studies show that MS patients often have increased rates of laminin degradation in the CNS, which could increase inflammation by making migration of T lymphocytes and macrophages into CNS easier. Studying how these variants affect protein structures or expression patterns which may potentially translate into weaker BBB structure could reveal potential therapeutic targets for reducing neuroinflammation in MS (53,54).

Alterations in Laminin-511 complex that consists of *LAMA5* and *LAMB1*, could have an impact on BBB stability as described above via disrupting interactions between laminins and endothelial integrins which can weaken BBB allowing immune cells infiltration and causing chronic inflammation. High impact variants in *LAMA5* and *LAMB1* that result in dysfunctional laminins could also impact the remyelination process. On the other hand, *LAMA1*, which encodes the laminin α1 subunit, doesn't seems to have a direct link to MS susceptibility and disease progression based on the current literature but studying it is role in CNS development and membrane integrity might shed light on it is role in MS (55,56).

More research is needed to characterize the function and impact of the variants detected in *LAMA5*, *LAMB1* and *LAMA1* genes as well as other genes found in related pathways that can potentially affect MS susceptibility and to understand their functional effects. Studying how these variants impact protein interactions, structural integrity of BBB and ECM, and migration of immune cells will allow us to better understand how laminin pathways are involved in MS. By fully characterizing laminin gene variants and their effects, we can work toward potential targeted therapies that modulate ECM interactions and protect CNS integrity.

5.6.2 ECM and hemidesmosome assembly

Our segregation analysis and downstream pathway enrichment analysis have pointed to pathways related with ECM integrity and hemidesmosome assembly which is mainly driven by the *DST*, *ITGB4*, and *PLEC* genes. Proteins encoded by these genes modulate cytoskeletal organization and cell-matrix interactions which are crucial for maintaining cell stability and functionality in CNS. A highly organized and maintained ECM in CNS is critical for maintaining the optimal environment for neuronal and glial cells to function properly. In MS pathophysiology, disruptions in ECM structure have been well-studied, with major molecular and structural changes observed in ECM components within MS lesions (57).

One of the most significant alterations is the upregulation of basement membrane components such as laminins (as described in the previous section) and collagen IV,

whose levels are found to be consistently elevated in MS lesions and the surrounding areas. Laminins interact with integrins, including the α6β4 complex (formed by *ITGB4* protein), variations in this protein can alter the nature of these interaction patterns which might result in ECM remodeling that could be linked to inflammation in MS. Another pathway we detected with our pathway enrichment analysis is the hemidesmosome assembly pathway. Previous studies suggest that this pathway might play an important role in the CNS, given that many of its components—such as integrin α6β4, plectin, and dystonin—are well expressed in neural cells. Integrin α6β4, formed by *ITGB4*, acts as a receptor protein for laminins in the basement membrane and promotes cells to stick to the ECM. In the CNS, *ITGB4* is found to be expressed in astrocytes, Schwann cells, neurons, and neural stem cells, so its role probably goes beyond just structural support to include modulating cell signaling during inflammation (53,58).

In MS, ECM dysregulation and faulty cell-matrix interactions can alter the BBB structure dramatically and induce neuroinflammation. Also, the cytoskeletal structure in CNS cells relies on plectin and dystonin proteins, which are encoded by *PLEC* and *DST*, respectively. Alterations and disruptions within these proteins could destabilize the cytoskeleton, affecting the integrity of axons and glial cell function—both of which are disrupted in MS. *DST* gene was also identified as one of the top candidate genes in an Italian family-based study by Mascia et al. that used a similar segregation and burden based approach to identify potential genes associated with MS (6,59).

Interestingly, plectin protein coded by the *PLEC* gene is also a vitamin D receptor super enhancer. Vitamin D deficiency is a well known environmental factor for MS susceptibility and genetic variants in this gene that alters the structure and function of plectin protein can potentially disrupt vitamin D binding and increase MS susceptibility (60).

Our findings here highlight the potential role of *DST*, *ITGB4*, and *PLEC* in increased MS risk, mainly through their structural role in cell adhesion, cytoskeletal stability, and ECM interactions. Future research confirming these links through

functional studies, particularly looking at how disruptions of hemidesmosome components can contribute to inflammation and neurodegeneration in MS would allow us to discover novel therapeutic targets up and downstream of these candidate genes.

5.6.3 Collagen formation

Collagen formation and assembly was another noticeable pathway that we detected in the pathway enrichment analysis. These collagen assembly pathways are mainly driven by the variants segregating within collagen genes that encode a wide range of collagen family proteins such as *COL18A1*, *COLGALT2* and *COL15A1*.

The association between collagen assembly and MS appears to be through both structural and immune modulation functions. In MS lesions, perivascular fibrosis has been observed which is characterized by an accumulation of fibrillar collagens. Structural networks formed by these fibrillar collagens around blood vessels have been previously associated with chronic CNS inflammation. The perivascular deposition of collagens in lesions suggests that collagen synthesis outweighs its degradation which results in an imbalance and accumulation. Studies comparing collagen levels between demyelinated active lesions to inactive lesions found significant increase in collagen levels, especially for type IV collagen levels. Involvement of collagen deposition in MS further supports the role and importance of ECM in MS pathophysiology since collagens, especially type IV collagen is a major component of basal membrane (BM) which is a specialized form of ECM (61,62).

Although collagen formation and deposition are known markers in active MS lesions, it is not clear how the variants segregating with MS patients in our cohort, which are mostly deleterious, can increase MS susceptibility. One potential explanation could be that disruptions in collagen formation alter the BM structure in a way that allows the movement of immune cells into the CNS. However, further functional studies are required to understand the role of these variants and the impact of lowered expression of collagen genes on MS susceptibility and the modulation of immune cells in the CNS.

5.6.4 *ITPR1* and *CD109*

In addition to identifying enriched pathways from our top genes with variants showing both complete and incomplete segregation patterns, we found variants that completely segregate in two or more independent families. Given their very low frequency in the population, detecting these variants in multiple families may indicate a true association with the disease, especially for variants with large effect sizes in genes associated with MS through relevant pathways. Our final list includes 12 genes with at least one LFRV segregating in more than one family. Based on our literature search, we prioritized two of these genes, *ITPR1* and *CD109*, due to their involvement in pathways and molecular mechanisms relevant to MS.

The inositol 1,4,5-trisphosphate receptor type 1 (*ITPR1*) gene encodes a receptor protein for inositol 1,4,5-trisphosphate (IP3) and upon binding of IP3, it modulates the release of calcium ions which makes *ITPR1* an important part of the calcium signalling pathway. Calcium signalling is a critical molecular process for a plethora of cellular functions and cell components such as activation of immune cells, neurotransmitter release in neurons, and oligodendrocytes. Regulation of intracellular calcium levels is vital for the development and function of the nervous and immune systems that are also focal to MS pathogenesis. Disruptions of *ITPR1* function via deleterious variants that might negatively impact these vital cellular processes, eventually contributing to MS susceptibility and pathology (63).

ITPR1 has not been directly associated with MS previously, but there are studies that have linked this gene to autoimmune cerebellar ataxia, a condition where the immune system attacks the cerebellum part of the brain, causing impaired balance and coordination. Although these findings do not indicate a direct link between *ITPR1* gene and MS, they might suggest that *ITPR1* plays an important role in neuronal function via autoimmune reaction, which is also compromised in MS patients. Further supporting this potential association is the presence of autoantibodies against *ITPR1* in patients with other neurological autoimmune conditions, highlighting its role in neuroinflammation, a critical component of MS pathology. Therefore, investigating the molecular function of *ITPR1* and its disruptions via damaging variants found by

our analysis in MS susceptibility and disease progression could potentially identify new therapeutic targets for MS (64).

Another promising candidate gene with a variant segregating completely in more than one family associated with elevated MS risk is *CD109*. *CD109* protein is known to be involved in immune response and inflammation by modulating TGF-β and NF-κB pathways. CD109 negatively regulates TGF-β signalling by inhibiting some of the key components of this pathway. This is especially important as TGF-β signalling plays a pivotal role in inflammation. Some in vivo studies done using mice models have shown that overexpression of CD109 resulted in decreased inflammatory response which is consistent with the known regulatory role of this protein (65). However, one study has shown an opposite effect where increased CD109 expressions leads to inflammation, claiming negative regulation of TGF-β signalling is outweighed by increased ECM production in fibroblasts with elevated CD109 levels (66). Thus they suggested inhibition of CD109 as a potential therapeutic approach for rheumatoid arthritis (RA) which is another autoimmune disease. Although the association between inflammation and CD109 levels can vary, this variability likely stems from the highly dynamic and cell/tissue-specific nature of immune system regulation. Investigating CD109 in the context of MS risk could reveal novel aspects of this gene and its regulatory effects on the immune system.

6 CONCLUSION

In this study, we investigated the LFRV contribution to MS susceptibility using one of the largest familial MS cohorts to date consisting of 215 individuals from 59 multiply affected MS families from Turkey. Here, we aimed to discover novel LFRVs that have potentially disruptive effects on proteins that would explain familial nature of MS susceptibility in our families by whole-exome sequencing all of our study participants to get high quality variant information for exonic regions. By analyzing a comprehensive, not fully European admixture cohort, we aimed to address at least a fraction of the missing heritability in MS, which is approaching a detection limit due to large-scale studies primarily consisting of European samples.

Our multifaceted analysis approach by combining a traditional segregation analysis, family-based rare variant burden testing and subsequent pathway enrichment analysis allowed us to not only prioritize candidate genes and variants but also discover known and unique molecular mechanisms underlying increase MS susceptibility. Our analysis of human leukocyte antigen (HLA) loci revealed that meanwhile some of the known MS associated HLA variants in MS risk such as *HLA-DRB1*15:01* has shown some degree of segregation with MS status in our cohort which is consistent with existing literature - we couldn't detect a noticeable segregation of most of the HLA variants associated with MS from the literature probably due to their small effect sizes. However, our focus on LFRVs with potentially moderate to large effect sizes revealed novel insights, particularly related to ECM structure and interactions as well as blood-brain barrier. Here laminin-related genes such as *LAMA5* and *LAMB1* and structurally important genes like *DST*, *PLEC*, and *COL16A1* emerged as potential contributors to MS pathogenesis, underscoring the significance of the extracellular matrix in maintaining neural integrity and the blood-brain barrier's role in modulating immune responses. Finally, we highlighted two important candidate genes with variants full segregating in multiple families: *ITPR1* and *C109* with potential molecular links to MS susceptibility.

Our findings support that rare and low-frequency variants contribute substantially to MS susceptibility and studying diverse ancestral groups is especially important where the genetic architecture of MS remains poorly characterized for detecting new LFRVs that can potentially reveal novel disease biology. On top of that, the pathway enrichment analysis approach provided valuable context, suggesting that disruptions in the structural integrity of neural and vascular tissues may be central to MS development. These findings support a model in which both genetic predisposition and compromised neural architecture contribute to MS, paving the way for further functional validation studies which can potentially reveal novel drug targets.

In conclusion, we believe this study offers novel insights into the genetic architecture of MS by identifying potential disease-associated LFRVs utilizing a unique familial cohort. Even though MS is a very complex disease with highly polygenic genetic nature combined with multifactorial environmental factors to create disease susceptibility, these results can be a valuable resource for future research efforts on exploring the functional consequences of the identified genes and pathways. By incorporating large-scale GWAS and family-based studies from multi-ancestry cohorts, we would have a more comprehensive understanding of MS heritability and the broader genetic landscape of autoimmune diseases.

7 REFERENCES

1. Harroud A, Stridh P, McCauley JL, Saarela J, van den Bosch AMR, Engelenburg HJ, et al. Locus for severity implicates CNS resilience in progression of multiple sclerosis. Nature. 2023;619(7969).
2. Ghasemi N, Razavi S, Nikzad E. Multiple sclerosis: Pathogenesis, symptoms, diagnoses and cell-based therapy. Cell J. 2017;19(1).
3. Alfredsson L, Olsson T. Lifestyle and environmental factors in multiple sclerosis. Cold Spring Harb Perspect Med. 2019;9(4).
4. Thompson AJ, Baranzini S, Geurts J, Hemmer B, Ciccarelli O. Seminar: Multiple Sclerosis. The Lancet. 2018;391(10130).
5. Patsopoulos NA, Baranzini SE, Santaniello A, Shoostari P, Cotsapas C, Wong G, et al. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. Science. 2019;365(6460).
6. Mascia E, Clarelli F, Zauli A, Guaschino C, Sorosina M, Barizzone N, et al. Burden of rare coding variants in an Italian cohort of familial multiple sclerosis. J Neuroimmunol. 2022;362.
7. Mitrovič M, Patsopoulos NA, Beecham AH, Dankowski T, Goris A, Dubois B, et al. Low-Frequency and Rare-Coding Variation Contributes to Multiple Sclerosis Risk. Cell. 2018;175(6).
8. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: Study designs and statistical tests. Am J Hum Genet. 2014;95(1).
9. Peterson RE, Kuchenbaecker K, Walters RK, Chen CY, Popejoy AB, Periyasamy S, et al. Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. Cell. 2019 Oct 17;179(3):589–603.
10. Loma I, Heyman R. Multiple Sclerosis: Pathogenesis and Treatment. Curr Neuropharmacol. 2011;9(3).
11. van Noort JM, Baker D, Kipp M, Amor S. The pathogenesis of multiple sclerosis: a series of unfortunate events. Clin Exp Immunol. 2023;214(1).
12. Waubant E, Lucas R, Mowry E, Graves J, Olsson T, Alfredsson L, et al. Environmental and genetic risk factors for MS: an integrated review. Ann Clin Transl Neurol. 2019;6(9).
13. Handel AE, Williamson AJ, Disanto G, Handunnetthi L, Giovannoni G, Ramagopalan SV. An updated meta-analysis of risk of multiple sclerosis following infectious mononucleosis. PLoS ONE. 2010;5(9).
14. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era - Concepts and misconceptions. Nat Rev Genet. 2008;9(4).
15. Barry CJS, Walker VM, Cheesman R, Smith GD, Morris TT, Davies NM. How to estimate heritability: a guide for genetic epidemiologists. Int J Epidemiol. 2023;52(2).
16. Westerlind H, Ramanujam R, Uvehag D, Kuja-Halkola R, Boman M, Bottai M, et al. Modest familial risks for multiple sclerosis: A registry-based study of the population of Sweden. Brain. 2014;137(3).
17. Hollenbach JA, Oksenberg JR. The immunogenetics of multiple sclerosis: A comprehensive review. J Autoimmun. 2015;64.

18. Martin R, Sospedra M, Eiermann T, Olsson T. Multiple sclerosis: doubling down on MHC. Trends Genet. 2021;37(9).
19. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265).
20. Wang Y, Kanai M, Tan T, Kamariza M, Tsuo K, Yuan K, et al. Polygenic prediction across populations is influenced by ancestry, genetic architecture, and methodology. Cell Genomics. 2023;3(10).
21. Everest E, Ahangari M, Uygunoglu U, Tutuncu M, Bulbul A, Saip S, et al. Investigating the role of common and rare variants in multiplex multiple sclerosis families reveals an increased burden of common risk variation. Sci Rep. 2022;12(1).
22. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet. 1993;52(3).
23. Sebro R, Rogus JJ. The power of the transmission disequilibrium test in the presence of population stratification. Eur J Hum Genet. 2010;18(9).
24. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. Am J Hum Genet. 2013;92(6).
25. Qi W, Allen AS, Li YJ. Family-based association tests for rare variants with censored traits. PLoS ONE. 2019;14(1).
26. Li D, Rotter JI, Guo X. A generalized least-squares framework for rare-variant analysis in family data. In: BMC Proceedings. 2014.
27. Schaid DJ, Mcdonnell SK, Sinnwell JP, Thibodeau SN. Multiple Genetic Variant Association Testing by Collapsing and Kernel Methods With Pedigree or Population Structured Data. Genet Epidemiol. 2013;37(5).
28. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43).
29. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1).
30. Andrews S. FastQC - A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Babraham Bioinforma. 2010;
31. Krueger F, James F, Ewels P, Afyounian E, Schuster-Boeckler B. FelixKrueger/TrimGalore. Babraham Bioinforma. 2021;
32. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14).
33. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Curr Protoc Bioinforma. 2013;(SUPL.43).
34. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al.

Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2017.
35. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10(2).
36. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17(1).
37. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16).
38. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med. 2020;12(1).
39. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: Predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1).
40. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;(SUPPL.76).
41. Li C, Zhi D, Wang K, Liu X. MetaRNN: differentiating rare pathogenic and rare benign missense SNVs and InDels using deep learning. Genome Med. 2022;14(1).
42. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016;99(4).
43. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22).
44. Pedersen BS, Quinlan AR. Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy. Am J Hum Genet. 2017;100(3).
45. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15).
46. Kawaguchi S, Higasa K, Shimizu M, Yamada R, Matsuda F. HLA-HD: An accurate HLA typing algorithm for next-generation sequencing data. Hum Mutat. 2017;38(7).
47. Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genomic mutational constraint map using variation in 76,156 human genomes. Nature. 2024;625(7993).
48. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44(1).
49. Kolberg L, Raudvere U, Kuzmin I, Adler P, Vilo J, Peterson H. G:Profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). Nucleic Acids Res. 2023;51(W1).
50. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021;49(D1).

51. Menegatti J, Schub D, Schäfer M, Grässer FA, Ruprecht K. HLA-DRB1*15:01 is a co-receptor for Epstein–Barr virus, linking genetic and environmental risk factors for multiple sclerosis. Eur J Immunol. 2021;51(9).
52. Batzdorf CS, Morr AS, Bertalan G, Sack I, Silva RV, Infante-Duarte C. Sexual Dimorphism in Extracellular Matrix Composition and Viscoelasticity of the Healthy and Inflamed Mouse Brain. Biology. 2022;11(2).
53. Halder SK, Sapkota A, Milner R. The importance of laminin at the blood-brain barrier. Neural Regen Res. 2023;18(12).
54. Sobel RA, Hinojoza JR, Maeda A, Chen M. Endothelial cell integrin laminin receptor expression in multiple sclerosis lesions. Am J Pathol. 1998;153(2).
55. Zapata-Acevedo JF, García-Pérez V, Cabezas-Pérez R, Losada-Barragán M, Vargas-Sánchez K, González-Reyes RE. Laminin as a Biomarker of Blood–Brain Barrier Disruption under Neuroinflammation: A Systematic Review. Int J Mol Sci. 2022;23(12).
56. Ichikawa-Tomikawa N, Ogawa J, Douet V, Xu Z, Kamikubo Y, Sakurai T, et al. Laminin α1 is essential for mouse cerebellar development. Matrix Biol. 2012;31(1).
57. Van Horssen J, Dijkstra CD, De Vries HE. The extracellular matrix in multiple sclerosis pathology. J Neurochem. 2007;103(4).
58. Halder SK, Sapkota A, Milner R. The impact of genetic manipulation of laminin and integrins at the blood–brain barrier. Fluids Barriers CNS. 2022;19(1).
59. Dalpé G, Leclerc N, Vallée A, Messer A, Mathieu M, De Repentigny Y, et al. Dystonin is essential for maintaining neuronal cytoskeleton organization. Mol Cell Neurosci. 1998;10(5–6).
60. Singh RB, Juneja L, Wilson DW, De Meester F, Wilczynska A, Takahashi T, et al. Diet and Lifestyle Guidelines for Immunomodulation with Reference to Corona Virus Pandemic: A Scientific Statement of the International College of Nutrition. World Heart J. 2021;13(4).
61. Juhl P, Domsic RT, Bay-Jensen AC, Karsdal M, Siebuhr AS, Franchimont N, et al. SAT0200 Serum Biomarkers of Collagen Turnover in Early and Late Diffuse Systemic Sclerosis. Ann Rheum Dis. 2016;75(Suppl 2).
62. Mohan H, Krumbholz M, Sharma R, Eisele S, Junker A, Sixt M, et al. Extracellular matrix in multiple sclerosis lesions: Fibrillar collagens, biglycan and decorin are upregulated and associated with infiltrating immune cells. Brain Pathol. 2010;20(5).
63. Gambardella J, Lombardi A, Morelli MB, Ferrara J, Santulli G. Inositol 1,4,5-trisphosphate receptors in human disease: A comprehensive update. J Clin Med. 2020;9(4).
64. Jarius S, Scharf M, Begemann N, Stöcker W, Probst C, Serysheva II, et al. Antibodies to the inositol 1,4,5-trisphosphate receptor type 1 (ITPR1) in cerebellar ataxia. J Neuroinflammation. 2014;11(1).
65. Bizet AA, Liu K, Tran-Khanh N, Saksena A, Vorstenbosch J, Finnson KW, et al. The TGF-β co-receptor, CD109, promotes internalization and degradation of TGF-β receptors. Biochim Biophys Acta - Mol Cell Res. 2011;1813(5).
66. Song G, Feng T, Zhao R, Lu Q, Diao Y, Guo Q, et al. CD109 regulates the inflammatory response and is required for the pathogenesis of rheumatoid arthritis. Ann Rheum Dis. 2019;78(12).

8 CURRICULUM VITAE