Fakultät für Medizin
Technische Universität München

Linking multiple sclerosis genetic risk variants to gene expression data

Dunja Kurtoić

Vollständiger Abdruck der von der Fakultät für Medizin der Technischen Universität München zur Erlangung des akademischen Grades einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Mikael Simons
Prüfende der Dissertation: 1. apl. Prof. Dr. Bertram Müller-Myhsok
2. Prof. Dr. Julien Gagneur

Die Dissertation wurde am 5. August 2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Medizin am 15. März 2022 angenommen.

Abstract

Genome-wide association studies have recently reported over 230 multiple sclerosis (MS) risk loci. In MS-related research, however, ways are still being sought to translate the genetic associations into functional, biological mechanisms. In this thesis, the task of studying MS etiology was tackled from two points of view: through the lens of animal models for MS and by examining the influence of MS-associated genetic variants on gene expression levels in individuals diagnosed with MS.

Firstly, the two currently used murine models for MS, namely the MOG-induced experimental autoimmune encephalomyelitis (EAE) model and the spontaneous opticospinal EAE (OSE) model were compared, and their relationship to human MS risk genes and T cell biology was examined. It was observed that the changes in gene expression in OSE mice were more prominent, with stronger signals from the adaptive immune system than when using the MOG-EAE model. In addition, the overrepresentation of human MS risk genes was more extensive among transcripts differentially expressed when the OSE model was used, especially in T_H1 cells. Therefore, when studying the functional role of MS risk genes and pathways during disease onset and their interaction with the environment, the spontaneous OSE might constitute a better model of human MS than the MOG-EAE.

Secondly, a workflow has been proposed with the aim to identify biological pathways mediating the effect of genetic variation in the early stage of the disease. The multi-level workflow enabled the comparison of groups of MS patients differing in their genetic background, with differential network analysis as the centerpiece of the analysis. By accounting for inter-individual variation due to clinical, demographic, and epidemiological factors, as well as for different cell type proportions present in the whole blood, it was possible to examine whether the variation left in the data was explainable by the genetic background. In the sample of KKNMS patients, it was observed that the exclusive effect of a single MS-associated variant had a comparatively low influence on gene expression levels. The gene expression variation in gene co-expression modules estimated from the immune-system-related genes was weakly explained by the genetic variants. In most comparisons, module structure was highly preserved between the groups. However, methods estimating conditional independence between genes were able to propose a potentially interesting result. The differential connectivity analysis of networks estimated based on Gaussian graphical models suggests the involvement of the rs6689470 genetic variant associated with MS risk in the actin-dependent cytoskeleton reorganization, a process important in B cell activation, in the early stages of MS. If this pathway were hampered in MS patients carrying the risk allele, the changed regulation of this biological unit could possibly lead to aberrant B cell activation resulting in evasion of self-reactive T cells into the periphery and potentially into the CNS. However, the involvement of rs6689470 variant in immune system regulation in the early stage of the disease needs to be further validated in an independent sample of MS patients.

Zusammenfassung

Vor kurzem sind in genomweiten Assoziationsstudien über 230 mit Multiple Sklerose (MS) assozierte Risikogenorte gefunden worden. Die MS-orientierte Forschung entwickelt, wie andere Gebiet der Forschung auch, weitere neuartige Methoden, damit wir die gefundenen genetischen Assoziationen im Kontext funktionaler biologischer Mechanismen erklären können. In dieser Arbeit wurde die Aufgabe, die MS-Ätiologie zu untersuchen, aus zwei Blickwinkeln angegangen: Als erstes wurden die Tiermodelle für MS erforscht und als zweites wurde der Einfluss der MS-assoziierten genetischen Varianten auf die Genexpression von Patienten mit MS Diagnose geprüft.

Zunächst wurden die beiden derzeit verwendeten Mausmodelle für MS, nämlich die MOG-induzierte experimentelle Enzephalomyelitis (EAE) und die spontane opticospinale EAE (OSE), verglichen und ihre Beziehung zu menschlichen MS-Risikogenen und der T-Zell-Biologie untersucht. Es wurde beobachtet, dass die Änderungen in Genexpression in OSE Mäuse prominenter waren, mit stärkeren Signalen des adaptiven Immunsystems als im MOG-EAE Model. Darüber hinaus war die Überrepräsentation von menschlichen MS-Risikogenen unter den in der OSE differenziell exprimierten Transkripten, insbesondere in T_H1-Zellen, größer. Daher könnte die spontane OSE bei der Untersuchung der funktionellen Rolle von MS-Risikogenen und -Wegen während des Krankheitsausbruchs und ihrer Interaktion mit der Umwelt ein besseres Modell der menschlichen MS darstellen als die MOG-EAE.

Zweitens wurde in dieser Dissertation ein Arbeitsablauf vorgeschlagen mit dem Ziel der Identifikation der biologischen Pathways, die die Wirkung der genetischen Variation auf die MS Suszeptibilität während des Frühstadiums der Krankheit vermitteln. Der mehrstufige Arbeitsablauf ermöglicht den Gruppenvergleich von MS Patienten mit unterschiedlichem genetischem Hintergrund. Durch die Berücksichtigung interindividueller Variationen aufgrund klinischer, demographischer und epidemiologischer Faktoren sowie unterschiedlicher Zelltyp-Anteile im Vollblut konnte untersucht werden, ob die in den Daten verbliebene Variation durch den genetischen Hintergrund erklärbar ist. In der Stichprobe der KKNMS-Patienten zeigen die Individuen robuste Genexpressionsniveaus, meist unabhängig von MS-assoziierten Varianten. Die Genexpressionsvariation in den aus den immunsystembezogenen Genen geschätzten Koexpressionsmodulen wurde nur in geringem Maße durch die genetischen Varianten erklärt, und in den meisten Vergleichen war die Modulstruktur zwischen den Gruppen in hohem Maße erhalten.

Allerdings konnten Methoden, die die bedingte Unabhängigkeit zwischen den Genen schätzen, ein möglicherweise interessantes Ergebnis identifizieren. Die differenzielle Konnektivitätsanalyse von Netzwerken, die auf der Basis von Gaußschen graphischen Modellen geschätzt wurden, legt die Beteiligung der genetischen Variante rs6689470 an der aktinabhängigen Reorganisation des Zytoskeletts nahe, einem Prozess, der für die Aktivierung der B-Zellen in den frühen Stadien der MS wichtig ist.

Wenn dieser Signalweg bei MS-Patienten, die das Risiko-Allel tragen, gestört ist, könnte die veränderte Regulation dieser biologischen Einheit zu B-Zell-Aktivierung aberranter Art führen, die zu einem Ausweichen von selbstreaktiven T-Zellen in die Peripherie und möglicherweise in das ZNS führt. Die Beteiligung der Variante rs6689470 an der

Regulation des Immunsystems im Frühstadium der Erkrankung muss allerdings in einer unabhängigen Stichprobe von MS-Patienten weiter validiert werden.

Acknowledgements

The process of working on my Ph.D. project and writing this thesis spans four and a half years of my life. During that time many things have significantly changed, both privately and professionally. Hereby, I would like to express gratitude towards my first supervisor, apl. Prof. Dr. med. Bertram Müller-Myhsok. Without him, pursuing my Ph.D. would not be such a stress-free yet successful experience. He taught me many things related to my job, but also taught me how to be more patient and to believe in what I can do. Thank you, Bertram, immensely, for your support. I would also like to say thank you to my second supervisor, Univ. Prof. Dr. Julien Gagneur, for providing me with a stress-free environment to learn and grow and for always being kind.

I would like to say thank you to my mentor, Dr. rer. nat. Till F. M. Andlauer, who was guiding me through my Ph.D. endeavor from the day one, literally. He helped me become the scientist I am today, to always do my best, to always question and seek for answers and to be thorough in my research. Time will show how much I learned from him. Together with Bertram, he taught me life-relevant skills and helped me open my mind towards diversity, fairness, and acceptance. I would also like to thank the members of our Statistical Genetics group, to all alumni and those who are still part of the group, for their support, for making me laugh, and for keeping me company in our everyday challenges.

I would like to thank my sister in law, who selflessly invested her time to thoroughly read my thesis and helped me with the grammar and the language.

Thank you, Martin. It took a lot of courage, love, faith, and trust to take that huge step (forward) and move to Munich with me, so that I could pursue my Ph.D. As it turned out, it was one of our best decisions. Thank you for being by my side, all this time.

Thank you, Sebastian. For making every single day one-hundred times better, for giving me so much courage. You made me grow in directions I didn’t know existed.

Lastly, but not any less importantly, I would like to express gratitude towards my parents and my sister, for their everlasting support and love.

Contents

List of Figures

List of Tables

Acronyms

AF	allele frequency
APC	antigen presenting cells
BF	Bayes factor
bp	base pair
C	MS-associated variant carriers
CI	confidence interval
CIS	clinically isolated syndrome
CNS	central nervous system
CSF	cerebrospinal fluid
DC	differential connectivity
DEA	differential expression analysis
DMT	disease modifying treatment
EAE	experimental encephalomyelitis
EBV	Epstein-Barr virus
eQTL	expression quantitative loci
FDR	false discovery rate
GGM	Gaussian graphical model
GWAS	genome wide association studies
HLA	human leukocyte antigen
HGNC	Human Genome Organization Gene Nomenclature Committee
IFN	interferon beta
IMSGC	International Multiple Sclerosis Genetics Consortium
kb	kilobase
KEGG	Kyoto Encyclopedia of Genes and Genomes
KKNMS	Krankheitsbezogene Kompetenznetz Multiple Sklerose
LD	linkage disequilibrium
log(FC)	logarithm of fold change
MA	minor allele
MAF	minor allele frequency
ME	module eigengene
MHC	major histocompatibility complex
ML	maximum likelihood
MOG	myelin oligodendrocyte glycoprotein
MS	multiple sclerosis
NC	MS-associated variant noncarriers
NK cells	natural killer cells
OR	odds ratio
OSE	opticospinal encephalomyelitis
PC	principal component
PCA	principal component analysis
PPMS	primary progressive multiple sclerosis
Q-Q	Quantile-Quantile
QC	quality control
RRMS	relapse remitting multiple sclerosis
SNP	single nucleotide polymorphism
SPMS	secondary progressive multiple sclerosis

T_H	T-helper cells
TOM	topological overlap measure
WGCNA	Weighted Gene Co-Expression Network Analysis

1 | Introduction

1.1 Introduction to multiple sclerosis

Multiple sclerosis (MS) is an autoimmune disease affecting the central nervous system (CNS) resulting in increased disability in affected individuals (Thompson *et al.*, 2018). MS is a complex disease where the coalescent effects of genetics, environment, and epigenetics confer the susceptibility to the disease. According to monozygotic twin studies, approximately 30% of MS risk can be attributed to genetics, thus making the disease partially heritable (Dyment, Dessa Sadnovich, & Ebers, 1997). MS follows the "common disease/common variant" model in which the common human disorders are caused by a set of common alleles, that is, alleles with a high allele frequency in the population (O'Gorman, Lin, Stankovich, & Broadley, 2013). This implies that many genetic variants have small contributions to the disease susceptibility, but their cumulative impact shapes the disease emergence, with the HLA-DRB1*15:01 allele constituting the strongest genetic risk for MS (Patsopoulos, 2018).

1.2 Pathology, disease phenotypes, and MS treatment

During MS development, the inflammation process results in lesions in the CNS, which are present in the white and grey matter, brain stem, spinal cord, and optic nerve (Reich, Lucchinetti, & Calabresi, 2018). Lesions can usually be detected on a magnetic resonance imaging (MRI) scan. The tissue is attacked by the cells of adaptive and innate immune system, directed to the myelin surrounding neurons, resulting in demyelination. Demyelination impairs the transduction of nerve signals, which, if affecting the optic nerve, causes problems with sight, which is often one of the first symptoms prior to developing MS (Ebers, 1985). Development of MS has different courses. In 85% of young adults it starts with a single episode, *i.e.*, clinically isolated syndrome (CIS) of the optic nerve, brain stem, or spinal cord (D. Miller, Barkhof, Montalban, Thompson, & Filippi, 2005). Three clinical subtypes of the disease have been defined: relapsing-remitting MS (RRMS), in which patients experience discrete repeated attacks followed by remissions, secondary progressive MS (SPMS), which is a secondary phase of RRMS in which deficits develop continuously without relapse, and the primary progressive MS (PPMS), in which the disease develops steadily from the onset. Clinical onsets of MS are distinguished based on the phenotype (tissue lesions, relapse, progression, atrophy) observed on the MRI scan (Thompson et al., 2018). Diagnosis is further supported by the cerebrospinal fluid (CSF) analysis in which the oligoclonal bands, unique antibodies produced in the CSF, probably by the B cells of the immune system (Reich et al., 2018), can be detected. The clinical course is important for selecting the optimal treatment. Effective therapies have been developed for RRMS, which affects 85-90% of MS patients, and partially effective for PPMS and SPMS. Due to important insights of B cell importance in MS development, B cell-aimed treatments have been developed. One example is ocrelizumab, a humanized monoclonal antibody directed at the CD20 molecule on the B cell surface, which has been effectively used to prevent relapses and (silent) progression in RRMS and PPMS patients. Ocrelizumab is one of the many disease modifying treatments (DMT), which reshape the course of MS by modulation of the immune function (Hauser & Cree, 2020).

1.3 Epidemiology and environmental risk factors

When looking at the gender distribution of MS patients in the population, one can observe that the majority of MS patients are women, with gender ratios in most of the European studies ranging from 1.1:1 to 3:1 (Kingwell et al., 2013). Average age of developing the disease is 30 years (Kimura, 2020). MS affects more than 2,3 million people worldwide (Thompson et al., 2018), with latitude positively correlated with MS risk (higher MS incidence further away from the equator). Distance from the equator indicates the potentially important role of the sun exposure. Ascherio and the colleagues (Ascherio et al., 2014) have suggested the importance of vitamin D supplementing in the early treatment of MS and other studies exploring the relationship between vitamin D-associated SNPs and risk for MS supported the importance of vitamin D on MS susceptibility (Mokry et al., 2015; Rhead et al., 2016). Epstein-Barr virus (EBV) infection is another important environmental factor associated with MS. This virus increases the probability of subsequently developing the disease threefold, in an age-dependent manner (Levin et al., 2005). It has been suggested that the EBV infection can mediate the autoreactive T cell response in the CNS in a twofold manner: the EBV-infected B cells can activate aberrant T cell response in the periphery, and stimulate the T cell autoreactivity in the CNS afterwards (Bar-Or et al., 2020). Other viral infections have also been suggested to interfere with mechanisms which normally limit T cell autoreactivity (Reich et al., 2018). Additionally, smoking and obesity interact with genetic factors as well, thereby increasing the MS risk (Hedström et al., 2014, 2011).

1.4 Hallmarks of immune response in MS patients

Immune cells reside in low concentration in the CNS of healthy individuals where they exert mostly surveilling functions. However, in patients with CNS inflammation, the concentration of immune cells is increased manyfold. In an MRI study, researchers reported finding of CD4$^+$ T cells, macrophages, B cells/plasma cells, and dendrocytes in the brain lesions of MS patients (Absinta, Sati, & Reich, 2016). The transport of activated lymphocytes from the periphery to the CNS parenchyma is facilitated by the more permeable blood-brain-barrier (BBB) observed in the MS patients (Chase Huizar, Raphael, & Forsthuber, 2020). Traditionally, T cells were considered the main drivers of human MS. The development of T cells starts in thymus, where these cells differentiate and go through positive and negative selection. First, CD4$^+$CD8$^+$ T cells recognizing the complex consisting of MHC and the peptide are selected in the thymus cortex. This process is called positive selection. In the process, cells correspondingly differentiate into CD4- and CD8-single positive T cells. Next, in the medulla of the thymus, these cells interact with antigen presenting cells (APCs), *e.g.*, the medullary thymic epithelial cells or the dendritic cells, whereby the most of the T cells recognizing the self-peptide-MHC complexes, *i.e.*, the autoreactive T cells, are removed. Some of the autoreactive T cells are additionally going through the agonist selection. In the agonist selection, the autoreactive CD4 T cells differentiate into regulatory T cells (T_{reg}), expressing the Foxp3. The role of such Foxp3$^+$ T_{reg} cells is to control the peripheral immune tolerance (Takaba & Takayanagi, 2017). It has recently been shown that B cells residing in thymus also contribute to T cell selection. It is therefore possible that they influence a T cell repertoire which will react with autoantibodies of the CNS (Jelcic et al., 2018).

Initially, CD4$^+$ T cells were considered responsible for MS emergence, because of the MHC class II restriction and because of their role in experimental autoimmune encephalomyelitis (EAE) induction, a commonly used animal model for MS. CD4$^+$ T cells are activated after the contact with MHC class II molecules on APCs. After the activation, they can differentiate into T$_H$1 and T$_H$2 cells, but also into T$_H$17 cells (Seder & Ahmed, 2003). Furthermore, the T$_H$17 cells have also been found in the blood and the brains of patients with MS. T$_H$1 and T$_H$17 cells have been shown to be involved in different autoimmune diseases, *e.g.*, MS, and may drive different immunopathologies (Damsker, Hansen, & Caspi, 2010). However, the depletion of CD4$^+$ T cells in MS patients did not improve the relapse rates in MS patients (Van Oosten et al., 1997). On the other hand, the depletion of both CD4$^+$ T cells and the CD8$^+$ T cells, together with an anti-CD52 monoclonal antibody led to successful reduction in relapses and new lesions (Paolillo et al., 1999). Thereby, the importance of CD8$^+$ T cells in MS has emerged. Contrary to CD4$^+$ T cells, CD8$^+$ T cells mostly recognize antigens presented by MHC class I molecules (Wong & Pamer, 2003). The implications of T cells in MS pathology have further been supported by GWAS, where for example *IL7R* and *IL2RA* genes, important in T cell differentiation and their expansion and apoptosis, were found to be associated with MS risk (J.A. Hollenbach & Oksenberg, 2016; Maier, Lowe, Cooper, Downes, & Anderson, 2009).

Both the adaptive and innate arms of the immune system play important roles in disease development. The adaptive immune system response includes T cells and B cells. In MS patients, defects in peripheral regulatory immune cell populations (*e.g.*, Foxp3$^+$ T$_{reg}$) promote differentiation of naïve T cells to pathogenic, creating autoreactive T cells. Increased frequency of interferon (IFN)-γ-secreting T$_{reg}$ cells has also been observed in comparison to healthy individuals (Axisa & Hafler, 2016), leading to a lower suppression of immune response. Therefore, the defense against autoimmune response in MS patients is weakened, and autoreactive T cells can proliferate, extravasate, and attack the myelin sheath in the CNS. The proliferation of the autoreactive T cells is further supported by the B cell activity. B cells express the receptors of the major histocompatibility complex (MHC) on their surface which present the processed myelin to CD8$^+$ and CD4$^+$ T cells thereby activating them. In addition, co-stimulatory molecules on the surface of B cells help promote the activation of pro-inflammatory T cells (Sabatino, Pröbstel, & Zamvil, 2019). The secretion of pro-inflammatory cytokines is another B cell function facilitating the expansion of the immune response. The mechanism of ocrelizumab is based on depleting the B cells expressing the CD20 molecule. As a consequence of this specific depletion, the supply of B cells from the periphery to the CNS is interrupted, B and T cell interaction is reduced as well as the secretion of the pro-inflammatory cytokines (Hauser & Cree, 2020). Due to ample important functions of B cells which have recently emerged, MS is no longer viewed as a primarily T-cell driven disease.

Furthermore, the cells of the innate immune system response, like macrophages and microglia also play important roles in supporting the autoimmune response. Macrophages secrete cytokines promoting the inflammatory response of T cells and B cells thereby mediating the destruction of the myelin sheath which surrounds axons. The role of natural killer (NK) cells and dendritic cells has also emerged recently (IMSGC, 2019b). Microglia, cells populating the CNS, are involved in myelin phagocytosis, in T cell antigen presentation and release of proinflammatory cytokines in active CNS lesions (Lassmann, Van Horssen, & Mahad, 2012).

1.5 Animal models for MS

Studying a complex disease has been supported by the creation of apt animal modes, such as mouse models. The first attempt to induce an inflammation process similar to the one happening in human MS, was described as early as 1933. Rivers and the colleagues (Rivers, Sprunt and Berry, 1933) injected monkeys with brain matter and observed brain inflammation of similar patterns as in active MS lesions. This paper marked the start of experimental autoimmune encephalomyelitis (EAE), after which researchers created various versions of EAE in different animal species. However, MS has variable clinical and pathological characteristics, the course of the disease in a patient is subject to change with disease progression. Neither one of the tested EAE models managed to recapitulate complex disease profiles observed in humans. Nonetheless, the importance of EAE is unquestionable in terms of studying cellular and molecular pathways, because many of the studied pathways were later found to be relevant for human MS as well. This also enables to test the potential medical treatments for MS, even though EAE cannot fully represent human MS in this aspect (Ben-nun et al., 2014). Many disease modifying treatments (DMTs) for MS were identified and all DMTs have been tested in animals suffering from EAE, such as IFN-β formulations, glatimer acetate (GA), natalizumab, ocrelizumab, and others (Glatigny & Bettelli, 2018). Very recently, a process of neddylation, analogous to ubiquitination, has been studied in mice suffering from EAE. Researchers observed that the inhibition of neddylation led to decreased EAE severity and suggested neddylation as a new therapeutic target (Kim et al., 2021).

1.5.1 Induced and spontaneous EAE models in mice

With the emerging role of genetics in MS susceptibility, it is important to reassess the two currently used EAE models, the MOG-induced EAE, in which the disease is triggered by active induction, and the spontaneous EAE model, *e.g.,* the opticospinal EAE (OSE). In actively induced EAE, the disease emerges after injecting myelin-derived antigens, such as myelin oligodendrocyte glycoprotein (MOG). The disease develops rapidly, contrary to the human MS (Krishnamoorthy, Holz, & Wekerle, 2007) . On the other hand, spontaneously induced EAE, *e.g.*, the OSE, develops in transgenic mice expressing a T cell receptor (TCR) which recognizes the MOG peptide. These mice also carry B cells with MOG-specific receptors (Glatigny & Bettelli, 2018). The double transgenic nature therefore allows the B and T cell interaction on several levels to produce the disease (Krishnamoorthy, Lassmann, Wekerle, & Holz, 2006). The OSE mice develop a disorder similar to neuromyelitis optica, a variant of MS where lesions are present in the optic nerve and the spinal cord only, omitting the cerebrum and cerebellum (Jarius et al., 2020; O'Riordan et al., 1996).

1.5.2 Roles of T_H1 and T_H17 cells in EAE development

It has been well known that myelin-reactive T cells with T_H1 phenotype can induce EAE, as well as to support the inflammatory response by secreting the IFN-γ, a cytokine which activates macrophages (Merrill et al., 1992). However, the role of T_H1 cells remains controversial, because some studies suggest that IFN-γ and T_H1 are crucial for EAE development, while other research showed that mice not expressing the IFN-γ can become more susceptible to EAE (Ferber et al., 1996), therefore stating that the IFN-γ is not crucial for EAE induction. Naïve T cells stimulated with tumor growth factor β (TGF- β) and IL-

6 will differentiate into another type of T cells, the T_H17 cells, which have also demonstrated their pathogenicity, that is, the potency to induce the EAE (Grifka-Walk, Lalor, & Segal, 2013). Both T_H1 and T_H17 can induce EAE by transfer, but the mechanisms of disease induction differ due to different spectrum of produced cytokines (Jäger, Dardalhon, Sobel, Bettelli, & Kuchroo, 2009). These knowledges together suggested that the heterogeneity of lesions present in patients with MS could be a consequence of different autoreactive T cell subtypes. Due to the important roles of T cells in MS emergence (described in the previous paragraphs), some of which were also suggested by GWAS (IMSGC, 2013, 2019b), it would be important to decipher which of the EAE models can better capture the role of T_H1 and T_H17 cells.

1.6 Genetic influence on MS susceptibility

Genome wide associations studies (GWAS) have revealed over 230 genome-wide loci associated with MS risk (IMSGC, 2019a). Most of the loci map to the highly polymorphic MHC region, coding for human leukocyte antigen (HLA) genes. There are two major classes of HLA genes, the class I and class II. Both classes code for the molecules on the cell surface, included in adaptive immune response. These molecules participate in antigen internalization, processing, and presentation of the peptides to T cells (Jill A. Hollenbach & Oksenberg, 2015). The classical class I HLA molecules (HLA-A, HLA-B, and HLA-C) are found on all cells with nucleus and they present peptides to the $CD8^+$ T cells. These peptides are mostly derived from endogenous proteins (*e.g.,* viral peptides). On the other hand, the classical class II HLA molecules, HLA-DR, HLA-DQ, and HLA-DP are found on APCs like B cells, dendritic cells, and macrophages and they present peptides to $CD4^+$ T cells (Jill A. Hollenbach & Oksenberg, 2015). The non-classical class II genes, the *DM* and *DO*, are involved in the peptide binding groove editing, influencing the binding and the release of the peptides (Welsh & Sadegh-Nasseri, 2020). Wucherpfennig and Sethi (Wucherpfennig & Sethi, 2011) suggested that the high level of polymorphisms in MHC region is one of the main factors linking the HLA genes to human diseases. The polymorphisms affecting the structure of the peptide binding groove might play the key role in deciding which self-peptides will be presented to T cells, therefore either promoting or preventing the autoimmune response (Wucherpfennig & Sethi, 2011).

The DRB1*15:01 allele of the class II *DRB1* gene constitutes the strongest risk for MS, with odds ratio (OR) of 7 or more in homozygous high risk carriers, and between 3.5 and 5 in heterozygotes (Baranzini & Oksenberg, 2017). The association between the DRB1*15:01 allele and expression of *DRB1* gene has been shown before (Alcina et al., 2012). Alcina and the colleagues (Alcina et al., 2012) suggest that the higher expression of class II HLA genes could contribute to higher number of HLA heterodimers exposed on the cell surface. As a consequence, this could promote stronger activation signals for T cells, contributing to the inflammatory response. Furthermore, a study comparing expression levels of genes in the brain tissue between the groups of MS patients carrying the DRB1*15:01 allele and those not carrying the allele showed that the variant influenced the expression of nine genes, including the *DRB1* and *IL18R1,* the interleukin receptor (Enz et al., 2020).

The first genetic variants discovered outside of the MHC region were found in genes *IL2RA* and *IL7RA* (IMSGC, 2007), both of which code for receptor subunits found on immune cells, *e.g.*, on T cells. Many polymorphisms in genes with immunological functions like

CXCR5 (C-X-C motif chemokine receptor 5; IMSGC, 2007) or *TNFRSF1A* (tumor necrosis factor receptor) were also found to be associated with MS (IMSGC, 2007). A large proportion of MS-associated variants is located in the noncoding regions of the genome, some even distant from any gene (Baranzini & Oksenberg, 2017), thus making it harder to explain the background of their association with MS.

It has been several decades since the first HLA genes associated with MS have been found, but their role in disease pathology still remains ambiguous, even though studies with many thousands of individuals have been performed (Baranzini & Oksenberg, 2017). One way to functionally annotate GWAS findings is to examine the effect of variants on expression levels of gene in the proximity of the variant, that is, to find the expression quantitative trait loci (eQTL). But, such SNP-gene associations never act alone, they are a part of the bigger system and probably have further downstream effects. Several gene ontology and network analyses of MS associated variants were performed revealing an overrepresentation of immune-cell associated genes, and in particular those which are T cell-associated (IMSGC, 2013, 2019a; IMSGC & WTCCC, 2011; Patsopoulos & De Bakker, 2011), thereby showing the potential of network based approaches in studying a complex disease like MS as well as underlining the complexity of the immune response producing the disease.

2 | Gene Expression in Spontaneous Experimental Autoimmune Encephalomyelitis is Linked to Human Multiple Sclerosis Risk Genes

This chapter describes the work published in the journal Frontiers in Immunology, section Multiple Sclerosis and Neuroimmunology, in September 2020 (Faber et al., 2020). I am the shared first author of the publication.

2.1 Research questions

The importance of genetic component in human MS has become prominent in the past 20 years, which calls for a reassessment of the currently used mouse models for studying human MS. Can we find significant enrichments of human MS risk genes in genes specific for either actively induced or spontaneously developing EAE? Which of the two models, the spontaneous opticospinal EAE or the MOG-induced EAE models human MS more faithfully? What is the scope of gene expression differences in T_H1 and T_H17 cell specific transcripts in both EAE models?

2.2 Motivation

Mouse models are commonly used to study the pathophysiology of human diseases, including MS. Both actively and spontaneously induced EAE models are valuable animal representatives of human MS, yet translation of EAE research to mechanisms of MS in humans remains controversial. At the moment, there are no experimental models covering the complete spectrum of clinical, pathological, and immune characteristics of MS. In the context of this thesis, the gene expression of EAE mice for which EAE was actively induced by injecting the $MOG_{35\text{-}55}$ peptide was compared to double transgenic mice developing the EAE spontaneously, in order to examine the potential discrepancies. Comparing the role of T_H cells in both models as well as finding the extent of the overlap between human MS risk genes and the differentially expressed genes could indicate which of the two models resembles human MS better.

2.3 Background

Animal models enable the research of human autoimmune disorders and are of paramount importance especially when the human tissue is not available or hard to extract, like, for example, brain tissue in MS research. Yet, it is still not clear whether mouse models for human diseases can adequately mirror complex disorders like the human MS, for which both the environmental and the genetic component contribute to disease susceptibility. The mouse models should ideally develop the disorder spontaneously and correctly represent at least part of the disease characteristics (Krishnamoorthy et al., 2007). In this project, we wanted to compare the MOG-induced EAE to opticospinal EAE (OSE). Differential expression analysis and pathway overrepresentation analysis were employed with the aim of investigating the following two major aspects: First, to understand the relationship between human MS and EAE models better, it was examined to which extent two currently used murine models resemble the immune system processes important in human MS. Second, the differential expression of T_H-cell specific transcripts was investigated in actively induced EAE and OSE. Finally, the scope of overrepresentation of human MS risk genes in both models was explored *via* enrichment tests and permutation analysis.

2.3.1 Exploring the scope of gene expression variability in the data

A common way to explore differences in gene expression between groups of interest is to apply linear models for microarray data (limma). This approach is implemented in the *limma* package in *R* (Ritchie et al., 2015). The method enables the analysis of gene expression data as a whole, not just by pairwise gene comparisons. Such methodology has been proven advantageous especially in a small sample data setting because it facilitates borrowing information across genes, therefore resulting in more stable estimates of variance (Smyth 2004).

Following the paper from Smyth (Smyth 2004), the gene expression data is represented as a numerical matrix where rows of the matrix are individual transcripts' gene expression levels and columns are the samples. To be able to fit the linear model, a design matrix D is created, representing the distribution of different transcripts across the groups. Next, the matrix of contrasts C provides information on which groups we want to compare. We assume that

$$E[y_g] = D\alpha_g \qquad (2.1)$$

where y_g contains gene expression data for the gene *g,* D is the design matrix and α_g is the vector of the coefficients. By fitting the model to the response for each gene, coefficient estimators $\tilde{\alpha}_g$ and an estimator s_g^2 of the standard deviation σ_g^2 are calculated as

$$var(\tilde{\alpha}_g) = V_g s_g^2 \qquad (2.2)$$

where V_g is the unscaled covariance matrix not depending on s_g^2. From the linear model and the contrast matrix C, we obtain the β_g, that is, the contrast estimator

$$\tilde{\beta}_g = \mathrm{C}^T \tilde{\alpha}_g \tag{2.3}$$

with estimated covariance matrices

$$var(\tilde{\beta}_g) = C^T V_g C s_g^2 \ . \tag{2.4}$$

Next, we test which of the contrast estimators are different than zero, that is, whether the expression of the gene g is significantly different for the given contrast. To assess the difference in gene expression, the empirical Bayes approach is applied. The empirical Bayes approach enables borrowing information across genes, which results in more precise estimates of gene expression variance. Gene expression variance is moderated in the following way. First, the average variance of all genes on the array is calculated and it constitutes the prior variance. Next, expression variance is calculated for each gene, *i.e.*, the posterior variance. The posterior variance of each gene is then shrunken towards the average value to increase the variance estimation precision, yielding the $\tilde{s}_g$. The moderated *t* statistic is calculated by using the shrunken variance values:

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{v_{gj}}} \tag{2.5}$$

where $\hat{\beta}_{gj}$ is the estimated contrast coefficient for gene g and v_{gj} is *j*th diagonal element of $C^T V_g C$. In other words, $\sqrt{v_{gj}}$ is the unscaled standard deviation of the *j*th contrast for gene g. Moderated *t* statistics follows *t*-distribution on $f_0 + f_j$ degrees of freedom, where extra degrees of freedom f_0 are added due to the extra information borrowed from the whole gene set for inference about each individual gene. Therefore, the corresponding *p* values can be obtained as well, which enables ranking genes according to the scope of their differential gene expression.

2.3.2 Overrepresentation analysis

Over-representation analysis (ORA) is the first-generation approach to the functional analysis of microarray gene expression data. The method enables a statistical evaluation of gene set enrichment in a given pathway. Statistical evaluation is employed *via* the hypergeometric test (or Fisher's exact test), where the degree of enrichment is calculated as a probability indicating that certain genes in the gene set are detected more often than expected by chance. The probability is given by

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{i}} \tag{2.6}$$

where N is the gene universe (genes in the background distribution), M is the number of genes within that distribution which are annotated to the GO node of interest, *n* is the size of our list of interest and *k* is the number of genes in that list annotated to the node (Boyle et al., 2004).

2.4 Materials and methods

This section presents and describes the Materials and Methods relating to and published in the research paper by Faber, Kurtoic, and the colleagues (Faber et al., 2020).

2.4.1 Materials

2.4.1.1 Mice

The study used mice with the *C57BL/6* background which were bred in the animal facilities of the Max Planck Institute of Biochemistry and Neurobiology, Martinsried, Germany. These mice lineages were either genetically manipulated, producing the double-transgenic 2D2 (TCR_{MOG}) × IgH_{MOG} OSE mice, or immunized subcutaneously with MOG peptide consisting of 35-55 amino acids. For the analysis of EAE models, only female mice were used. The standard 5-point scale (Krishnamoorthy et al., 2006; S. D. Miller, Karpus, & Davidson, 2010) was used to score the clinical signs of EAE in mice: 0: healthy animal; 1: animal with a flaccid tail; 2: animal with impaired righting reflex and/or gait; 3: animal with one paralyzed hind leg; 4: animal with both hind legs paralyzed; 5: moribund animal or death of the animal after preceding clinical disease. Following the ethically approved protocol from the animal welfare committee of the government of Upper Bavaria, animals were sacrificed when they reached the score 4.

2.4.1.2 T cell differentiation

The spleen from four OSE mice (mixed gender) was used to extract the T cells. T cells were then polarized *in vitro* as described in the research paper by Domingues and the colleagues (Domingues et al., 2010). In summary, four batches of four mice were used to isolate erythrocyte-lysed spleen cells. To generate the T_H1 cells, cells were cultured in the presence of a MOG peptide (amino acids 1-125), IL-12, IL-1, and anti-IL-4. Additionally, the IL-2 was added after three days. Cells were cultured in the presence of a MOG peptide (amino acids 1-125), TGF-β1, IL-6, IL-23, anti-IL-4, and anti-IFN-γ. The IL-23 was added to the culture after three days, generating the T_H17 cells. Both types of cells were re-stimulated after six days and harvested after nine and 12 days. Naïve T_H0 cells were harvested on day 0. Flow cytometry, ELISA, and quantitative real-time PCR were used to determine the success of polarization.

2.4.1.3 Microarrays

The RNA from the spinal cord of healthy and diseased EAE mice was analyzed on the Sentrix BeadChip ArrayMouseWG-6 v2 (Illumina, San Diego, USA). Four chips (24 samples, four per experimental group) were hybridized. The RNA from the T_H cells was analyzed on the Sentrix BeadChip Array MouseWG-6 v1.1 (Illumina, San Diego, USA). Altogether, three chips were used (18 samples from four separate experiments: 4 × T_H0, 7 × T_H1, 7 × T_H17). Samples and chips from both experiments, that is, the spinal cord and the spleen, were processed in parallel. All microarrays fulfilled Illumina's recommendations for quality control (QC). Further QC was done in *R* v3.2.2 (R Core Team, 2020).

Bead summary gene expression data for the EAE models analysis was loaded using the *beadarray* package (Dunning, Smith, Ritchie, & Tavare, 2007), normalized and transformed with the help of *lumi* (Du, Kibbe, & Lin, 2008) and *vsn* packages (Huber, Von Heydebreck, Sültmann, Poustka, & Vingron, 2002). Probes with a detection p value > 0.05 in >10% of the samples were removed. The *illuminaMousev2.db* package enabled filtering out probes based on their annotated quality. Probes annotated as “no match” or “bad” were removed. After the QC, the dataset consisted of 21,483 transcripts. For the T_H cells analysis, the *limma* package (Ritchie et al., 2015) was used to load the summary data. The data was processed as described previously, with the *illuminaMousev1p1.db* package used to filter out probes based on their annotated quality. After the procedure, 17,858 transcripts remained.

2.4.2 Methods

2.4.2.1 *k-means clustering*

k-means is a method for unsupervised clustering, for which the number of clusters needs to be specified before the run. Cluster centers are assigned randomly in the first run, and in the following runs, the data point is clustered with its nearest mean, thereby minimizing the within-cluster variance (Macqueen, 1967). The *kmeans* function in R was employed to capture the variance of four clusters in the data, representing the healthy mice (OSE$_0$, CFA, and WT), the mice with the OSE score 1 (OSE$_1$), OSE score 4 (OSE$_4$), and MOG 4 (MOG$_4$). The clustering was run for one-hundred independent times, and for each run, cluster centers were assigned randomly.

2.4.2.2 *Differential gene expression*

To explore the gene expression differences between spontaneous and induced EAE, six mouse types (with four mice each) were examined: the wildtype (WT); healthy OSE controls (OSE$_0$); OSE mice with disease score 1 (OSE$_1$); OSE mice with disease score 4 (OSE$_4$). These are followed by the MOG EAE control mice, which are healthy mice injected with complete Freund's adjuvant but not with a MOG peptide (CFA); and lastly, the MOG$_{35\text{-}55}$ EAE mice with disease score 4 (MOG$_4$), which are the *C57BL/6* wildtype mice injected with adjuvant and the MOG$_{35\text{-}55}$ peptide. The six mouse types were used to build a design matrix for the differential expression analysis. Four chips were used to measure the gene expression, and each chip contained one sample per mouse type. To account for the random effects due to sample positioning on the chip, chip labels were added to the linear model *via* the *duplicateCorrelation* function from the *limma* package. By fitting the linear models and applying the moderated *t*-tests, the following five contrasts were compared: MOG$_4$-CFA, CFA-WT, OSE$_4$-OSE$_0$, OSE$_1$-OSE$_0$, and OSE$_4$-WT. For the T$_H$ cell analysis, the data from day 9 of the three cell types was used to create a design matrix: naïve T$_H$0, T$_H$1 and T$_H$17. Four mouse pools were included as random effects in the model. The difference in gene expression was analyzed in two contrasts: T$_H$1 *vs.* T$_H$0 and T$_H$17 *vs.* T$_H$0.

2.4.2.3 *Overrepresentation analysis*

ORA was applied on differentially expressed sets of genes when comparing EAE models and genes differentially expressed in the T$_H$ cell comparison. The analyses were conducted using the *WebGestalt* v2019 package in *R* (Y. Liao, Wang, Jaehnig, Shi, & Zhang, 2019), with *genome protein-coding* genes as a background. For this purpose, DE transcripts were uniquely mapped to Entrez IDs. The hypergeometric test was used to determine the significance levels, and Benjamini-Hochberg FDR was used to account for multiple testing (Benjamini & Hochberg, 1995).

2.4.2.4 *Enrichment tests – permutation analysis*

In order to explore in which sets of differentially expressed genes of EAE models we find an enrichment of, *e.g.,* MS susceptibility genes or T$_H$-specific genes, enrichment tests in *R* were implemented as described hereafter. For each tested group, differentially expressed transcripts were mapped to unique set of Entrez IDs. In the first step, the number of differentially expressed genes was determined, and the random set of genes of the same size was selected. In the second step, the test set was selected, which comprised either MS susceptibility genes or T_H-specific genes. Then the overlap between the genes in the random set and the test set was determined. Random gene set shuffling was repeated 100,000 times. Enrichment *p*-value was calculated by counting the number of times the overlap between the random set of genes and the test set was equal to or bigger than the overlap between differentially expressed genes and test set, and dividing by the number of permutations. When testing for the enrichment of human MS susceptibility genes, the most recent list of such genes was used. IMSGC Consortium (IMSGC, 2019b) published a list of 558 MS susceptibility genes outside of the MHC region (Supplementary Table 18 of the cited paper). Genes *CTB-50L17.10*, *RP11-345J4.5*, *JAZF1-AS1*, *ZEB1-AS1*, *GATA3- AS1*, *SSTR5-AS1*, and *RPL34-AS1* were excluded from the list, leaving 551 putative MS susceptibility genes described in the IMSGC publication.

2.5 Results

This section presents and describes the Results published in the research paper by Faber, Kurtoic, and the colleagues (Faber et al., 2020).

2.5.1 Disease-relevant gene expression changes successfully identified by k-means clustering

Gene expression profiles of the total spine cord samples from two EAE models, OSE and MOG_{35-55} EAE, were compared. In OSE mice, the autoimmune response in the CNS manifested itself predominantly in the lumbar part of the spinal cord. MOG_{35-55} peptide was injected in *C57BL/6* wildtype animals, thereby actively inducing the EAE. The gene expression variation patterns between the healthy (OSE_0, CFA, and WT) and the diseased (OSE score 1 (OSE_1), OSE score 4 (OSE_4), and MOG 4 (MOG_4)) animals were examined by analyzing the first two principal components (PCs) of gene expression data. The two PCs showed clear separation between healthy and diseased animals, with healthy mice clustering together (Figure 2.1 **A**). More variation in gene expression levels was observed in OSE_4 mice than in MOG_4 mice. This reflects the spontaneous nature of the OSE development, while the course of EAE in MOG-induced animals is usually more stereotypic (Krishnamoorthy et al., 2006). To further support these findings, the unsupervised k-means clustering was run with one-hundred different random starting centers. In 97 out of 100 replications, healthy and diseased animals were put into different clusters (Table 2.1, Figure 2.1 **A**). The first two PCs were employed to visualize the cluster distribution. In the most frequent cluster solution (34/100), all healthy mice were clustered in cluster 1, OSE_1 animals were grouped in cluster 2, all OSE_4 animals in cluster 3, while cluster 4 comprised of the rest of the animals.

Table 2.1 ***Frequencies of six different cluster solutions after 100 k-means runs and whether healthy mice were clustered apart from the diseased mice.***
Figure source: the supplement of the research paper by Faber, Kurtoic, and the colleagues (Faber et al., 2020)

Clustering solution	Frequency (out of 100 runs)	Healthy and diseased animals clustered apart
Solution 1	34	TRUE
Solution 2	30	TRUE
Solution 3	22	TRUE
Solution 4	10	TRUE
Solution 5	3	FALSE
Solution 6	1	TRUE

2.5.2 OSE mice exert more prominent gene expression changes

To examine gene expression differences between OSE and MOG EAE mice, five contrasts were introduced: OSE_1-OSE_0, OSE_4-OSE_0, MOG_4-CFA and the two control contrast CFA-WT and OSE_0-WT (Figure 2.1 **B**). There were no differentially expressed transcripts in the control contrast CFA-WT. Transcript mapping to the *T cell receptor alpha chain* (*Tcra*) was the only transcript differentially expressed between the OSE_0 and WT mice, and was also upregulated in all the other contrasts, except for the contrast CFA-WT. There were more significantly differentially expressed transcripts between OSE_4 and OSE_0 mice (*n* = 5,555) than between MOG_4 and CFA (*n* = 3,182). The expression of altogether 864 transcripts differed significantly between MOG_4 and OSE_4 animals (Figure 2.1 **B**). Furthermore, in OSE_4-OSE_0 contrast only, there were 4.88 × more transcripts differentially expressed when compared to the MOG_4 -CFA contrast (Figure 2.1 **B**). More prominent global gene expression changes were observed within OSE_4-OSE_0 contrast than in OSE_1-OSE_0 (binomial test: *p* value = 1.4×10^{-65} for all transcripts, *p* value = 9.9×10^{-119} for transcripts differentially expressed in both contrasts, Figure 2.1 **C**) or MOG_4-CFA (*p* value = 5.8×10^{-3} for all, *p* value = 2.7×10^{-221} for differentially expressed transcripts, Figure 2.1 **D**).

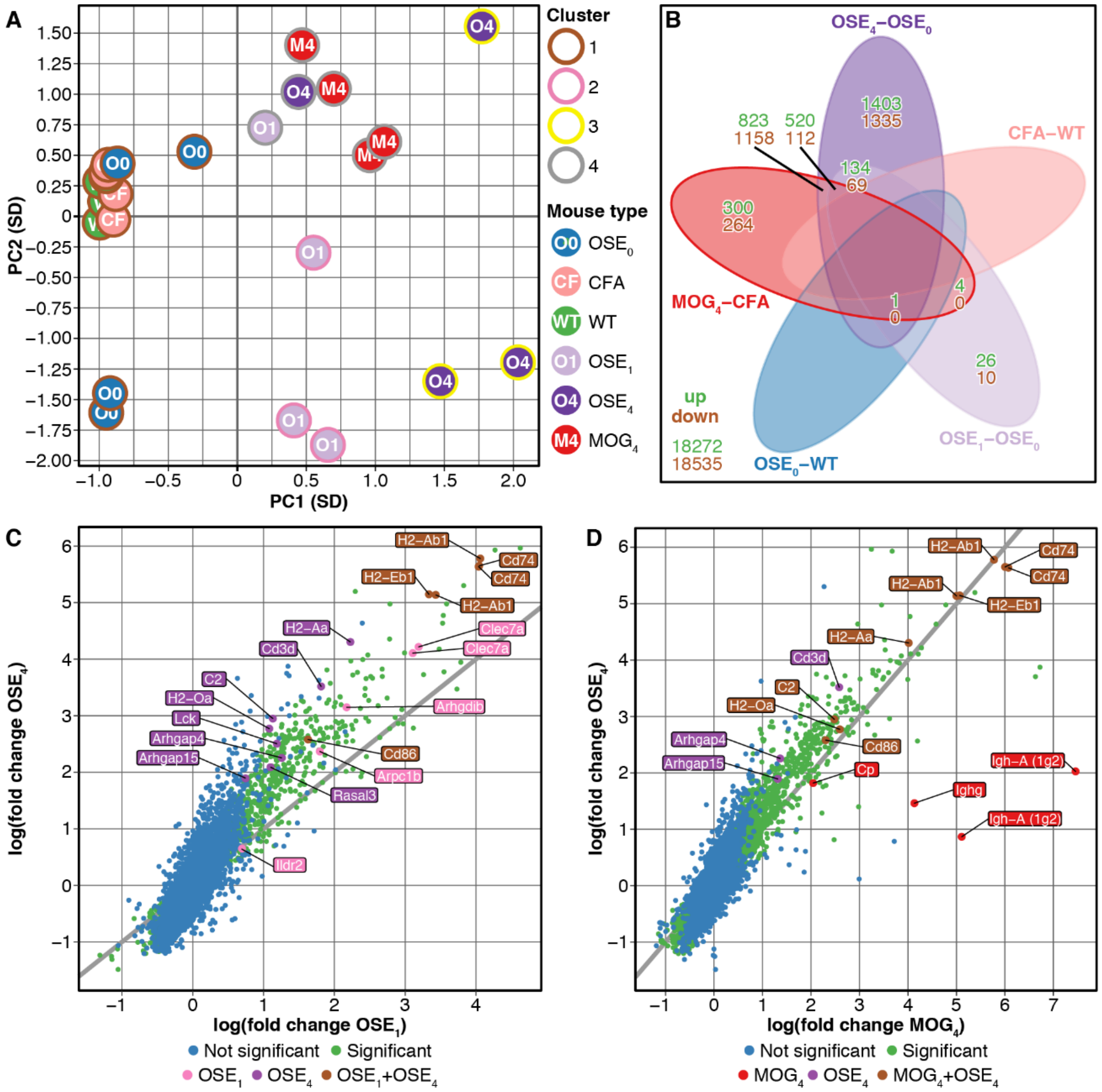

Figure 2.1 ***Differential expression analysis of OSE and MOG EAE animal models****.*
Figure from the research paper by Faber, Kurtoic, and the colleagues (Faber et al., 2020)**.**
(A) *The first two principal components (PCs) of gene expression data grouped by k-means clustering (the most frequent cluster solution, 34/100). In 97 out of 100 runs, all WT, OSE$_0$, and CFA mice were grouped in a separate cluster from the diseased animals. PC – principal component; SD – standard deviations.* ***(B)*** *Venn diagram visualizing the number of differentially expressed transcripts in five analyzed contrasts.* ***(C, D)*** *OSE$_4$ mice showed stronger fold changes of gene expression levels than* ***(C)*** *OSE$_1$ and* ***(D)*** *MOG$_4$ mice, each compared to its control (OSE$_0$ and CFA, respectively). On the plots, for each group, the top 10 differentially expressed genes with Entrez IDs are labelled. For some genes, there were two probes hitting the gene and both were present in the top 10 differentially expressed transcripts. In such cases, both of the probes were plotted, but the gene name itself was counted once. Group-color legend: differentially expressed in OSE$_1$ only (light magenta), differentially expressed in OSE$_4$ only (dark magenta), differentially expressed in MOG$_4$ only (red), differentially expressed in* ***(C)*** *both OSE$_1$ and OSE$_4$ or* ***(D)*** *both MOG$_4$ and OSE$_4$ (brown; with higher expression levels observed for OSE$_4$).*

2.5.3 OSE-related genes especially enriched in immune system processes

In order to further describe the expression changes in the two EAE models, ORA was applied on the differentially expressed transcripts (Supplementary Table 3 in the paper by Faber, Kurtoic, and the colleagues (Faber et al., 2020)) divided into groups, which was presented in the Table 2.2 (Figure 2.2).

Table 2.2 ***Description and abbreviations of groups representing differentially expressed transcripts.***

Group description	Abbreviation
transcripts differentially expressed for both contrasts OSE_4-OSE_0 and MOG_4-CFA, but not in the two control contrasts OSE_0-WT or CFA-WT.	*common disease transcripts* (CDT)
differentially expressed for the OSE_4-OSE_0 contrast, but not for the MOG_4-CFA or the control contrasts	*OSE_4-specific transcripts* (OSE_4sp)
differentially expressed in OSE_1-OSE_0 but not in the control contrasts	OSE_1-expressed transcripts (OSE_1ex)
differentially expressed for MOG_4-CFA, but not for OSE_4-OSE_0 or the control contrasts	*MOG_4-specific transcripts* (MOG_4sp)

For the CDT set, 1,379 redundant GO biological processes remained significant after adjusting for multiple testing. Gene sets *immune response, regulation of immune system process,* and *T cell activation* were among top associated terms (adjusted *p* value $< 2 \times 10^{-16}$, Figure 2.2). These processes were also significant in the OSE_4sp (adjusted *p* value $\leq 3.5 \times 10^{-2}$), together with other immune system related terms. On the other hand, no immune system related process was significant for MOG_4sp. Hence, the gene expression changes of the immune system were more strongly triggered in the OSE model than in the MOG-induced EAE.

2.5.4 Mice with mild disease score of 1 show activation of adaptive immune system

As already mentioned, mice developing OSE exert a slower clinical disease course with more variability between animals, unlike MOG EAE mice, which develop the disease more rapidly (Krishnamoorthy et al., 2007). For that reason, the OSE poses a good model to study the disease at different stages. Mice with a mild disease score of 1 (OSE_1) were compared to OSE_0 mice, and 34 transcripts were found to be differentially expressed in specifically OSE_1 animals, and not in any other contrast. However, no significant GO biological processes were associated with the 34 transcripts (OSE_1-specific transcripts (OSE_1sp)), which are potentially representative of changes happening in the early stage of the disease. Nonetheless, it was observed that the transcripts differentially regulated in the OSE_1 and OSE_4 showed the same direction of regulation compared to OSE_0 (binomial test *p* value = 4.36×10^{-252}, 95% CI 0.995-1.0, Supplementary Table 3 in the paper by Faber, Kurtoic, and the colleagues (Faber et al., 2020)). Furthermore, for transcripts differentially expressed in OSE_1-OSE_0 but not in the control contrasts (OSE_1-expressed transcripts (OSE_1ex), Table 2.2) and association with 805 GO terms was detected, after correcting for multiple testing. The three GO terms previously highlighted were present among them (adjusted *p* value $< 2 \times 10^{-16}$, Figure 2.2). Interestingly, gene sets *B cell mediated immunity* and *antigen processing and presentation* were significantly overrepresented in both CDT gene set and OSE_1ex, a finding potentially revealing a role of B cells also in mildly affected OSE mice.

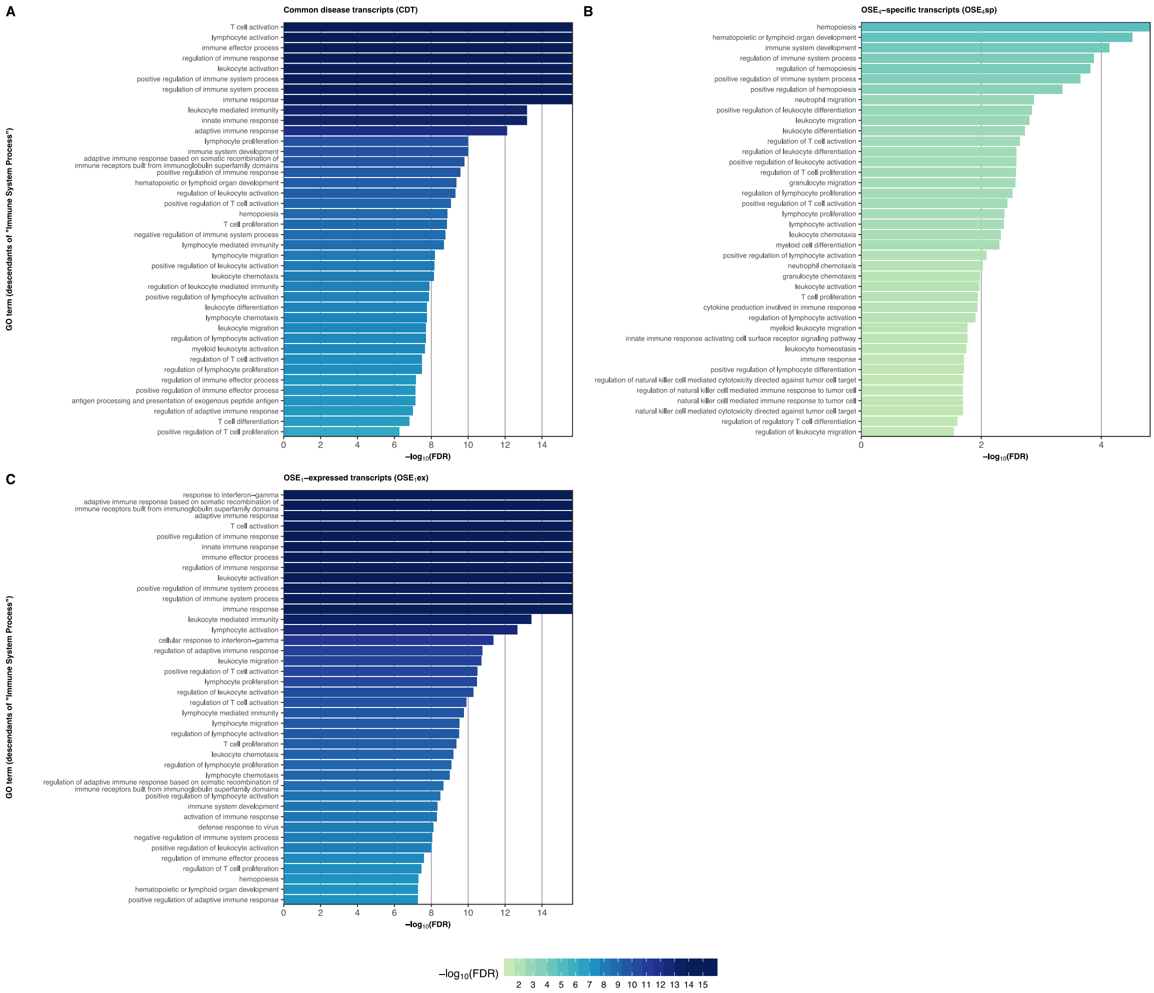

Figure 2.2 ***The GO enrichment analysis of differentially expressed genes.***
Figure source: the supplement of the research paper by Faber, Kurtoic, and the colleagues (Faber et al., 2020)
Plots show top 40 GO terms which are descendants of the term Immune System Process associated with differentially expressed transcripts grouped in the following groups ***(A)*** *CDT,* ***(B)*** *OSE$_4$sp,* ***(C)*** *OSE$_1$ex. There were no GO terms that are descendants of the term Immune System Process significantly overrepresented for the group MOG$_4$sp. The -log10(FDR) is estimated from the hypergeometric test* via *the ORA analyses. Bars of the plot are colored corresponding to the -log10(FDR). Darker color corresponds to the lower FDR value. The descriptions of group abbreviations are found in Table 2.2.*

2.5.5 Overrepresentation of MS susceptibility genes among transcripts expressed in OSE

There have been over 230 MS-associated genetic variants identified *via* GWA studies (Andlauer *et al.,* 2016; IMSGC, 2019b). Recently, a genetic association study using the genetic data from over 47,000 MS cases and over 68,000 healthy individuals was performed (IMSGC, 2019b). In the study, the functional impact of uncovered variants was evaluated in-depth, presenting a list of 551 putative human MS susceptibility candidate genes. In the data set analyzed in this work, gene expression from 499 transcripts mapping to 265 genes was available. PCA was conducted on these transcripts, where first component explained 75.7% of the variance in expression of these transcripts. Therefore, first PC was used to analyze whether the expression of MS susceptibility genes was increased in the EAE models. It was observed that the first PC was significantly higher in all diseased groups than in controls. Therefore, the levels of MS-associated genes were highly expressed in EAE, with highest levels observed in OSE$_4$ mice (Figure 2.3 **A**, Supplementary Table 6 in the paper by Faber, Kurtoic, and the colleagues (Faber et al., 2020)). To further support this, individual MS risk genes, *e.g.*, *H2-Ab1*, *Cd52*, and *Cd86* (Andlauer et al., 2016; IMSGC, 2019b) were analyzed, together with the putative MS-associated genes like *Cd74*. All of them were among transcripts exerting the lowest differential expression *p* values and significantly upregulated in all diseased mouse types, namely the OSE$_1$, OSE$_4$, and MOG$_4$ (Figure 2.1 **C-D**, Figure 2.3 **B-D**, Supplementary Table 5 in the paper by Faber, Kurtoic, and the colleagues (Faber et al., 2020)). Additionally, differentially expressed genes from sets CDT, OSE$_4$sp, and OSE$_1$ex were significantly enriched for MS risk genes, while genes in the MOG$_4$sp set did not show such enrichment (Table 2.3). Hence, in comparison to the MOG induced EAE model, the OSE model might be more closely connected to the human MS etiology.

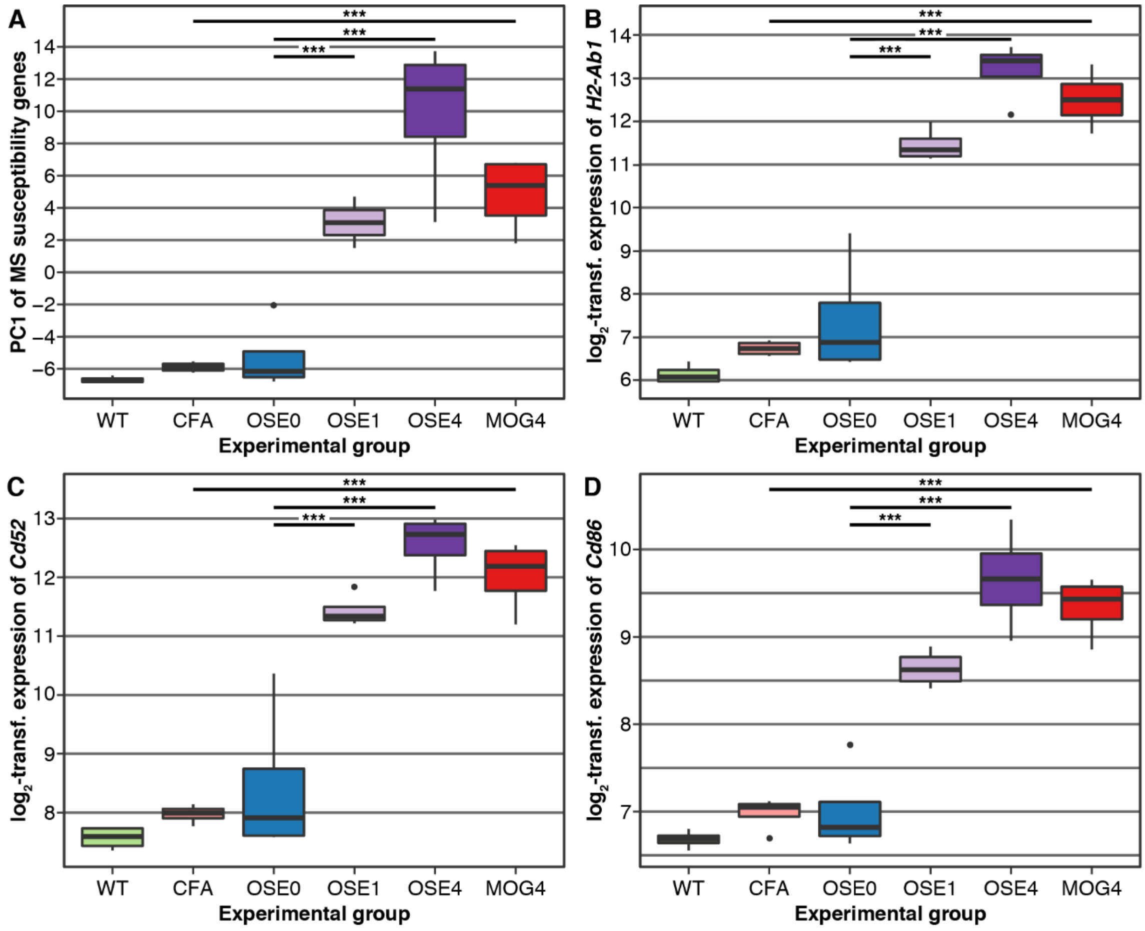

Figure 2.3 ***MS risk genes show a higher expression in diseased EAE mice, particularly in the OSE.***
Figure source: the research paper by Faber, Kurtoic, and the colleagues (Faber et al., 2020)
(A) *First PC of gene expression levels of MS risk genes. Diseased animals exert higher MS risk gene expression levels. Examples of expression levels of three putative MS risk genes,* H2-Ab1, Cd52, *and* Cd86. ***(B-D)*** *Significantly higher gene expression levels of all three MS risk genes were observed in diseased mice, with strongest effect for OSE$_4$ animals. PC = principal component (y-axis unit: standard deviations); Significance levels: *** adjusted* p *value < 0.001.*

Table 2.3 ***Enrichments of MS susceptibility genes.***
Table source: the research paper by Faber, Kurtoic, and the colleagues (Faber et al., 2020)
The 265 MS risk genes present in the data were tested for enrichment via *permutation analysis. Enrichment* p *values were computed based on 100,000 permutations and adjusted for multiple testing* via *the Holm-Bonferroni correction. Significant results are presented in bold font (adj.* p *value < 0.05). The descriptions of the DE transcript groups can be found in Table 2.2. DE = differentially expressed; WT = wildtype; adj. p-value = adjusted p value.*

DE transcript group	DE genes	Overlapping genes	*p* value	adj. *p* value
CDT	2014	68	**$<1\times10^{-5}$**	**$<4\times10^{-5}$**
OSE$_4$sp	2362	68	**4.4×10^{-4}**	**8.8×10^{-4}**
MOG$_4$sp	469	11	3.2×10^{-1}	3.2×10^{-1}
OSE$_1$ex	693	34	**1.0×10^{-5}**	**4.0×10^{-5}**

2.5.6 Gene expression of transcripts specific for T$_H$1 cells is more prominent in OSE mice

With T$_H$ cell differentiation pathway being highlighted as a crucial pathway involved in MS etiology (IMSGC & WTCCC, 2011), the relationship between gene expression changes in EAE models and T$_H$ cell differentiation was further examined. T cells from the spleen of OSE mice were *in vitro* polarized. The gene expression of T$_H$1 and T$_H$17 cells was compared to the gene expression of naïve T$_H$0 cells. There were 8 × more transcripts differentially expressed specifically in T$_H$1 than in T$_H$17 cells (Figure 2.4 **A**). All transcripts differentially expressed in T$_H$1 and T$_H$17 cells were regulated in the same direction.

Differential expression analysis revealed 1,080 transcripts significantly differentially expressed in T$_H$1 cells and 145 transcripts in T$_H$17 cells. It was further examined whether the T$_H$1- and T$_H$17-specific transcripts are more strongly expressed in diseased mice than in controls. The PCA was applied on the sets T$_H$1- and T$_H$17-specific transcripts, in which case the first component explained 49.6% and 68.6% of the variance, respectively. In both cases, the PC1 was significantly higher in all diseased groups than in the controls, with the difference being most prominent in OSE$_4$ animals (Figure 2.4 **B-C**, Supplementary Table 7 in the paper by Faber, Kurtoic, and the colleagues (Faber et al., 2020)). Signature molecules for T$_H$1 cells were examined as well, and *Tbx21* (*T-bet*) was significantly upregulated in all diseased mice. In OSE$_4$ mice, *Ifng* was upregulated as well (Figure 2.4 **D**, Supplementary Table 8 in the paper by Faber, Kurtoic, and the colleagues (Faber et al., 2020)). The same analysis was performed for marker molecules of T$_H$17 cells, where only *Il17f* was upregulated in OSE$_4$. In addition, *Rorc* and *Il17a* were tested. However, none of them were differentially expressed (Figure 2.4 **E**, Supplementary Table 8 in the paper by Faber, Kurtoic, and the colleagues (Faber et al., 2020)).

Furthermore, the gene expression of T_H1 and T_H17 specific transcripts was examined in CDT, OSE_4sp, OSE_1ex, and MOG_4sp gene sets. After adjusting for multiple hypothesis testing, the transcripts specific for CDT, OSE_4sp, and OSE_1ex groups were significantly enriched for both T_H1 and T_H17-specific transcripts (Table 2.4). In the MOG_4sp gene set, the sizes of the overlaps were lower and significant for T_H1-specific transcripts only. These data indicate that T_H cell-mediated immune responses associated with MS seem to be more prominent in OSE than in MOG-induced model.

Table 2.4 ***Examining the extent of overlap between T_H-specific transcripts and four gene sets, namely the CDT, OSE_4sp, OSE_1ex, and MOG_4sp.***
Table source: the research paper by Faber, Kurtoic, and the colleagues (Faber et al., 2020)
P values were computed based on 100,000 permutations. Enrichments significant after the Holm-Bonferroni multiple testing correction for eight tests are presented in bold font (adjusted p-value < 0.05). The descriptions of the DE transcript groups can be found in Table 2.2. DE = differentially expressed; WT = wildtype.

DE transcript group	DE genes	Cell type	Overlapping genes	*p* value	adj. *p* value
CDT	2014	T_H1	150	**$<1\times10^{-5}$**	**$<8\times10^{-5}$**
		T_H17	28	**2.0×10^{-2}**	**4.0×10^{-2}**
OSE_4sp	2362	T_H1	195	**$<1\times10^{-5}$**	**$<8\times10^{-5}$**
		T_H17	36	**2.0×10^{-3}**	**8.0×10^{-3}**
MOG_4sp	469	T_H1	35	**1.1×10^{-2}**	**3.3×10^{-2}**
		T_H17	7	9.8×10^{-2}	9.8×10^{-2}
OSE_1ex	693	T_H1	61	**2.0×10^{-5}**	**1.2×10^{-4}**
		T_H17	16	**1.0×10^{-3}**	**5.0×10^{-3}**

Next, it was explored whether EAE-associated genes differentially expressed in T_H1 or T_H17 cells were related to human MS. Each of the four gene sets was intersected with both T_H1 and T_H17-specific genes, generating altogether eight additional gene sets. Pathway analyses revealed that pathways involved in immune response were overrepresented for CDT, OSE_4sp, and OSE_1ex genes intersected with T_H1-specific genes (Figure 2.5). For the sets representing intersects with T_H17-specific genes no terms were found. The same result was observed for intersection between MOG_4sp gene set and T_H1-specific genes.

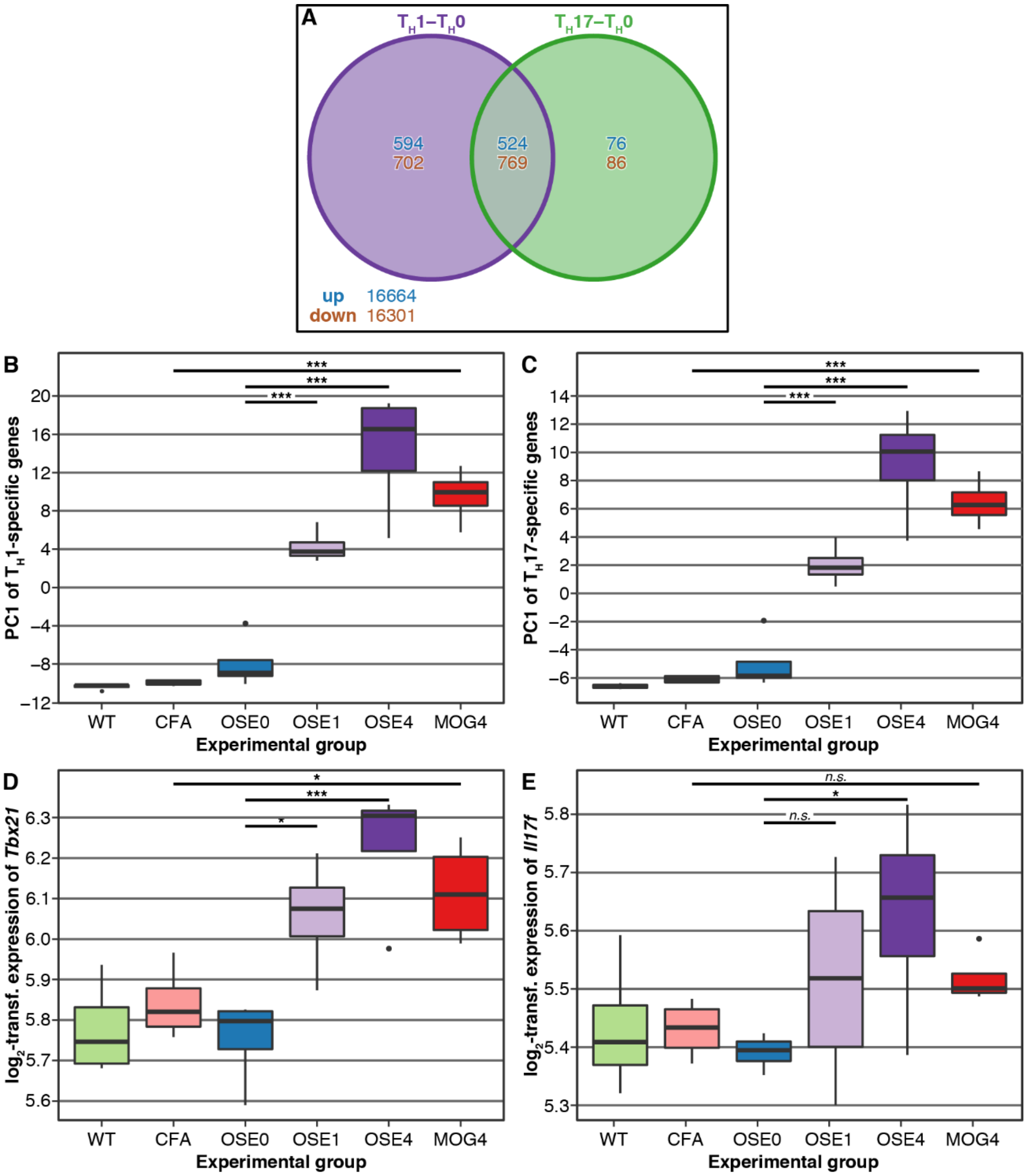

Figure 2.4 ***T_H1 and T_H17 specific genes show higher expression in disease EAE mice, especially in OSE_4.***
Figure source: the research paper by Faber, Kurtoic, and the colleagues (Faber et al., 2020)
(A) *Venn diagram showing the number of differentially expressed genes in two contrasts, T_H1-T_H0 and T_H17-T_H0. Here, up- and downregulated transcripts were analyzed separately, and transcripts differentially expressed in opposing directions are therefore included in the counts. PCA was run on the transcripts differentially expressed in T_H1* ***(B)*** *and T_H17 cells* ***(C)****. In diseased mice levels of T-cell-specific transcripts were higher (Supplementary Table 7 in the paper by Faber, Kurtoic, and the colleagues* (Faber et al., 2020)*). Gene expression of T_H1 signature molecule Tbx21* ***(D)*** *was significantly higher in diseased mice. On the other hand, when signature molecules for T_H17 were analyzed, only Il17f showed differential expression, an effect present in OSE_4 mice only. Significance levels: * adjusted p <0.05, ** adjusted p <0.01, *** adjusted p <0.001.*

Lastly, the enrichment of MS risk genes was examined in the eight lists. After correcting for multiple testing, the CDT and OSE$_1$ex genes differentially expressed in T$_H$1 cells were significantly enriched for the MS risk genes (p value $< 7 \times 10^{-4}$, Table 2.5). The enrichments for OSE$_4$sp, T$_H$17-specific gene sets and MOG$_4$sp were not considered significant after the multiple testing procedure. In conclusion, gene expression changes observed in OSE model include risk genes for human MS, an effect observed especially in T$_H$1 cells. In MOG-induced EAE, the effect was of a much smaller scope.

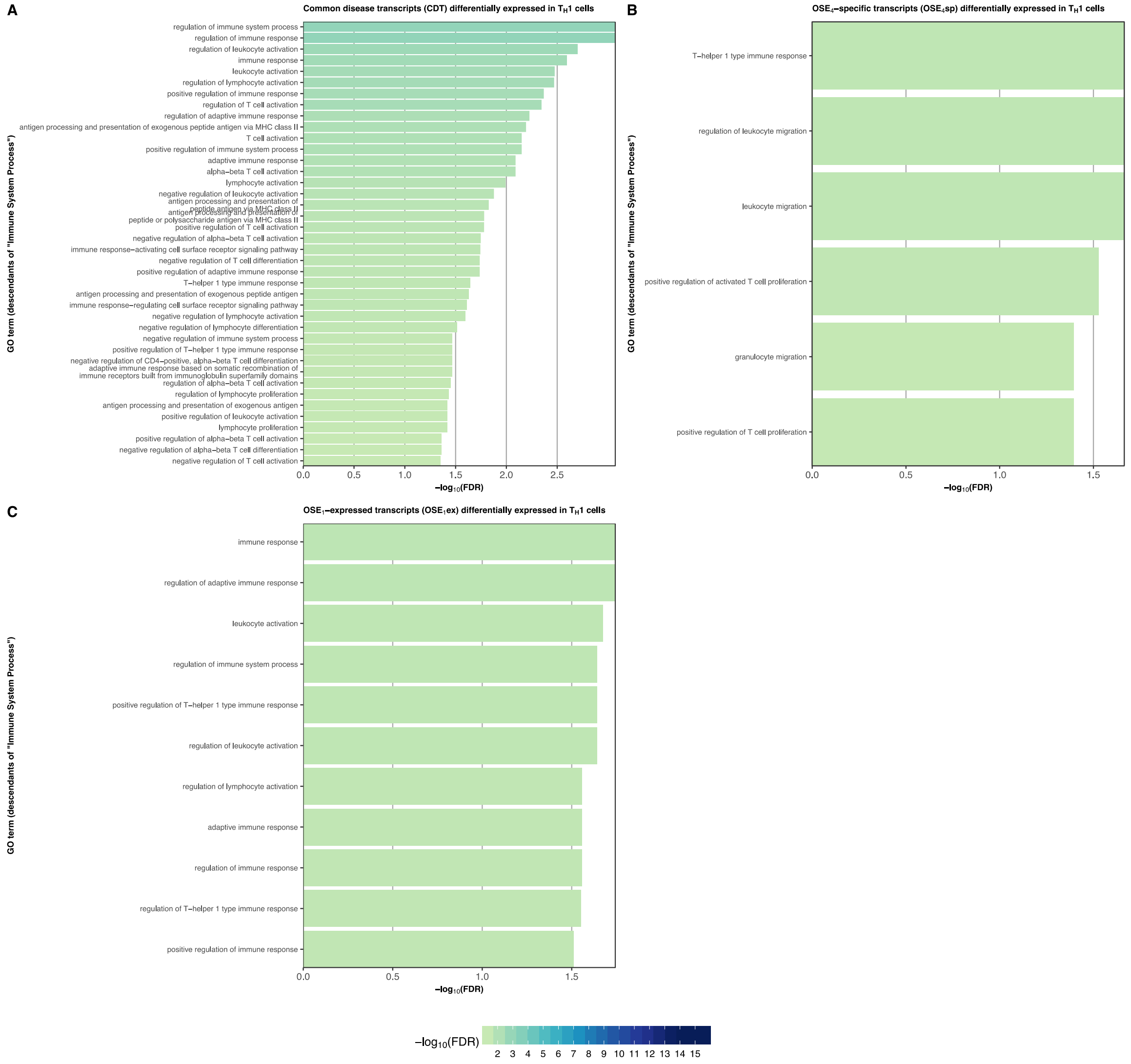

Figure 2.5 ***The GO enrichment analysis of differentially expressed transcript groups intersected with T$_H$1-specific transcripts.***

Figure source: the supplement of the research paper by Faber, Kurtoic, and the colleagues (Faber et al., 2020)

*The plots show the top 40 overrepresented GO terms that are descendants of the term Immune System Process for the transcript groups (**A**) CDT intersected with TH1-specific genes, (**B**) OSE4sp intersected with TH1-specific genes, (**C**) OSE$_1$ex intersected with T$_H$1-specific genes. Note that no GO terms were significantly overrepresented for any T$_H$17-specific or MOG$_4$sp genes. The descriptions of the DE transcript groups can be found in Table 2.2. The -log10(FDR) from hypergeometric tests is shown on the x-axis and used for coloring the plots (darker colors represent lower FDRs).*

Table 2.5 ***Enrichment of the putative MS risk genes among T_H cell-specific transcripts.***
Table source: the research paper by Faber, Kurtoic, and the colleagues (Faber et al., 2020)
The p values are estimated based on 100,000 permutations. Results significant after applying the Holm-Bonferroni method for multiple hypothesis testing (0.05/8) are presented in bold font (adjusted p *value < 0.05). The descriptions of the DE transcript groups can be found in Table 2.2. DE = differentially expressed; WT = wildtype.*

DE transcript group	Cell type	EAE T_H cell list size	Overlapping genes	*p*-value	adj. *p*-value
CDT	T_H1	150	10	**6.5×10^{-4}**	**5.2×10^{-3}**
	T_H17	30	3	2.1×10^{-2}	1.1×10^{-1}
OSE_4sp	T_H1	215	10	9.7×10^{-3}	5.8×10^{-2}
	T_H17	41	3	4.7×10^{-2}	1.6×10^{-1}
MOG_4sp	T_H1	37	1	5.1×10^{-1}	5.1×10^{-1}
	T_H17	7	1	1.3×10^{-1}	2.6×10^{-1}
OSE_1ex	T_H1	60	6	**1.1×10^{-3}**	**7.7×10^{-3}**
	T_H17	16	2	3.9×10^{-2}	1.6×10^{-1}

3 | A differential network approach to explore the influence of genetic variation on gene expression in the early stages of multiple sclerosis

3.1 Research question

Over 230 genomic loci associated with multiple sclerosis susceptibility have been identified in genome-wide association studies. However, we still lack the knowledge about the underlying biology of these genetic effects. In this work, it was hypothesized that genetic variation influences the regulation in gene co-expression networks. Contrasting such networks might provide deeper insights into fundamental differences between subgroups of MS patients and help identify patterns in the development of the disease.

3.2 Motivation

The translation of genome wide association studies into functionally relevant biology poses a tedious task. The differential network approach is considered a promising tool in that aspect because it enables the analysis of interactions between genes. Gaussian graphical models, which represent a method for sparse graph estimation estimating the conditional dependence between variables, has already shown promising results in metabolomics (Krumsiek, Suhre, Illig, Adamski, & Theis, 2011; Trinh Do et al., 2014), breast cancer research (Dobra et al., 2004), as well as in transcriptomics (de la Fuente, Bing, Hoeschele, & Mendes, 2004; Schäfer & Strimmer, 2005c). When studying the gene expression data, the method enables finding directly associated gene expression profiles, which might be crucial in detecting pathways mediating the effect of genetic variation on disease development. It is important to discern the mechanisms of genetic influence on gene expression in the early stages of the disease in order to better understand the disease etiology as well as to develop a more personalized route for MS diagnosis and treatment. The workflow for the differential network analysis presented in this work can be employed to analyze other complex diseases as well, and, for example, compare cases and controls or gene expression data from treated and untreated individuals in order to find differences in pathway regulation.

3.3 Finding patterns in gene expression data

In this work, the effect of genetic variation on gene expression data from MS patients was investigated, with all the patients in the early stages of the disease or diagnosed with clinically isolated syndrome (CIS). Groups of patients differing in their genetic background were contrasted using different approaches, with major interest in differential network analysis (Figure 3.1). MS-associated variants included in this study were preselected based on the previous GWAS (Andlauer et al., 2016), and patients were split into groups, depending on whether they carry the risk variant or not. Groups were compared on four different levels (Figure 3.1). First, differential gene expression was applied to grasp the scope of potential genetic effect on gene expression profiles. Additional layer of information can be added by performing the expression quantitative loci (eQTL) analysis, which is used to examine the impact of genetic variants on adjacent and nonadjacent gene expression. As an intermediate step, Pearson's correlation matrices were calculated for each group to examine the variation in pure correlation. Finally, the weighted gene co-expression network analysis (WGCNA) and the Gaussian graphical model (GGM) were employed to examine the effect of genetic variation on gene co-expression networks with the aim of identifying biologically meaningful subunits mediating the influence of genetic variation on the disease.

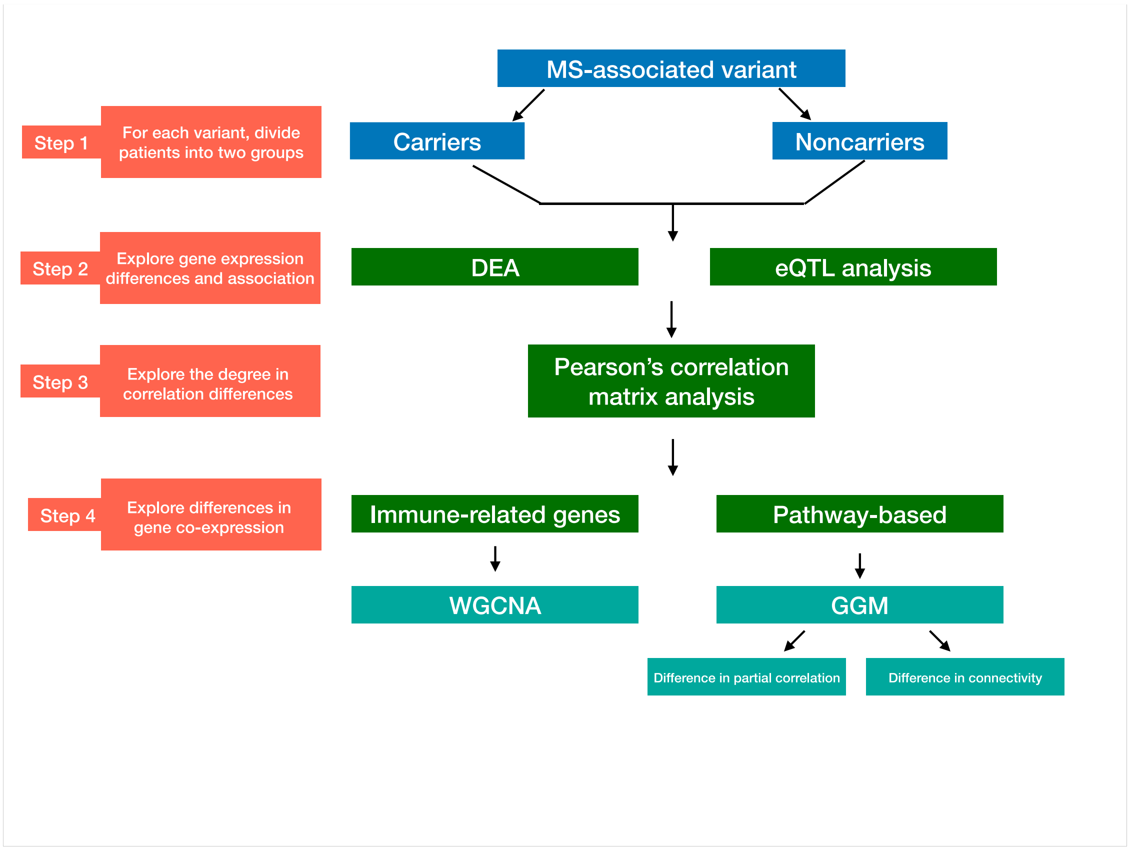
MS-associated variant
Step 1
For each variant, divide patients into two groups
Carriers
Noncarriers
Step 2
Explore gene expression differences and association
DEA
eQTL analysis
Step 3
Explore the degree in correlation differences
Pearson's correlation matrix analysis
Step 4
Explore differences in gene co-expression
Immune-related genes
Pathway-based
WGCNA
GGM
Difference in partial correlation
Difference in connectivity

Figure 3.1 ***Representation of the analysis workflow in four steps.***
For each MS-associated variant analyzed, patients were divided into two groups to explore the variant influence on gene expression. Carriers = dosage > 0.5; non-carriers = dosage <= 0.5. DEA = differential gene expression analysis, WGCNA = weighted gene co-expression analysis, GGM = Gaussian graphical model. The figure was created with the Keynote application on MacBook Pro.

3.3.1 Introduction to networks – biological networks

The information in our body flows from the source of the trigger through the chain of chemical reactions and never stops moving. Therefore, our body system can be described as a network where each trigger and each response have a certain role and position. Such network is complex and dynamic, because there are various types of triggers and responses, which constantly intertwine. We can explore metabolic reactions, protein-protein interactions, interactions between genes, interactions between single nucleotide polymorphisms (SNPs), etc. However, the basic structure of all networks remains unchanged regardless of the matter being studied. Each network consists of nodes and links between the nodes. A graphical representation of a network is called a graph in mathematical terms. In a graph, an edge is drawn between the two nodes if two nodes are associated (Barabási & Oltvai, 2004). Measures of associations vary, a commonly used one being Pearson's correlation and its variations. A network is further characterized by node degree or node connectivity – the number of connections a node in a network has formed with other nodes. For example, in a random network, the node degree follows the Poisson distribution, where all nodes have approximately the same number of edges. On the other hand, scale free networks are defined by a power law distribution where most of the nodes have very few connections and a small proportion of nodes forming a high number of connections holds the network together. Therefore, in such networks, there is no typical node which could be used to characterize the rest of the nodes (Barabási & Oltvai, 2004), *i.e.*, no characteristic scale to describe the node degree values (Ravasz, Somera, Mongru, Oltvai, & Barabási, 2002). The power law distribution is defined as

$$P(k) \sim k^{-\gamma} \tag{3.1}$$

where *k* stands for the number of connections of a node and *P(k)* is the probability of a node having *k* connections. In networks free of scale, new nodes are added to already established subnetworks, a property observed in biological networks as well. Thus, biological networks are considered scale free. This is a very interesting observation, because scale free networks show tolerance against errors while being vulnerable to deletions (B. Zhang & Horvath, 2005). That is, removal of a small number of important nodes may be detrimental to the network structure. Studying biological processes in the context of graphs represents an important tool to explore changes in the system, both prominent and subtle.

Biological networks are commonly estimated from gene expression data, because genes involved in the same biological pathway are either controlled by the same transcription factor or functionally related otherwise, and therefore have similar expression levels. These are the hallmarks of the guilt-by-association principle (Wolfe, Kohane, & Butte, 2005). In this work, gene co-expression networks were estimated using whole blood gene expression data from MS patients. Whole blood is commonly used as a proxy for studying disorders of the central nervous system (Wittenberg, Greene, Vértes, Drevets, & Bullmore, 2020), due to the correlation in gene transcription between human blood and the CNS.

Two approaches will be employed to analyze the influence of MS-associated variants on gene co-regulation: the WGCNA and the GGM. The major difference between the two is that the GGMs account for the effect of other genes on pairwise gene associations *via* conditional independence approach. In that way, two genes have a (partial) correlation different than zero if they are conditionally dependent after the effects of other genes in the network have been removed. The WGCNA, on the other hand, estimates the data driven soft threshold which removes spurious pairwise correlations between genes and in that way finds clusters of interconnected genes, exerting similar functions in the cell.

3.3.1.1 Weighted gene co-expression network analysis (WGCNA)

The WGCNA enables analysis of gene co-expression networks by finding highly correlated subnetworks – modules. It has been shown that genes which cluster tightly within the module exert similar functions in the cell (Fuller et al., 2007). Following Langfelder and Horvath (Langfelder & Horvath, 2008), a correlation network was constructed on the basis of Pearson's correlation between expression levels of genes in the gene set. Let X be an $n \times m$ matrix, where row indices correspond to network nodes (i = 1, …, n) and the column indices (j = 1, …, m) correspond to sample measurements

$$X = [x_{ij}] = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}. \tag{3.2}$$

A network module is a set of rows of X which are highly associated. To detect such substructures, the adjacency matrix a_{ij} needs to be determined. The adjacency matrix fully specifies the network, and its elements take the values [0,1]. First, an intermediate matrix describing co-expression similarity s_{ij} is defined as an absolute value of the Pearson's correlation coefficient

$$s_{ij} = |\, cor(x_i, x_j)| \,. \tag{3.3}$$

By applying the threshold, we transform the intermediate matrix into the adjacency matrix. The cut off can be selected by applying either hard or soft thresholding. Hard thresholding introduces τ such that

$$a_{ij} = \begin{cases} 1 & \text{if } s_{ij} \geq \tau; \\ 0 & \text{otherwise.} \end{cases} \tag{3.4}$$

In that case, two genes are linked (a_{ij} = 1) if their absolute correlation exceeds the hard threshold τ. Adjacency matrices with such binary elements are representative of unweighted networks. However, to be able to reflect the continuous nature of the co-expression information, an adjacency matrix needs to be able to take continuous values between 0 and 1 as well. This is achieved by raising the co-expression similarity to a power β

$$a_{ij} = s_{ij}^{\beta}. \tag{3.5}$$

Hence, in weighted networks, $0 \leq a_{ij} \leq 1$. The selection of hard and soft threshold can be guided by following the scale free criterion, *i.e.*, finding such a cut off to satisfy the Eq. (3.5). Once the adjacency matrix is calculated, the network is defined (Langfelder & Horvath, 2008). The next step is to detect highly correlated network substructures – modules. Among other approaches, the topological overlap measure (TOM) can be employed (Langfelder & Horvath, 2008). The TOM enables examining gene pairs in the context of the network as a whole, not as an isolated quantity, as is the case when calculating the adjacency matrix. Genes are described as highly topologically similar if they connect to the same groups of other genes in the network. Application of the TOM results in a similarity matrix $\Omega = [\omega_{ij}]$ whereby each input describes the strength of topological overlap for each gene pair

$$\omega_{ij} = \frac{l_{ij} + a_{ij}}{\min\{k_i,\ k_j\} + 1 - a_{ij}} \tag{3.6}$$

where $l_{ij} = \sum_u a_{iu} a_{uj}$, and $k_i = \sum_u a_{iu}$ denotes the connectivity of the node i. ω_{ij} is defined on the interval [0, 1], whereby $\omega_{ij} = 1$ if the node with fewer connections satisfies the following conditions: (i) neighbors of two nodes completely overlap and (ii) the two nodes are connected. $\omega_{ij} = 0$ if two nodes i and j do not share neighbors and are not connected (B. Zhang & Horvath, 2005). Next, the dissimilarity matrix is calculated

$$d_{ij}^{\omega} = 1 - \omega_{ij}. \tag{3.7}$$

The dissimilarity matrix d is used as a distance matrix for hierarchical clustering. The results of the hierarchical clustering can be visualized by a dendrogram where tree branches correspond to modules, *i.e.*, groups of genes showing high topological overlap. Traditionally, the dendrograms are cut on a fixed height, whereby the user defines the height of the cutting point. The challenge here is selecting the cutting point in the tree. One alternative to fixed branch height cut is the dynamic tree cut method (Langfelder, Zhang, & Horvath, 2009). Dynamic cut tree method is based on analyzing the shape of the branches on the dendrogram. A series of cluster decompositions and combinations is run to find the stable number of clusters. Dendrogram heights and reference heights l_m, l_u and l_d are defined. First, the tree is cut on the height l_m, which is typically very high, resulting in a small number of big clusters. The dendrogram of each cluster is further cut separately, according to the reference heights. If new clusters are created during the process, the algorithm is repeated. The procedure is run until no new clusters appear. The resulting modules can be finally processed by merging those which still show high degree of correlation. The variation of each module can be summarized by calculating the module eigengene (ME), which is the first principal component of gene expression matrix of a given module. MEs can then be used as module representatives and the correlation with phenotypic traits can be explored. In such a way, the multiple testing problem is greatly alleviated.

3.3.1.1.1 Module preservation

Module preservation is a procedure of comparing module topology between the reference and the test group (Langfelder, Luo, Oldham, & Horvath, 2011). If we have strong evidence that the topology of the module is preserved between the reference and the test network, we can say that the respective module does not show group-specific network characteristics. Depending on the question, applying such technique can help in finding modules which distinguish different conditions, *e.g.*, the human brain and blood tissue (Cai et al., 2010) or human and mouse brains (J. A. Miller, Horvath, & Geschwind, 2010). Network statistics which reflect the potential module preservation are as follows: density-based preservation statistics, separability-based and connectivity-based. Density-based preservation statistics determine whether module nodes remain highly connected in the test network, while separability-based statistics determine whether network modules remain separated from another in the test network (Langfelder et al., 2011). Furthermore, according to the paper by Langfelder and the colleagues (Langfelder *et al.*, 2011), density-based statistics outperform the separability-based approaches in discriminating module preservation. Connectivity-based preservation statistics compare the node connectivity pattern of the two networks of interest. When calculating different preservation statistics, the results often don't fully match, and it is therefore useful to aggregate different module preservation statistics into a summarized preservation statistics, the $Z_{summary}$

$$Z_{summary} = \frac{Z_{density} + Z_{connectivity}}{2}. \tag{3.8}$$

A higher $Z_{summary}$ statistics suggests that we have stronger evidence that our module is group specific. Langfelder and the colleagues (Langfelder *et al.*, 2011) hence suggest the following scale to describe the strength of $Z_{summary}$ statistics: if $Z_{summary} > 10$, there is strong evidence that the module is preserved; if $2 < Z_{summary} < 10$ the evidence is weak to moderate and if $Z_{summary} < 2$, there is no evidence that the module is preserved.

3.3.1.2 Gaussian graphical model (GGM)

Genes could be weakly correlated in terms of the Pearson's correlation, but highly related on the partial correlation level, for which the influence of other genes in the gene set is taken into account (Dobra et al., 2004). The approach employed in this work was the Gaussian graphical models (GGM) (Schäfer and Strimmer, 2005a), a specific graph-based method that explores the conditional independency between each pair of genes in a dataset. This method reveals direct interactions between genes based on their partial correlation, thereby removing potentially spurious connections between genes. The GGM has already shown good performance in inferring meaningful sub-networks originating from yeast gene expression data (de la Fuente et al., 2004). Furthermore, networks were inferred from breast cancer gene expression data providing evidence that genes with crucial roles in tumor growth and transcription factors form high number of direct connections (Schäfer & Strimmer, 2005b).

3.3.1.2.1 Estimation of partial correlation coefficients in a small sample setting

Obtaining a stable estimate of the covariance matrix precedes the partial correlation estimation and poses a problem in a big data setting. According to the standard graphical theory (Whittaker, 1990), the matrix of partial correlations $\tilde{P} = (\tilde{p}_{ij})$ is related to the inverse of the covariance matrix Σ. The standard graphical theory cannot be applied in a "small *n,* big *p*" data setting, where *n* is the number of samples and *p* the number of predictors. In bioinformatics, the accurate and reliable estimate of the population covariance matrix is obtained using maximum likelihood where the covariance matrix S^{ML} is estimated (Schäfer & Strimmer, 2005b). Further following the paper by Schäfer and Strimmer (Schäfer & Strimmer, 2005b), the entries of such matrices are defined as

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x_\iota})(x_{kj} - \bar{x_\jmath}) \tag{3.9}$$

where

$$\overline{x_\iota} = \frac{1}{n}\sum_{k=1}^{n} x_{ki} \tag{3.10}$$

with *k* being the k-th observation of the variable X_i. Such data often suffers from a n<<p problem, whereby the number of predictors tremendously surpasses the number of samples. This also creates an issue with the empirical covariance matrix estimation, resulting in an ill-posed matrix which cannot be inverted. Such matrix is also a very poor estimate of the true population covariance matrix (Schäfer & Strimmer, 2005a). Schäfer and Strimmer (Schäfer and Strimmer, 2005b) further applied the Ledoit-Wolf lemma (Ledoit & Wolf, 2003) to combat this problem and also implemented their approach in the *GeneNet* package in *R*. The Ledoit-Wolf theorem enables the shrinkage of the covariance matrix by estimating the shrinkage parameter in an analytic way. Therefore, a regularized estimate of covariance matrix is obtained, which is well-posed. Following this, the estimated covariance matrix $\hat{\Sigma}$ can be inverted. Through the intermediate matrix $\widehat{\Omega}$, partial correlations are obtained

$$\tilde{r}_{ij} = \tilde{\hat{\rho}}_{ij} = -\frac{\widehat{\omega}_{ij}}{\sqrt{\widehat{\omega}_{ii}\,\widehat{\omega}_{jj}}} \tag{3.11}$$

where

$$\widehat{\Omega} = \left(\widehat{\omega}_{ij}\right) = \hat{\Sigma}^{-1} \tag{3.12}$$

Once the partial correlation matrix is estimated, the proportion of null edges needs to be estimated as well. Based on the research from Schäfer and Strimmer (Schäfer and Strimmer, 2005a), the distribution of the observed partial correlations $\tilde{r}$ across edges is modelled as a mixture

$$f(\tilde{r}) = \eta_0 f_0(\tilde{r}\,; \kappa) + (1-\eta_0) f_A(\tilde{r}\,). \tag{3.13}$$

Here, f_0 is the null distribution, η_0 is the unknown proportion of the indirect edges (“null edges”), f_A is the distribution of observed partial correlations assigned to the actually existing edges. The characteristics of the null distribution are given in the paper by Hotelling in 1953 (Hotelling, 1953). Local false discovery rates (fdr) are then calculated as follows:

$$p(null\ edge|\tilde{r}) = fdr(\tilde{r}) = \frac{\hat{\eta}_0 f_0(\tilde{r}; \kappa)}{\hat{f}(\tilde{r})} \tag{3.14}$$

That is, the posterior probability that an edge is null given the value of $\tilde{r}$. Therefore, for each pairwise partial correlation, posterior probabilities, and the local fdr are calculated, which can be the basis to select direct edges in the gene set. Such associations might be left undiscovered by using pure Pearson’s correlation and may indeed reflect important biological relationships.

3.3.2 The differential network approach

The differential network analysis provides formal statistical tests to explore the change in network structure between the two conditions of interest (Gill, Datta, & Datta, 2010). In the following paragraphs, the aspects of connectivity analysis of pathway-based GGM networks will be introduced.

3.3.2.1 GGM-based differential network analyses

In the previous research, GGM-based networks were compared by, for example, examining the partial correlation of an edge or connectivity patterns of the nodes in a network. Zannas and the colleagues (Zannas et al., 2019) inferred pathway-based GGM network and showed that the signaling in the NF-κB pathway is promoted by the expression changes of *FKBP5* gene, coding for a stress-responsive molecule (Zannas et al., 2019). In another example, the Gaussian graphical model was used to develop an approach which directly estimates the differential network. The approach includes rigorous statistical tests to determine the difference of conditional independence between two conditions (He et al., 2019). Furthermore, the analysis of connectivity profiles in pathway-specific GGM networks have been applied as well. For example, the analysis of connectivity of the Signal NOTCH1 pathway in mice has shown that the *Kat2b* gene has an important role in the craniofacial development in mice. In neuroblastoma tumor samples, the comparison of connectivity profiles between the gene expression of clinically high-risk (HR) neuroblastoma patients and that of non HR neuroblastoma patients showed that weaker connectivity of proto-oncogene SRC leads to metastatic behavior in HR patients (Grimes, Potter, & Datta, 2019).

3.3.2.2 Differential network approach examining node connectivity

Node connectivity or the node degree is the sum of all connections of a node in a network. The degree reflects the role of the gene, where functionally more important genes tend to have more connections. Therefore, by inspecting the connectivity of nodes in two different conditions, one could detect important changes in network regulation. Nodes with a substantial change in connectivity are suspected to have an important role in the disease phenotype. Exploring node connectivity in addition to differential gene expression has revealed genes which were overlooked when differential expression was studied alone (Leonardson et al., 2009; Reverter et al., 2006). This indeed makes sense, because regulatory function of a gene can change even without the change in gene's expression. A mutation or a post-translational modification can lead to altered gene function, thereby changing expression of other interacting genes (de la Fuente, 2010).

A very important feature when employing the GGM as a network estimator is a clear distinction between direct and indirect edges. One can use the following criteria to distinguish between the two types of edges: the partial correlation coefficient, the posterior probability of an edge being direct, or a q value (FDR adjusted p value). Each of the measures is edge specific, *i.e.*, unique for each gene pair in the analyzed network. However, none of the mentioned measures enables incorporating the knowledge of the pathway structure in the purpose of constructing a prior. The incorporation of a good quality prior in the analysis can be crucial, if such a prior does exist. In this work, it was decided to select direct edges in the networks by using the Bayesian approach to hypothesis testing, which involves the incorporation of prior knowledge when estimating the number of direct edges in a pathway. The approach will be explained in the next paragraphs.

3.3.2.2.1 Bayes factor: incorporating informative prior to select direct edges in a network

One can assume that there is a higher probability that genes in a pathway-based network should form more direct connections, than those in a network consisting of a random set of genes, simply because we know that those genes are already annotated to exert a similar function. A statistical approach one can use to incorporate such prior knowledge is the Bayesian approach to hypothesis testing. In the context of this project, we are comparing two Gaussian graphical models. Both models consist of a certain set of parameters, among which is the partial correlation. In the simpler model (M_0), this correlation is estimated to be zero. On the other hand, in the more complex model (M_1), the correlation coefficient is different than zero. In other words, the simpler model describes an indirect edge between two nodes, while the complex model describes a direct edge.
Following Kass and Raftery (Kass and Raftery, 1995), we begin with the data D, assumed to have risen under M_0 or M_1, with two probability densities pr(D | M_0) or pr(D | M_1). The a priori probabilities are given as pr(M_1) and pr(M_0) = 1 – pr(M_1). Hence, the a posteriori probabilities are defined as pr(M_0 | D) and pr(M_1 | D). That is, every prior gets transformed into a posterior after observing the data. Further, based on the Bayes theorem it follows

$$pr(M_k\ |D) = \frac{pr(D\ |M_k)\ pr(M_k)}{pr(D\ |M_0)\ pr(M_0) + pr(D\ |\ M_1)pr(M_1)} \tag{3.15}$$

where k = (1, 2), so that

$$\frac{pr(M_0|D)}{pr(M_1|D)} = \frac{pr(D\ |\ M_1)}{pr(D\ |\ M_0)} \times \frac{pr(M_1)}{pr(M_0)} \tag{3.16}$$

Equation 3.16 explains the transformation of the prior beliefs about the model to posterior beliefs, by incorporating the predictive updating factor – the Bayes factor (BF)

$$BF_{10} = \frac{pr(D\ |\ M_1)}{pr(D\ |\ M_0)}\ . \tag{3.17}$$

Thus, the Bayes factor represents the ratio between posterior and prior odds of a model. It is the summary of the evidence provided by the data concerning one scientific model or theory (Kass & Raftery, 1995). In this specific example, the B_{10} is weighing the evidence *against* the null model. This is what “10” stands for, it denotes which model is in the numerator versus the denominator. There is a consensus regarding Bayes factor categorization, presented in Table 3.1.

Table 3.1 ***Bayes factor categorization (adjusted from Kass and Raftery, 1995).***

$\log_{10}(BF_{10})$	BF_{10}	Evidence against M_0
0 to 1/2	1 to 3.2	Anecdotal
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

We can use the Bayes factor categorization to evaluate the strength of the evidence that an edge is direct in a pathway-specific GGM network. The GGM provides us with a posterior probability that an edge is direct based on the estimated model. It is then straightforward to calculate the prior probability

$$pr(D|M_0) = 1 - pr(D|M_1). \tag{3.18}$$

From there, the Bayes factor follows

$$BF_{10} = \frac{pr(D \mid M_1)}{pr(D \mid M_0)} = \frac{pr(D \mid M_1)}{1-pr(D \mid M_1)}. \tag{3.19}$$

In that way, for each edge in the gene set, a corresponding BF is calculated. The strength of the BF corresponds to the strength of evidence in favor of M_1 , that is, in favor of an edge being direct.

A widely used frequentists approach to hypothesis testing relies on *p* values to determine whether one has enough evidence to reject the null model or null hypothesis, depending on the context. In this approach, a *p* value of 0.05 denotes that if we reject the null hypothesis, we would make a type I error in 5% of the times (that is, reject the null hypothesis when it is actually true). The corresponding BF to the 0.05 *p* value is ≈ 3 ($10^{-1/2}$), meaning that we have 3 times stronger evidence for M_1. Down the same line, the BF of 10 corresponds to the *p* value of 0.01 (Jeffreys, 1961). Targeted levels of evidence are researcher-defined (Kass & Raftery, 1995).

3.4 Materials and methods

3.4.1 Materials

3.4.1.1 The KKNMS cohort

German Competence Network Multiple Sclerosis (*Krankheitsbezogene Kompetenznetz Multiple Sclerose*, KKNMS) is an interdisciplinary network focused on improving MS diagnostics, treatment and patient care, as well as providing a better understanding of the underlying genetic effects. The KKNMS cohort consists of 1,019 patients diagnosed with clinically isolated syndrome (CIS) or an early-stage MS, in which patients are followed-up longitudinally. The KKNMS cohort study inclusion criteria are: at least 18 years of age, clinically isolated syndrome within the past 6 months or early relapsing-remitting MS (RRMS) for at most 2 years after initial symptoms and no previous long-term treatment for MS (Johnen et al., 2019; von Bismarck et al., 2018).

In this research project, patients data at the baseline was examined, that is, their first visit to the hospital. After the QC, the gene expression data was available for 399 individuals. Among these 399 individuals, 317 individuals were genotyped for single nucleotide polymorphisms. Three individuals were excluded due to treatment that was still ongoing when the blood samples were collected, therefore affecting the gene expression too heavily. Table 3.2 summarizes the characteristics of a subset of KKNMS patients analyzed in this thesis. Most of the 314 individuals are females (68.5%), corresponding to the overall gender distribution of MS patients in the population. A small group of patients received the disease modifying treatment (DMT) at the time of the blood draw. The effects of DMT and other treatments on gene expression were accounted for in the regression analysis, as explained in the Methods section (3.4.2.2 Calculation of residual sum of squares). Nine percent of the patients were experiencing an active infection at the time of the blood draw, and nearly half of the patients had an ongoing relapse.

Table 3.2 ***Descriptive statistics, treatment information, and clinical characteristics of patients in the subset of patients from the KKNMS cohort analyzed in this work.***
DMT = disease modifying treatment.

Variable	Sample
N	314
Demographics	
Age, mean (SD)	34.3 (9.1)
Gender, N (%)	215 (68.5)
DMT	
Glatiramer acetate, N	3
Interferon Beta, N	3
Other treatment	
Vitamin-D, N	13
Analgesics, N	20
Anti-histamines, N	7
Mesalazin, N	3
Budesonide, N	2
Cortisol, N	42
Clinical status	
Active infection, N	9
Current relapse, N	44

3.4.1.2 Microarrays

Gene expression from the whole blood of MS patients was measured on Illumina HumanHT-12v4 Expression BeadChip. The data was loaded using the *beadarray* package (Dunning et al., 2007). The QC was performed in *R* v3.2.1. (R Core Team, 2020) using the packages *vsn* and *lumi* (Du et al., 2008; Huber et al., 2002). Probes showing detection *p* value less than 0.05 in more than 10% of the samples, probes that could not be mapped to a known transcript, or those that were identified as cross-hybridizing by the Re-Annotator pipeline (Arloth, Bader, Röh, & Altmann, 2015), were removed. Probes were mapped to unique Entrez identifiers (Entrez ID). For gene co-expression analysis, transcripts were further collapsed to the gene level, that is, each transcript mapped to exactly one HGNC gene symbol. Collapsing was based on the highest mean expression approach, *i.e.*, among multiple probes hitting the same gene, the probe with the highest gene expression mean was selected. This left 13,442 transcripts in the dataset from 314 individuals. For differential expression and eQTL analysis, the collapsing step was skipped, therefore leaving 19,420 transcripts. In order to identify technical batch effects, the PCA was run on the gene expression data. Next, the first two PCs were correlated with amplification round, amplification plate, and amplification plate column and row, as well as expression chip, as described in the research paper by Andlauer and the colleagues (Andlauer et al., 2016). The expression data were adjusted using the *ComBat* package (Johnson, Li, & Rabinovic, 2007).

3.4.1.3 Genotype data

Patients were genotyped on Illumina OmniExpress BeadChips. SNPs were imputed using the 1000 Genomes Phase 3 reference panel, whereas HLA alleles were imputed using the T1DGC reference panel. 19 variants associated with MS were analyzed in this research project, based on a previous research study, in which MS-associated variants and alleles were determined in the GWA study (Andlauer et al., 2016). This included four allelic changes in the MHC region and 15 minor frequency alleles of SNPs outside of MHC region (Table 3.3).

3.4.2 Methods

3.4.2.1 Study design

In order to examine the influence of an MS-associated variant on gene expression in MS patients, patients were split into groups pertaining their variant carrier status, based on their dosage data. Dosage data represent transformed posterior genotype probabilities which come from genotype imputation. Dosage is a continuous variable, with values ranging from 0 to 2. In this work, the dominant association model was used to split patients into groups of variant carriers and non-carriers. In the dominant model, heterozygous individuals and individuals carrying two copies of the risk allele (homozygous individuals) were grouped together (Bae, Perls, Steinberg, & Sebastiani, 2015). The following thresholds for dosage data were introduced: if a patient had a dosage lower than or equal to 0.5, the patient was labelled as a noncarrier of the respective MS-associated variant. If for an individual the dosage above 0.5 was estimated, the person was considered a variant carrier. The Table 3.3 summarizes the group sizes depending on the variant, as well as the related odds ratios. Odds ratios as well as (minor) allele frequencies ((M)AF) were calculated in a bigger sample comprising of KKNMS patients and six other samples of MS patients, all recruited across Germany (Andlauer et al., 2016). In the paper by Andlauer and the colleagues (Andlauer et al., 2016), altogether 16 SNPs outside of MHC and seven MHC variants were presented. The reasons why some of them weren't included in the analyses of this work are the following. First, the rs3104373 variant outside of the MHC region is the proxy SNP for the HLA-DRB1*15:01 allele. Second, the three HLA alleles, namely the HLA-B*38:01, HLA-DRB1*13:03, and HLA-DRB1*08:01 have allele frequencies < 5%, and the groups of patients carrying the allele would be too small, not providing enough power for complex network analyses.

Table 3.3 ***19 MS-associated variants analyzed in this work, their related odds ratios, and group sizes.***
Odds ratios as well as (minor) allele frequencies ((M)AFs) were calculated in a bigger sample comprising of KKNMS patients and six other samples of MS patients, all recruited across Germany (Andlauer et al., 2016). *MA = MS-associated variant and it's minor allele; Gene = (the associated) locus gene name; OR = odds ratio for developing MS if an individual carries the MS risk variant; (M)AF = (minor) allele frequency; Carriers = patients from the KKNMS cohort sample (314 individuals) with dosage data > 0.5; Noncarriers = patients from the KKNMS cohort sample (314 individuals) with dosage data <= 0.5.*

Genomic variation	Gene	OR	(M)AF	Carriers	Noncarriers
HLA allele					
HLA-A*02:01	*HLA-A*	0.68	28.6	107	207
HLA-DPB1*03:01	*HLA-DPB1*	1.33	10.3	74	240
HLA-DRB1*03:01	*HLA-DRB1*	1.29	12.2	77	237
HLA-DRB1*15:01	*HLA-DRB1*	2.85	14.8	158	156
Variants outside of MHC region (MA)					
rs10797431 (T)	*MMEL1*	0.85	34.0	155	159
rs1800693 (C)	*TNFRSF1A*	1.17	42.2	225	89
rs1891621 (G)	intergenic	0.87	46.4	219	95
rs2182410 (T)	*IL2RA*	0.83	38.1	161	153
rs2300747 (G)	*CD58*	0.75	12.3	55	259
rs2681424 (C)	*CD86*	0.86	49.7	220	94
rs2812197 (T)	*DLEU1*	0.85	38.3	179	135
rs2836425 (T)	*ERG*	1.25	12.4	92	222
rs34286592 (T)	*MAZ*	1.22	14.2	97	217
rs4364506 (A)	*L3MBTL3*	0.86	26.1	150	164
rs4925166 (T)	*SHMT1*	0.86	34.4	166	148
rs6498168 (T)	*CLEC16A*	1.22	35.4	187	127
rs6689470 (A)	*EVI5*	1.22	14.3	106	208
rs6859219 (A)	*ANKRD55*	0.85	22.1	113	201
rs7535818 (G)	*RGS1*	0.76	19.0	84	230

3.4.2.2 *Calculation of residual sum of squares*

By adjusting for the known sources of variance in our data, we are making sure that data stratification or differences we may find in our analysis are not due to factors we were able to adjust for. That is, we do the best we can to enable that the potential results mirror the contrasts we are interested in exploring, in this case, the different genetic background. One way of adjusting for known sources of variance is by fitting a linear model and then using the resulting residuals for the further data analysis. Residuals, that is, the residual sum of squares, represent the variation left unexplained after removing the sources of known variation *via* regression (James, Witten, Hastie, & Tibshirani, 2000). Following Chapter 3, Section 3.1.1. in the book *An Introduction to Statistical Learning* (James et al., 2000), a multivariate linear model was applied

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{3.20}$$

where Y is a dependent variable (or a response), β_0 is the intercept and β_j is the slope coefficient of the independent variable (covariate, predictor) X_j, and ϵ is the random error term. We interpret β_j as the average effect on Y of a one unit increase in X_j, holding other predictors fixed. The coefficients are estimated by minimizing the residual sum of squares, that is, the difference between observed values y_i and fitted values $\hat{y}_i$

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i). \tag{3.21}$$

In the KKNMS dataset, the expression level for each of the 13,442 probes was corrected with a set of twenty-seven covariates. Ten variables were dummy coded (0/1) and they referred to whether the patient was taking a DMT at the time of the blood draw (glatiramer acetate or interferon beta (IFN)), cortisol, budesonide, mesalazin, analgesics, vitamin-D or anti-histamines. It was further adjusted for an active infection and a current relapse as well (Table 3.2). Next, it was adjusted for the population structure by including eight components from the multi-dimensional scaling (MDS) analysis, as well as for the cell proportions in the gene expression data. Cell proportions of seven cell types were estimated using the *CellCode* package (Chikina, Zaslavsky, & Sealfon, 2015). The cell-type specific variation in gene expression was estimated for neutrophils, T-cells, monocytes, B-cells, natural killer (NK) cells, megakaryocytes, and erythrocytes. Lastly, adjustment for the age and gender differences was performed. As described above, twenty-seven variables were added as covariates in the multivariate linear model to explain the gene expression for each transcript in the data set separately. Model diagnostics was examined in the following way. First, principal component analysis (PCA) was applied on *log2* transformed gene expression data to reduce the dimensionality of the data. The first four principal components (PCs) explained roughly 39% of the overall variance in the data and they were used as representatives of the variation of data to test whether the model is stable. Hence, four models were fitted for each of the four selected PCs. By inspecting the residuals plot, Q-Q plot, variance inflation factor (vif), and the squared vif it was concluded that the data meets regression assumptions well with randomly distributed error and that there are no influential cases (outliers) in the dataset. The analysis of the variance inflation in variables was employed *via* the *car* package (Fox & Weisberg, 2020) and it showed no collinearity among covariates. The model with preselected covariates was fitted for each transcript separately. The residual sum of squares was obtained from each regression and used for further analyses.

3.4.2.3 Subsetting transcripts in WGCNA and GGM analysis

In both WGCNA and GGM analysis, transcripts were subsetted depending on the interest and type of the analysis. According to the authors of the WGCNA method, it is advisable to create a meaningful subset of genes in your data set to perform the network analysis. That is, running the analysis on all the genes from microarray could impede the analysis because of too much noise (Langfelder & Horvath, 2017).Therefore, it was decided to focus on immune system related genes, knowing that MS is mainly an immune-system driven disorder (IMSGC & WTCCC, 2011; International Multiple Sclerosis Genetics Consortium (IMSGC) et al., 2013). The AmiGO online tool (Carbon et al., 2009) was used to extract a list of all members of the Gene Ontology (GO) term “immune system process”(Ashburner et al., 2000; The Gene Ontology Consortium, 2020). Next, the *gene2go* list was used to extract all GO terms associated with each gene name in the data set. The *gene2go* list is a part of The National Center for Biotechnology Information (NCBI) database. By using the two lists, it was possible to subset transcripts in the KKNMS data set, therefore creating a list of immune system related transcripts. The list contains 2171 transcripts uniquely mapping to Entrez IDs. The WGCNA was therefore used to examine the influence of the

19 MS-associated variants on gene-gene interactions between 2171 immune system related transcripts.

On the other hand, the GGM analysis was applied in a pathway-based manner. The reasoning was the following. Inference of a big network where the number of variables (13,442 transcripts) exceeds the number of individuals (anywhere between 55 and 240, Table 3.3) would potentially lead to a poor association estimate due to a low power and therefore, high probability of committing a type II error (high number of false negatives). Hence, it was decided that the transcripts should be split into meaningful subunits. The data were first preprocessed in a way to only contain transcripts uniquely mapping to Entrez IDs, which left 13,442 transcripts in the data set. Afterwards, the KEGGREST package in *R* was used to divide genes into pathways according to the KEGG database (version August 2020, Kanehisa, Furumichi, Sato, Ishiguro-Watanabe, & Tanabe, 2020). Pathways containing less than ten genes were removed from the analysis because getting a stable estimate of GGM on such a small matrix was not possible. The subsetting resulted in 307 KEGG pathways, covering 5370 transcripts uniquely mapping to Entrez IDs. The smallest pathway in the data set contained 10 genes, whereas the biggest pathway consisted of 386 genes. To explore the extent of gene overlap between pathways, for each pathway the average number of genes it shares with the remaining 306 pathways was calculated. The average number of shared genes was further divided by pathway size to get a more standardized degree of overlap which does not depend on the size of the pathway (Figure 3.2).

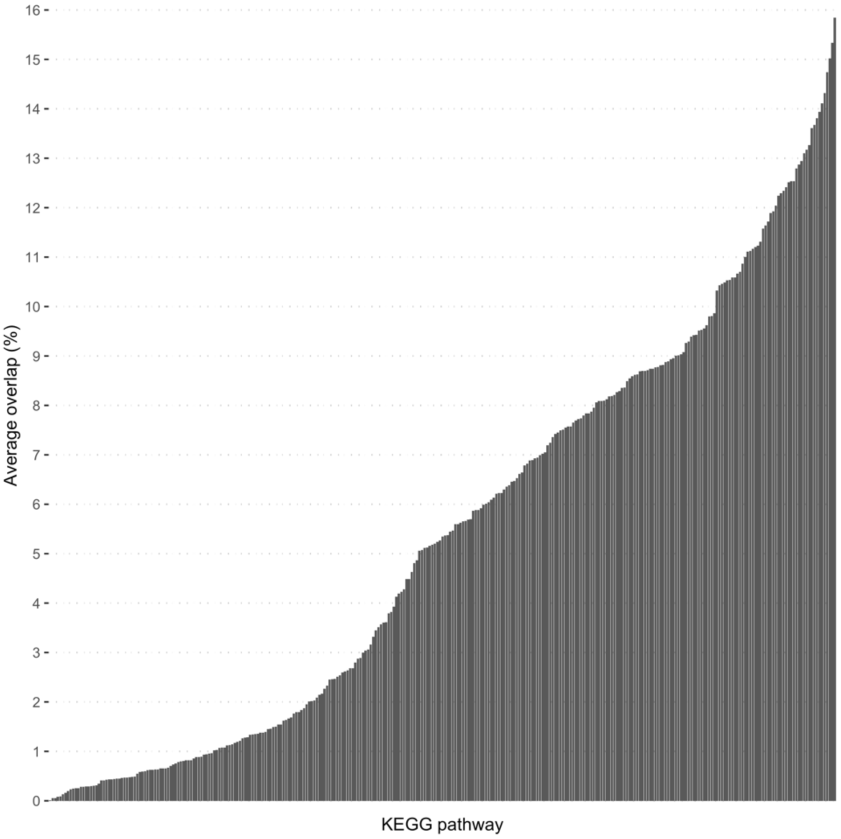

Figure 3.2 ***The average overlap in gene names for 307 KEGG pathways covering 5370 unique Entrez IDs in the KKNMS data set.***

Each bar in the plot represents one of the 307 analyzed pathways. The height of the bar indicates the percentage of how many genes on average is shared between the pathway and the rest of the pathways.

3.4.2.4 Grouped Benjamini-Hochberg (GBH) approach to control for the false discovery rate (FDR) in the data with group structure

When performing the cis-eQTL analysis and the enrichment tests for genes in modules to find associated GO terms, the grouped Benjamini-Hochberg (GBH) approach (Hu, Zhao, & Zhou, 2010) was used to control and adjust the FDR. The GBH approach enables including prior information of natural group structure among hypotheses, present when, for example, performing eQTL analysis for several SNPs in parallel. It has been previously discussed in the research paper by Efron (Efron, 2008) that ignoring the group structure naturally present in the data during the correction for the multiple hypothesis testing can lead to “overly conservative or overly liberal conclusions” (Hu et al., 2010).

Following the paper by Hu and the colleagues (Hu et al., 2010), let I_g be the index set of the gth group and I_N index set of all the hypotheses which satisfies

$$I_N = \bigcup_{g=1}^{K} I_g = \bigcup_{g=1}^{K} \left(I_{g,0} \bigcup I_{g,1}\right) \tag{3.22}$$

where $I_{g,0} = \{i \in I_g : H_i \, is \, true\}$ and $I_{g,1} = \{i \in I_g : H_i \, is \, false\}$, with $n_{g,0} = |I_{g,0}|$ and $n_{g,1} = n_g - n_{g,0}$. It further follows that the proportion of null hypotheses in the group is defined as $\pi_{g,0} = \frac{n_{g,0}}{n_g}$ and the proportion of discoveries as $\pi_{g,1} = \frac{n_{g,1}}{n_g}$. The overall proportion of null hypothesis, π_0, is then defined as

$$\pi_0 = \frac{1}{N} \sum_{g=1}^{K} n_g \, \pi_{g,0}. \tag{3.23}$$

In the next paragraphs, a so-called “oracle case” is introduced, where the proportion of null hypotheses in a group is bound between 0 and 1, that is, $\pi_{0,g} \in [0, 1]$.

The GBH procedure for the oracle case is as follows.

1. For each *p*-value in a group *g*, weighted *p*-values $P_{g,i}^{w}$ are calculated following

   $$P_{g,i}^{w} = \frac{\pi_{g,0}}{\pi_{g,1}} \times P_{g,i} \tag{3.24}$$

   where $\pi_{1,g}$ is the proportion of discoveries in the group, and $\pi_{1,g} = 1 - \pi_{0,g}$. If $\pi_{0,g} = 0$, then $P_{g,i}^{w} = \infty$. If $\pi_{0,g} = 0$ for all the hypotheses, stop. Otherwise, proceed with the next step.

2. Take all the weighted *p*-values across all tests *N,* and order them in ascending order so that the first *p*-value in the list is the lowest one, $P_{(1)}^{w} \leq \cdots \leq P_{(N)}^{w}$.

3. Finally, get the number of null hypotheses to be rejected

$$k = \max\left\{i : P_{(i)}^{w} \leq \frac{i \times \alpha^{w}}{N}\right\}. \tag{3.25}$$

Here, k is the maximum over i for which this inequality is true, and $\alpha^{w} = \frac{\alpha}{1-\pi_0}$.
The proportion of null hypotheses in the group $\pi_{g,0}$ will be estimated by employing the two-stage (TST) method (Benjamini, Krieger, & Yekutieli, 2006), which is an adaptive BH procedure enabling finite-sample FDR control for independent p-values. First, the BH procedure at level $\alpha' = \frac{\alpha}{1+\alpha}$ is applied on all p-values within each group. The number of rejections $r_{g,1}$ is thereby estimated. The TST estimator γ_g^{TST} of $\pi_{g,0}$ is then computed by calculating the following ratio

$$\gamma_g^{TST} = \frac{n_g - r_{g,1}}{n_g}. \tag{3.26}$$

The GBH approach is then employed at level α' with $\pi_{g,0}$ replaced by γ_g^{TST}, to estimate the k (Hu et al., 2010).

3.4.2.5 Differential expression analysis

Gene expression levels between the carriers and noncarriers were compared for each of the 19 MS-associated variants separately. Important clinical differences, the population structure, gender, age and cell composition were accounted for using the linear model where *log2* expression values for each of the 19420 probes were corrected and residuals were used to test for differentially expressed genes. The scope of gene expression difference in noncarriers in comparison to the carriers of the respective variant was explored using the *limma* package (Smyth, Michaud, & Scott, 2005). The *p* values resulting from group comparisons were adjusted using the Holm method (Holm, 1979). The Bonferroni method (Dunn, 1961) was employed to account for the number of independent tests, therefore, the 5%-significance level was adjusted to 0.05/19 = 0.00263.

3.4.2.6 eQTL analysis in the KKNMS data set

The eQTL analysis enabled exploring potential functional associations between SNPs and genes. Among 19 MS-associated variants, four allelic variations inside the MHC region were excluded due to a complex linkage disequilibrium (LD) structure and high SNP frequency (Lam, Shen, Tay, & Ren, 2017). For the remaining 15 SNPs, *cis*-eQTL analysis was run in the following manner. First, a subset of probes in close proximity to each SNP was selected. Probes mapping to genes 500,000 base pairs (bp) up- and down-stream from the respected variant were tested for association. The *log2* transformed gene expression data was used to test for associations. The linear model included important clinical differences, the population structure, gender, age and cell composition, as well as the SNP dosage data of the tested SNP. Each transcript was tested separately. The rate of false discoveries was controlled by applying the GBH and TST approaches explained in the earlier paragraphs. The approach yielded weighted p values adjusted for the proportion of null hypotheses in the SNP ($\pi_{g,0}$), for which the corresponding thresholds were estimated, based on the weighted p value index i and the adjusted alpha significance level α^{w}.

3.4.2.7 Differences in Pearson's correlation patterns

Groups of interest were also compared by examining the differences in Pearson's correlation between gene expression levels. Pearson's correlation is a measure of linear association between variables. The correlation pattern between genes can be subject to change, even if the expression level of a gene does not significantly differ between the two states. This could be the result of a mutation in the coding region of a gene, which affects a certain gene's function without altering its expression level (de la Fuente, 2010). Therefore, by contrasting gene correlation matrices between two states of interest, such distortions could potentially be found. When comparing gene expression correlation between two groups of interest, Pearson's correlation matrix was calculated, where each correlation coefficient r_{xy} represents Pearson's correlation between the expression levels of two probes, x and y. However, Pearson's correlation coefficients are defined on the [-1,1] interval, and the sample distribution of highly correlated values is skewed, with variance and skewness dependent on the value of underlying correlation in the population (ρ). Following Myers and Sirois (Myers & Sirois, 2014) the distribution of the correlation coefficients can be approximately normalized by applying the Fisher's *Z*-transformation (Fisher, 1921)

$$Z_r = \frac{1}{2} \times \ln\left(\frac{1+r}{1-r}\right). \tag{3.27}$$

The standard error is defined as

$$SE = \frac{1}{\sqrt{N-3}} \tag{3.28}$$

which is independent of the correlation coefficient. Using the following formula to calculate the test statistic

$$z = \frac{Z_{rx} - Z_{ry}}{\sqrt{\frac{1}{(N_x - 3) + (N_y - 3)}}} \tag{3.29}$$

one can easily compute a matrix of differences in normalized Pearson's correlation coefficients and obtain a *p* value to infer the significance of the observed difference.

Pearson's correlation matrices were computed using the *cor* function from the *WGCNA* package which enables fast correlation matrix computing (Langfelder & Horvath, 2012). The correlation was calculated on the data corrected for known sources of variance, *i.e.*, the residual sum of squares. The variant-specific noncarriers correlation matrix $\mathrm{NC}_{\mathrm{variant}}$ was subtracted from the carriers correlation matrix $\mathrm{C}_{\mathrm{variant}}$ to infer the variant-specific difference matrix (Eq. (3.30)).

$$NC_{variant} - C_{variant} = \begin{pmatrix} nc_{11} - c_{11} & \cdots & nc_{1j} - c_{1j} \\ \vdots & \ddots & \vdots \\ nc_{i1} - c_{i1} & \cdots & nc_{ij} - c_{ij} \end{pmatrix} \tag{3.30}$$

Fisher's *Z*-transformation was implemented in *R* to obtain the matrix of *z*-scores, that is, the normalized correlation differences for each of the 19 MS-associated SNPs. The corresponding two-sided *p* values were computed using the *pnorm* function. The resulting *p* values were corrected on two levels. Firstly, the Holm method was used to account for the number of tests executed on the SNP level, with *p.adjust* function in *R*. Secondly, the Bonferroni method was employed to adjust the significance for the number of MS-associated variants, aiming for a significance level of 5% after correction.

3.4.2.8 Gene co-expression analysis in the KKNMS data set

Gene co-expression networks in the KKNMS data set were built using two methods: the WGCNA, which builds weighted gene co-expression networks using Pearson's correlation, and the GGM, a method to estimate sparse covariance matrix to unveil the direct connections between genes. The former approach was data driven, that is, genes were grouped in modules based on their correlation patterns. The GGM approach was applied on genes which were previously grouped in pathways based on KEGG database (Kanehisa et al., 2020).

3.4.2.8.1 Estimating the gene co-expression modules *via* the WGCNA

Genetic variation underlying the MS etiology probably affects multiple genes that share the similar functionality or are regulated in a similar way. The weighted gene co-expression network analysis (WGCNA) was used to find clusters of highly connected genes in a data-driven fashion (Langfelder & Horvath, 2008). The analysis was run using the WGCNA (v1.66) package in *R.* The WGCNA method uses Pearson's correlation to distinguish blocks of genes that show a significantly higher correlation pattern than the other genes, usually uncovering groups that share a similar biological function. The blocks of highly correlated genes are called modules. All genes which did not show strong connectivity patterns were considered a part of an unassigned group of genes - the grey module. Considering prior information on biological pathways underlying the MS etiology, the gene expression data was subsetted to contain 2171 immune system related genes. We tested for a specific module-genotype association in two ways. First, a data-driven approach was applied where gene expression of all 314 patients and 2171 immune-related genes was used to construct highly connected modules. The expression variation of genes from a certain module was summarized by their first principal component, the module eigenvector. The module eigenvectors were tested for association with 19 MS-associated variants, potentially revealing the biological subunits affected by a specific variant. In the second approach, group-specific modules were built (MS-associated variant carriers and MS-associated variant non-carriers) and module preservation was compared depending on the carrier status. The analysis was run for each variant separately. The statistic summarizing the preservation level is the $Z_{summary}$ statistic (Langfelder et al., 2011). We were testing against the null hypothesis that there is no difference in the structure of the tested module between the reference and the test group. The bigger the Z statistic, the more preserved the module from a reference set is, when compared to the test set. For a Z value bigger than 10, we say that there is strong evidence for module structure preservation. For a Z value between 2 and 10, there is medium to weak evidence of the preservation of the module structure. Lastly, a Z value lower than 2 characterizes a module with very weak to no evidence of preservation, suggesting that there are topological differences between the groups.

Furthermore, to examine whether genes in the modules exert any biological function, modules were biologically annotated using the GOstats package in R (v2.48.0), which performs the overrepresentation analysis (ORA). Genes were submitted as unique Entrez IDs, and 2171 Entrez IDs were used as gene universe for the analysis. The hypergeometric test was applied to test for the significance of the hits. The rate of false discoveries was controlled by applying the GBH and TST approaches explained earlier. The approach yielded weighted p values adjusted for the proportion of null hypotheses in the group $\pi_{g,0}$, for which corresponding thresholds are estimated, based on the weighted p value index i and adjusted alpha significance level α^w.

3.4.2.8.2 Estimating pathway-specific sparse graphs with GGM

The Gaussian graphical models (GGM) (Schäfer & Strimmer, 2005b) approach was employed to estimate partial correlations between gene expression levels in the KKNMS dataset. It is a specific graph-based method that explores conditional independency between each pair of genes in a dataset, *i.e.*, it distinguishes direct interactions between genes from indirect ones. The *ggm.estimate.pcor* function from the GeneNet R package (v1.2.13) with the method *static* was applied to estimate the partial correlations between the transcripts. The method estimated a gene expression data-based graph, in which each node represented a transcript and each edge a direct association between the genes. First, genes were divided into pathways based on the KEGG data base (version August 2020, Kanehisa et al., 2020) in the following way. From the KKNMS data set, 5400 unique Entrez IDs were matched in the KEGG, spanning 326 pathways. Pathway sizes ranged from 1 to 386. It was decided to analyze pathways with at least 10 genes according to the previous similar study (Grimes et al., 2019) and given the fact that GGM reported errors with pathways having less than 7 genes. In the end, 307 pathways were selected, with sizes ranging from 10 to 386, and covering 5370 unique Entrez IDs. This enabled direct pin-pointing to an annotated biological subunit that was subject to change due to genetic variance, if such change existed. For each pathway, the GGM was estimated and then compared between the two groups, the MS-associated variant carriers and the MS-associated variant non-carriers, for each of the 19 analyzed variants separately. Differences in partial correlations in the same pathway between the two groups were inspected using two approaches, the *sum* and the *max*. To infer the global differences in gene expression correlation between the groups on a pathway level, all correlation coefficients for a pathway within the group were summed up. Then, the absolute difference in correlations between the two groups was obtained, yielding an empirical global measure of difference (the *sum*). On the other hand, examination of edge-specific effects was conducted by finding the biggest absolute difference in partial correlations between the groups in each pathway as a representative of pathway-level discrepancy, resulting in an empirical edge-specific measure of difference (the *max*).

To assess the significance of both empirical measures, namely the *sum* and the *max*, permutation tests were run in a way that genotype information was first shuffled 1000 times for each variant, and the two empirical measures were calculated for each pathway in each of the 1000 runs. Let v be the vector with elements representing the pathway-specific empirical measures calculated after shuffling the group labels of one variant for n times

$$v = \{v_1, v_2, \dots, v_n\} \tag{3.31}$$

where n stands for the number of permutations. For each element of v, we check if the element is bigger or equal than the original empirical measure calculated earlier for that pathway

$$I_{v_i} = \begin{cases} 1, v_i \geq q \\ 0, v_i < q \end{cases}, \qquad i = 1, \ldots, n \tag{3.32}$$

where I is the indicator function defined on the set v. The elements of I_{v_i} are summed and divided through by the number of permutations n

$$permutation \text{ p } value = \frac{\sum_{i=1}^{n} I_{v_i}}{n} \tag{3.33}$$

resulting in the permutation p value which represents the average number of permuted measures bigger or equal than the original measure. To account for multiple hypothesis testing on a pathway level, p values were corrected using the Holm method (Holm, 1979). The nominal significance of 5% was adjusted due to 19 simultaneously performed independent tests, *i.e.*, the number of tested SNPs. The Holm adjusted p values below a 0.00263 threshold were considered significant. For pathways showing statistical significance after 1000 permutations, genotypes were additionally shuffled 100,000 times, yielding a more precise posterior distribution of the two empirical measures of the network differentiality. Permutation p values from 100,000 permutations were again first corrected using the Holm method, with *p.adjust* function and the number of comparisons set to 307. Accordingly, the Bonferroni method was employed to adjust for the number of tested variants (19).

3.4.2.8.3 Difference in node connectivity between two pathway-specific networks

In order to investigate the connectivity patterns in a pathway-based network, Gaussian graphical models estimated in the previous analysis step were further analyzed. However, when examining connectivity, a focus was shifted from all the edges in a pathway, and towards direct edges only. The Bayes factor was used to select direct edges in a pathway-specific GGM network. For each edge, the Bayes factor was calculated based on the posterior probability that an edge is direct, that is, probability that the partial correlation between the two genes is different than zero (Eq. (3.16)). Direct edges were selected by applying a threshold on each edge-specific Bayes factor. Based on the Table 3.1, the targeted level of evidence was selected to be the geometric mean of the *substantial evidence against the M$_0$* category. Each edge with a Bayes factor equal to or bigger than 5.6 was considered direct. Once the edge-specific Bayes factor was calculated, nodes with direct connections were selected. Based on previous research on gene connectivity (Yang et al., 2014; J. Zhang et al., 2016), the top 5% of nodes with the highest connectivity in all pathways were further analyzed .

Therefore, by following the common guidelines on the Bayes factor cut off, it was possible to select direct edges and analyze node connectivity profiles in the data. For each edge in a pathway, g - 1 different Bayes factors were calculated, g being the number of genes in that pathway in the KKNMS data set. The Bayes factor was calculated using data from patients carrying the variant and patients not carrying the variant separately. Hence, two different Bayes factors were joint to each edge in the pathway for each variant. In the analysis of this setting, where groups of variant carriers and noncarriers are contrasted, there are altogether

38 groups among 19 MS-associated variants, consisting of different individuals combinations, based on their genotype. The connectivity was further calculated with two referencing directions: by selecting direct edges in the carrier group of a pathway and comparing their connectivity in the noncarrier group and by selecting direct edges in the noncarrier group of a pathway and comparing the connectivity of those nodes to their connectivity in the carrier group. To calculate the difference in connectivity (DC), the connectivity of a node in the reference group was subtracted from a connectivity of a node in the test group. The permutation analysis was run to infer the statistical significance of each gene's DC in the following way. All nodes with three or more direct edges were analyzed on a 1000 permutation level. Nodes which passed the multiple hypothesis threshold after 1000 permutations were analyzed on a 10,000 permutations level. Only genes which passed the 10,000 permutation multiple hypothesis testing threshold were analyzed on a 100,000 permutations level. The intermediate step of running 10,000 permutations prior to running the 100,000 permutations was implemented to combat memory issues and computational time (Figure 3.3). The assumption was that the number of pathways which will need 100,000 permutations run will be cut down after the 10,000 permutations.

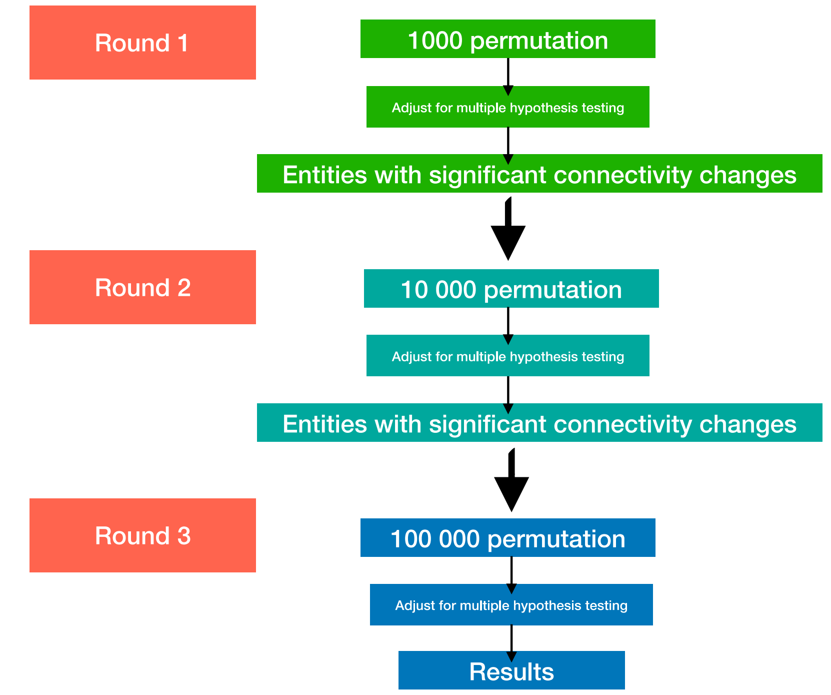
Round 1
1000 permutation
Adjust for multiple hypothesis testing
Entities with significant connectivity changes
Round 2
10 000 permutation
Adjust for multiple hypothesis testing
Entities with significant connectivity changes
Round 3
100 000 permutation
Adjust for multiple hypothesis testing
Results

Figure 3.3 ***Differential connectivity analysis workflow****.*
Entity is a node in a pathway-specific GGM network of one group (carriers or noncarriers of a variant). The workflow is created with Keynote application.

Each permutation run resulted in a permutation *p* value for each edge. Permutation *p* value was calculated by dividing the number of times a shuffled genotype group resulted in a DC bigger or equal to the true DC. This number was then divided by the number of permutations (the number of times a genotype was shuffled). To account for the multiple testing, the Holm method for multiple testing adjustment (Holm, 1979) was used to account for the number of pathways tested within the same group, then the Bonferroni method (Dunn, 1961) was applied to account for the number of tested MS-associated variants (19).

3.4.3 Graphical visualizations

Figures in the thesis were created using the *ggplot* package in R (Wickham, 2016), if not stated otherwise below the figures.

3.5 Results

3.5.1 Data preprocessing: removing known sources of variation

As explained in the Methods section (3.4.2.2 Calculation of residual sum of squares), the gene expression data was corrected *via* the multivariate linear regression and the resulting residual sum of squares (estimated transcript expression values versus the observed values) was used in the further analyses. To inspect whether twenty-seven known sources of variation were successfully regressed out, the PCA was first run on the *log2* transformed expression values and tested for association with 27 variables. Next, the PCA was applied on the residuals and the association between the residual sums of square and the 27 variables was tested again. Figure 3.4 shows correlation patterns before (left) and after (right) adjusting for covariates.

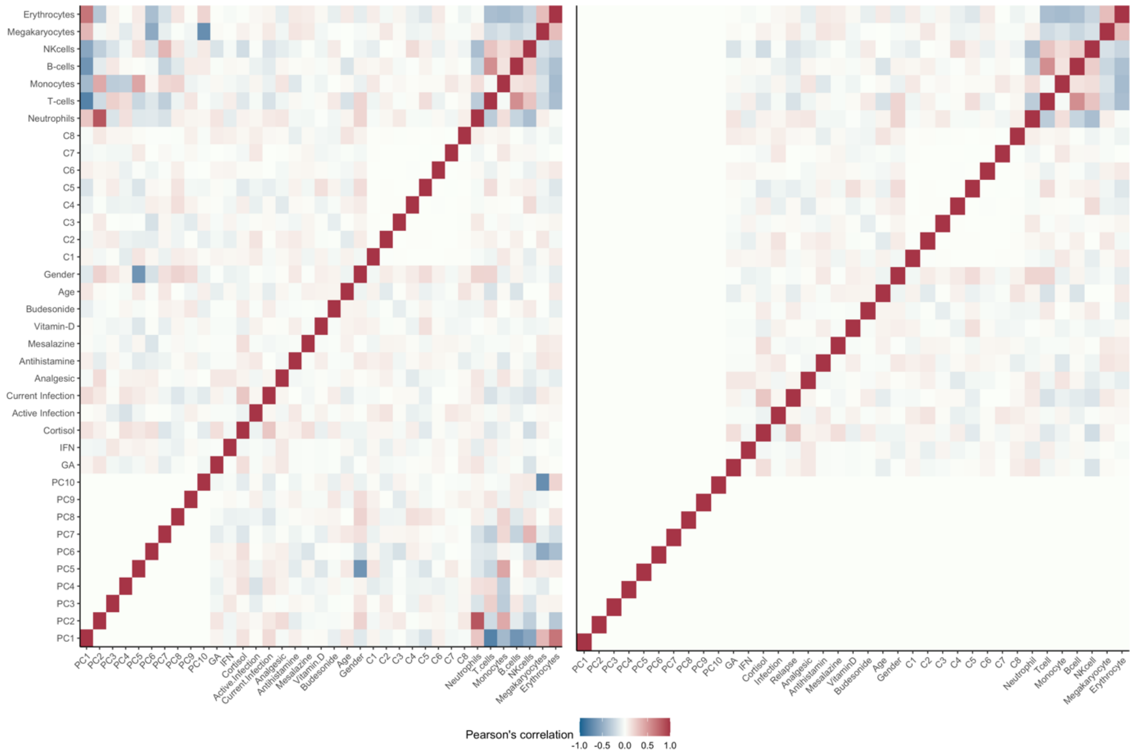

Figure 3.4 ***Correlation pattern between the first ten PCs of the gene expression data and twenty-seven covariates, before (left) and after regression (right).*** *Darker colors indicate stronger correlation coefficients. The plots were created using the* corrplot *package in R* (Wei & Simko, 2021).

3.5.2 Differential gene expression analysis reveals subtle differences between groups of variants

Gene expression profiles between groups of MS patients differing in their genetic background were compared. Moderated *t*-tests were applied to examine how many of 19,420 tested transcripts show differences in gene expression among 19 contrasts. Contrasts were designed to test the gene expression change relative to the carriers group.

Table 3.4 ***Differentially expressed genes among 19 analyzed contrasts.***
Six contrasts for which differentially expressed genes have been detected after adjusting for the multiple testing are presented. The direction of the logFC is relative to the allele carriers, that is, to the minor allele for SNPs outside of the MHC region. Adj. p *value is calculated using the Holm method. Genes with adjusted p-value below the adjusted Bonferroni threshold (0.00263) are considered significantly differentially expressed for the given contrast. Log(FC) = logarithm of fold change, 95% CI = 95% confidence interval for the log(FC), AveExpr = average expression of the gene across all samples, t = corresponding t value from the moderated t-test, B = B-statistic, log odds that the gene is differentially expressed.*

Variant	IlluminaID	Gene	*p* value	adj. *p* value	log(FC)	95% CI	AveExpr	t	B
HLA-DRB1*15:01	ILMN_1697499	HLA-DRB1	8.4×10^{-123}	1.62×10^{-118}	-6.6	-6.9 - -6.3	8.12	-38.79	211.90
HLA-DRB1*15:01	ILMN_1715169	HLA-DRB1	1.9×10^{-43}	1.83×10^{-39}	-3.5	-3.9 - -3.1	8.60	-16.19	78.97
rs4925166	ILMN_1811933	SHMT1	9.1×10^{-19}	1.77×10^{-14}	0.52	0.41 - 0.63	6.14	9.42	30.61
HLA-DRB1*03:01	ILMN_1808405	HLA-DQA1	5.3×10^{-10}	0.00001	0.64	0.44 - 0.83	9.94	6.41	12.64
rs10797431	ILMN_1718488	MMEL1	1.0×10^{-8}	0.00020	0.26	0.17 - 0.34	4.32	5.88	10.58
HLA-A*02:01	ILMN_2203950	HLA-A	1.5×10^{-8}	0.00028	-0.11	-0.15 - -0.08	14.48	-5.82	10.16
HLA-DRB1*03:01	ILMN_3249667	HLA-DQA1	1.7×10^{-7}	0.00160	0.59	0.37 - 0.80	10.40	5.35	7.87
HLA-DRB1*15:01	ILMN_2157441	HLA-DRA	3.6×10^{-7}	0.00232	-0.18	-0.25 - -0.11	13.58	-5.20	7.55

For four alleles in the MHC region it was observed that the allelic RNA expression levels were influenced by its variant (Table 3.4, Figure 3.5). A strong log fold change (log(FC)) was observed for the HLA-DRB1 locus, where HLA-DRB1*15:01 allele carriers were found to have markedly higher expression of two probes mapping to HLA-DRB1 gene (ILMN_1697499 log(FC)=-6.6, adj. *p* value 1.63 × 10^{-118}; ILMN_1715169 log(FC)=-3.5, adj. p value 1.83 × 10^{-39}). The gene expression of the *HLA-DQA* gene was found to be significantly different between groups of the HLA-DRB1*03:01 carriers and noncarriers (ILMN_1808405 log(FC)=0.64, adj. *p* value 1.02 × 10^{-5}; ILMN_3249667 log(FC)=0.59, adj. *p* value 1.60 × 10^{-3}). MS patients carrying the HLA-DRB1*15:01 variant showed significantly higher HLA-DRA gene expression levels (log(FC)=-0.18, adj. *p* value 2.32 × 10^{-3}) than the patients not carrying the respective variant (Figure 3.5). Similarly, it was observed that the carriers of the HLA-A*02:01 allele showed significantly higher gene expression for the HLA-A locus (log(FC)=-0.11, adj. *p* value 2.83 × 10^{-4}).

Furthermore, there were two variants outside of the MHC region for which we found potential influence on the gene expression profiles (Table 3.4, Figure 3.5). The gene expression of the SHMT1 locus was higher in patients not carrying the rs4925166 variant (log(FC)=0.52, adj. *p* value 1.77 × 10^{-14}). The same effect was observed for the association between the *MMEL1* gene and the rs1079743 variant (log(FC)=0.26, adj. *p* value 2.01 × 10^{-4}).

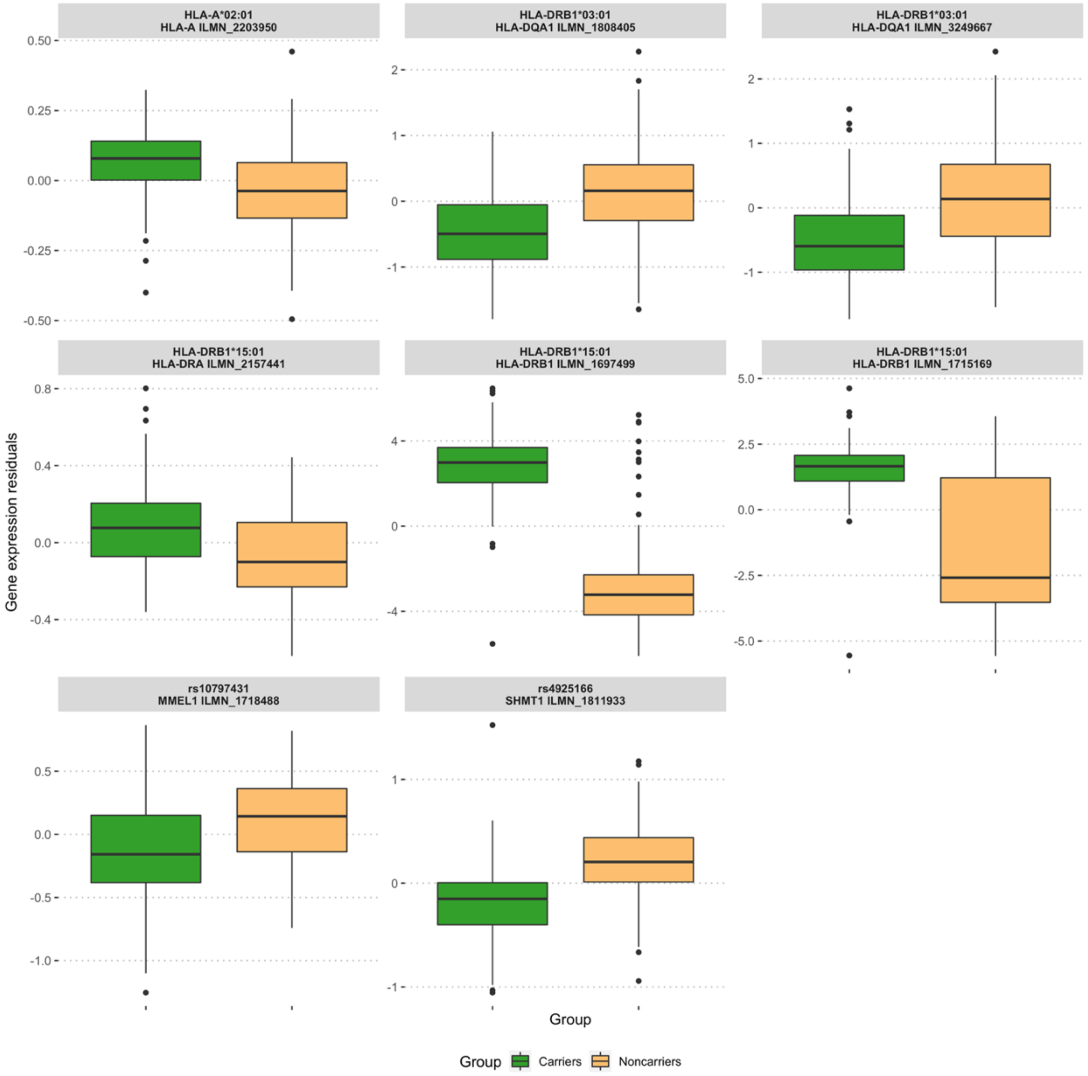

Figure 3.5 ***Gene expression of probes differentially expressed between groups of MS-associated variants.*** *The differences in gene expression levels of transcripts differentially expressed between variant groups, statistically significant after adjusting for multiple testing. The fold change differences were estimated* via *limma analyses (Table 3.4).*

3.5.3 eQTL analyses

Dosage data from 15 MS-associated variants outside of the MHC region was tested for linear associations with *log2* transformed gene expression of probes in a multivariate linear model. The *p* values from the regression were adjusted for the multiple hypothesis testing by applying the GBH procedure, with the TST approach employed to estimate the proportion of null hypotheses within the group (SNP). The results of eQTL analysis across 15 tested variants in the KKNMS sample are presented in the Table 3.5, and found associations represent already annotated eQTLs (Andlauer et al., 2016).

Table 3.5 ***Detected eQTL associations for SNPs outside of MHC region.***

Probes mapping to 500,000 bp up- and down-stream of a given SNP were included in the analysis. Association was considered significant if the w-p *value was smaller or equal to the threshold.* P *value =* p *value from the regression;* w-p *value = weighted* p *value corrected for the proportion of null hypotheses in a group, estimated from the GBH procedure; threshold –* p *value index-based threshold, estimated from GBH and TST procedure; index = order of the* p *value across all groups.*

Variant	IlluminaID	Gene	*p* value	*w-p* value	threshold	index
rs4925166	ILMN_1811933	*SHMT1*	8.035×10^{-27}	2.089×10^{-25}	0.00029	1
rs10797431	ILMN_1718488	*MMEL1*	2.286×10^{-10}	3.887×10^{-9}	0.00059	2
rs4364506	ILMN_1727495	*L3MBTL3*	4.377×10^{-7}	4.377×10^{-6}	0.00088	3
rs6859219	ILMN_2341724	*ANKRD55*	1.509×10^{-6}	1.207×10^{-5}	0.00118	4
rs1800693	ILMN_1685005	*TNFRSF1A*	5.563×10^{-6}	0.0001	0.00147	5
rs6859219	ILMN_1798947	*ANKRD55*	3.167×10^{-5}	0.0002	0.00177	6

3.5.4 *DHRS13 – UBOX5* gene pair showed a significant difference in Pearson’s correlation between groups of the rs2836425 variant

For each variant, two Pearson’s correlation matrices were calculated – one on the data from the variant noncarriers ($NC_{variant}$) and the other based on the data from the variant carriers ($C_{variant}$). Each matrix represented the correlation for the 90,336,961 different gene pair combinations of 13442 genes. Next, the Pearson’s correlation matrices were subtracted between groups of variants resulting in 19 matrices representing the degree of variant-related differences (following Eq. (3.30)). Differences were further standardized as explained in the Methods section, which enabled inferring the *p* values for each of the respective differences. After adjusting for the multiple hypothesis testing, one gene pair resulted in a significant difference in the correlation between groups of the rs2836425 variant. The rest of the differences did not sustain the multiple testing correction, steps of which were explained in the Methods section (3.4.2.7 Differences in Pearson’s correlation patterns). The measured difference in the correlation between the transcript mapping to *DHRS13* gene, coding for a dehydrogenase, and the transcript mapping to *UBOX5,* coding for proteins involved in ubiquitination pathway, was statistically significant. With $r_{noncarriers}= 0.388$ and $r_{carriers}= -0.435$, the measured difference *diff* = 0.822 was standardized *via* Fisher transformation, yielding the $Z = 6.961$ (*p*-value = 3.38×10^{-12}, adjusted *p* value = 3.05×10^{-4}). Next, all standardized differences for the rs2836425 variant were plotted (Figure 3.6). The difference of the DHRS13-UBOX5 gene pair was the strongest difference among all tested differences and is located in the tail of the distribution.

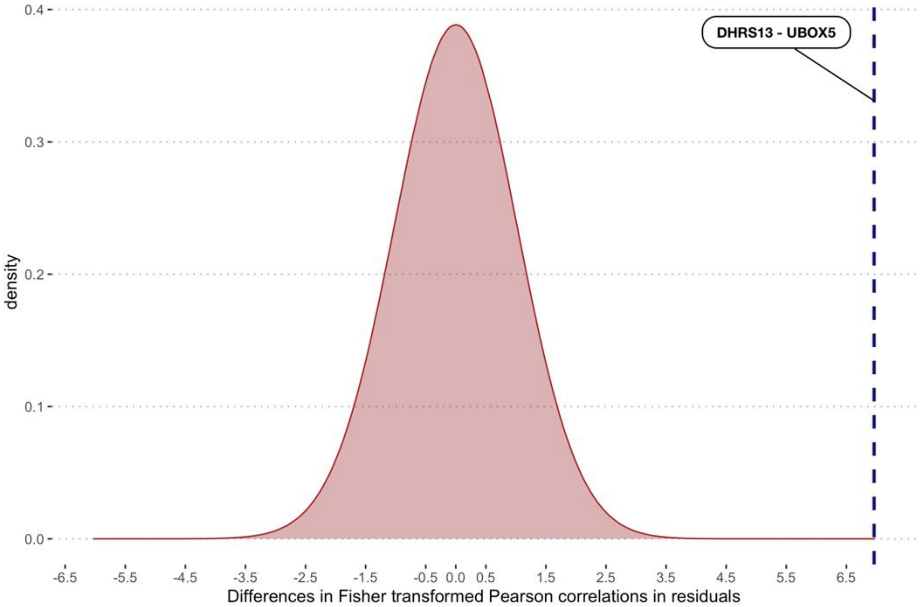

Figure 3.6 ***Distribution of standardized Pearson's correlation differences for all gene pairs between the patients carrying the rs2836425 variant and variant noncarriers.***
*The differences are calculated by comparing the correlations among 90336961 gene pairs between the rs2836425 variant noncarriers and carriers. The dashed vertical line represents the standardized difference in correlation for the gene pair DHRS13-UBOX5 (Z = 6.961). The difference was statistically significant after the multiple hypothesis testing (*p *value* 3.38×10^{-12}*, adjusted* p *value* 3.05×10^{-4}*).*

3.5.5 Module variance weakly explained by different genetic background

The WGCNA was applied on the residuals from the gene expression data of 2171 probes and 314 patients. Scale free topology criteria was satisfied with a soft threshold (power) of 3 (Figure 3.7). The determined power was used to estimate the topological overlap matrix (TOM) from the expression data.

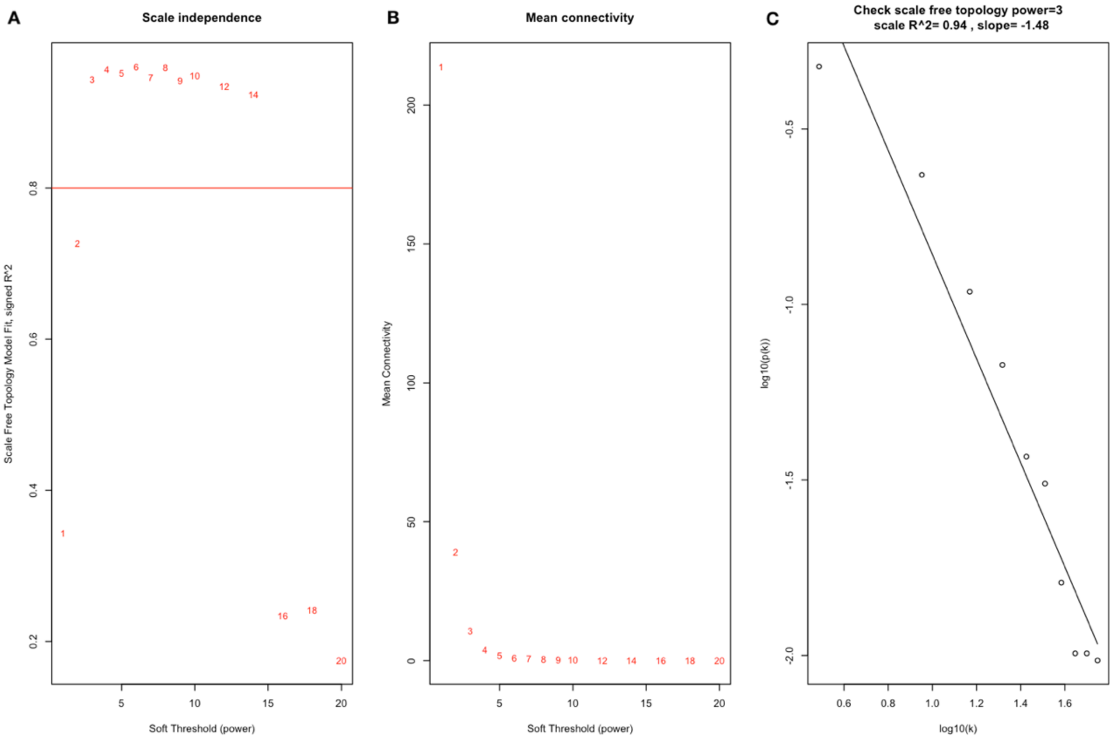

Figure 3.7 ***Determining the soft thresholding power satisfying the scale-free criteria.*** *(**A**) The analysis of scale independence for various soft thresholding powers (β). (**B**) Mean connectivity distribution depending on the power level. (**C**) Examining the linear relationship between log10(k) and log10(p(k)) for the selected power of 3. Plots were created with the* scaleFreePlot *function from the* WGCNA *package in R* (Langfelder & Horvath, 2008).

The hierarchical cluster analysis with average linkage criteria was performed on the TOM-based dissimilarity matrix. Next, the resulting dendrogram was cut at the lowest level (*dpSplit*=3), yielding eleven modules. Finally, the Pearson's correlation between modules was calculated and modules with r > 0.8 were merged. Most of the genes (1273/2171) were grouped into one of the ten resulting modules, while for the rest (898/2171) no significant association with other genes was found and they were labelled as grey (Figure 3.8).

Gene expression variance in modules was summarized with the PCA. For each module the first principal component represented the module eigengene (ME), *i.e.*, the direction of the variance in gene expression of a given module. The dosage data from the variants was correlated with the MEs to explore the influence of genetic risk factors on module gene expression variance. For each correlation coefficient, the *t* value was calculated enabling *p* value estimation based on the Student distribution. Correlation coefficients from all comparisons followed a normal distribution (Shapiro normality test, *W* = 0.993, *p* value = 0.463), with a mean of -0.0008, therefore providing negligible evidence for genotype-specific gene expression variation in the modules (Figure 3.9).

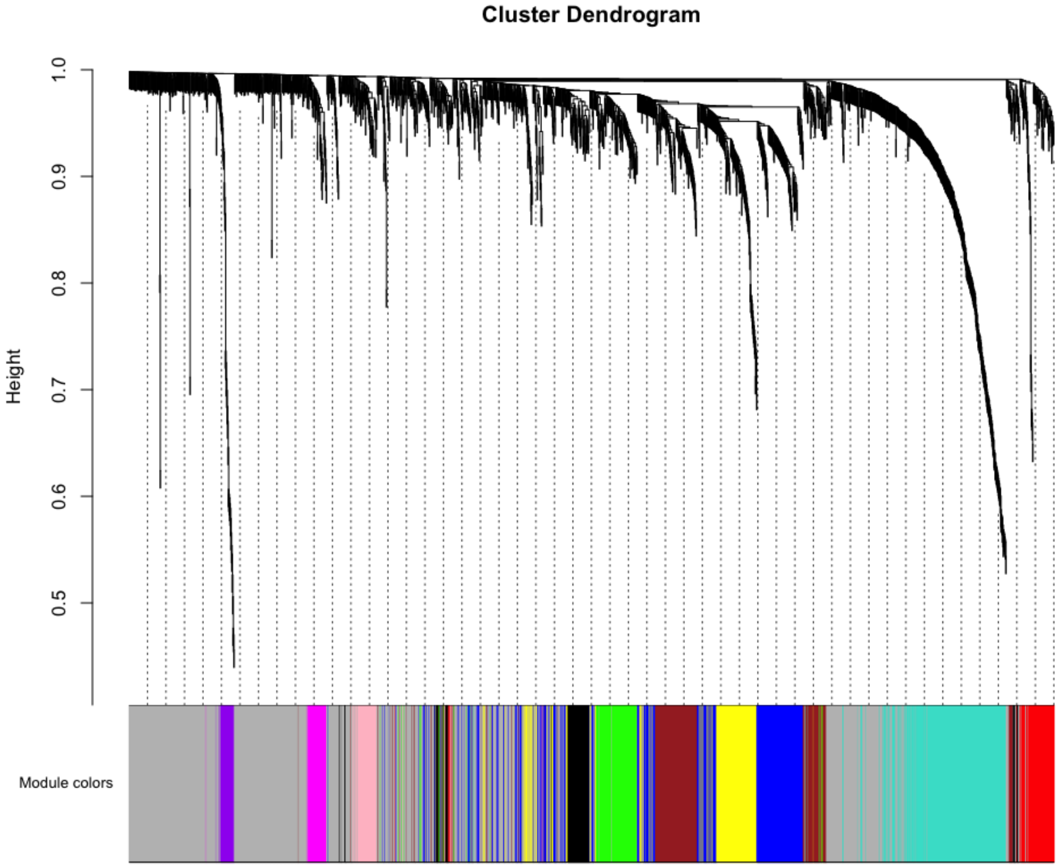

Figure 3.8 ***Dendrogram branches clustered into ten modules based on the TOM dissimilarity matrix from 314 MS patients.***
Each of the 2171 input genes has an associated color, corresponding to the module it was assigned to. Genes exerting low correlation with other genes are labelled grey. The plot was created using the plotDendroAndColors *function from the* WGCNA *package in R* (Langfelder & Horvath, 2008).

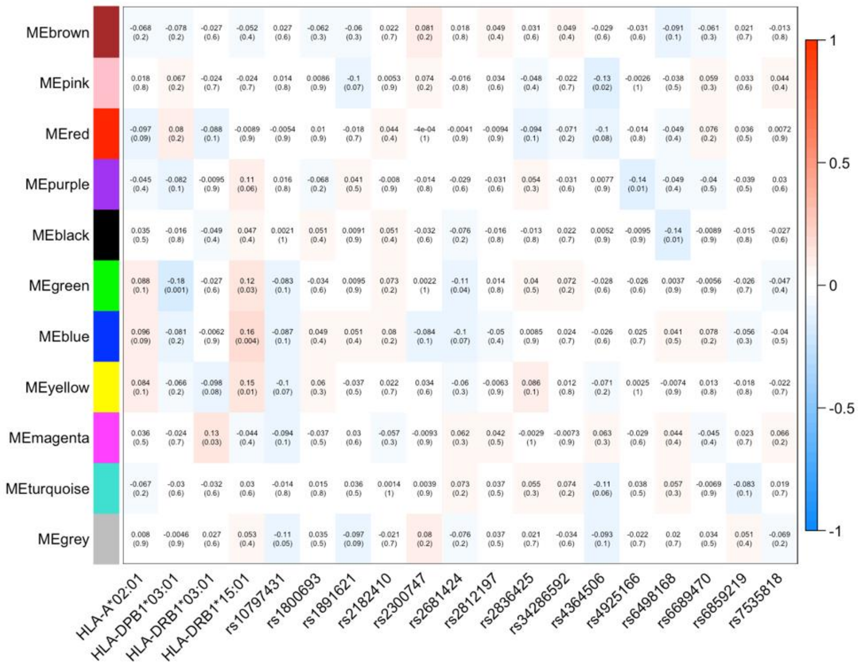

Figure 3.9 ***Correlation between module eigengenes (MEs) and dosage data from 19 MS-associated variants.*** *Correlation coefficients are presented in the first row of each cell. The* p *values for a certain correlation coefficient are given in the brackets in the second row (not adjusted for multiple testing). The grey module consists of unassigned genes. Stronger red or blue color corresponds to stronger associations. The plot was created with the* labeledHeatMap *package from the* WGCNA *package in R* (Langfelder & Horvath, 2008).

3.5.6 Highly preserved module structure between the groups of MS patients

Next, gene co-expression modules were estimated for variant carriers and noncarriers for each MS-associated variant separately, using data from 2171 residuals. Scale-free topology criteria was satisfied with a soft threshold (power) of 3 for most of the variant groups, with the power being 4 only for the carriers of the rs7535818 SNP. After estimating the TOM, the TOM-based dissimilarity matrix was used for hierarchical clustering, with average linkage criteria. The dynamic tree cut approach resulted in four different sets of modules depending on the height where the dendrogram was cut. Split-2 was selected for module preservation analysis because it resulted in a moderate number of modules (ten) in each group (Figure 3.10). Next, module preservation was examined in the following way. Modules were compared in two directions. First, the module from the carriers group was considered as a reference structure and it was examined whether a similar structure exists in the data from noncarriers of a given SNP, in order to investigate the degree of module preservation structure in carriers. Then the reference and test group were flipped and module preservation analysis was repeated to examine the scope of preservation of the module structure in the noncarriers.

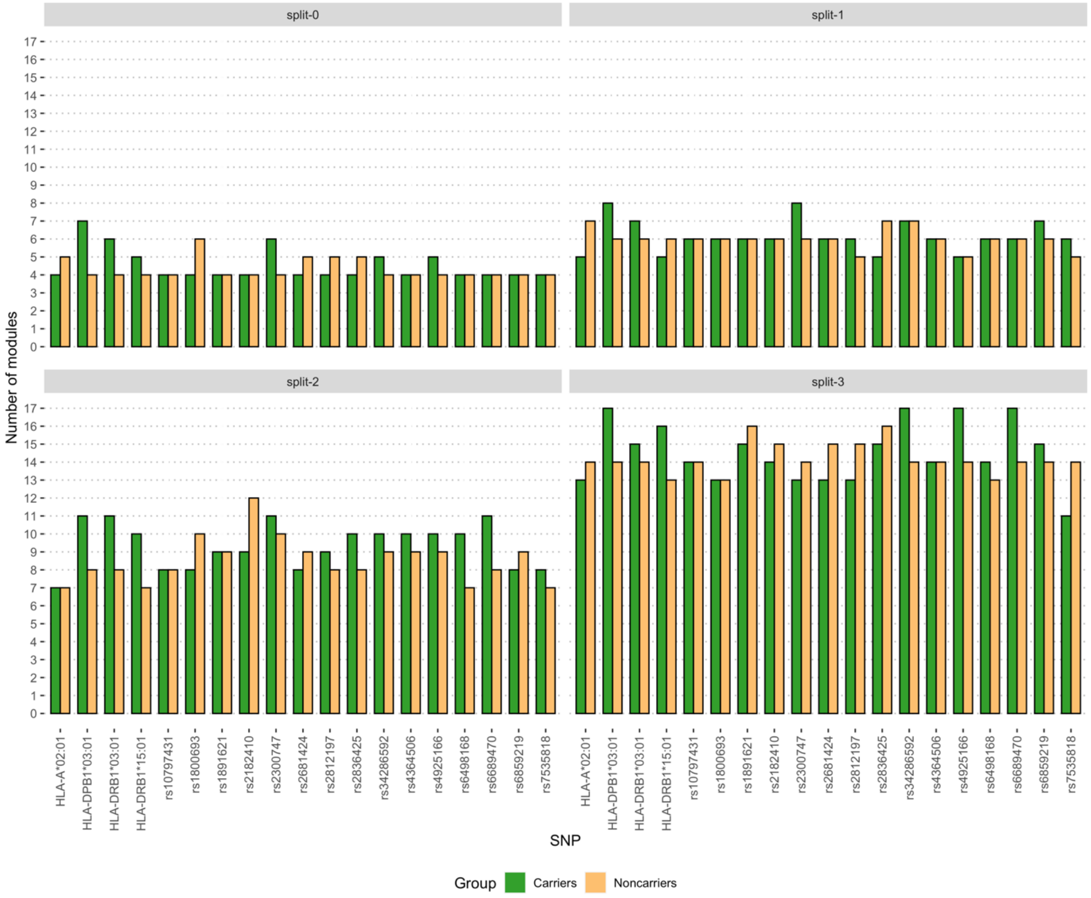

Figure 3.10 ***Number of found modules for each dendrogram cut height ("split"), per SNP and per group.***
The dendrogram obtained from hierarchical clustering of genes based on their interconnectedness after applying the WGCNA was cut on four different heights resulting in four splits. Split-0 was the least deep split (nearest to the main branching start) and split-3 is the deepest split (farthest from the main branching start). The WGCNA was run for each group of each analyzed MS-associated variant (19). Split-2 was selected for module preservation analysis.

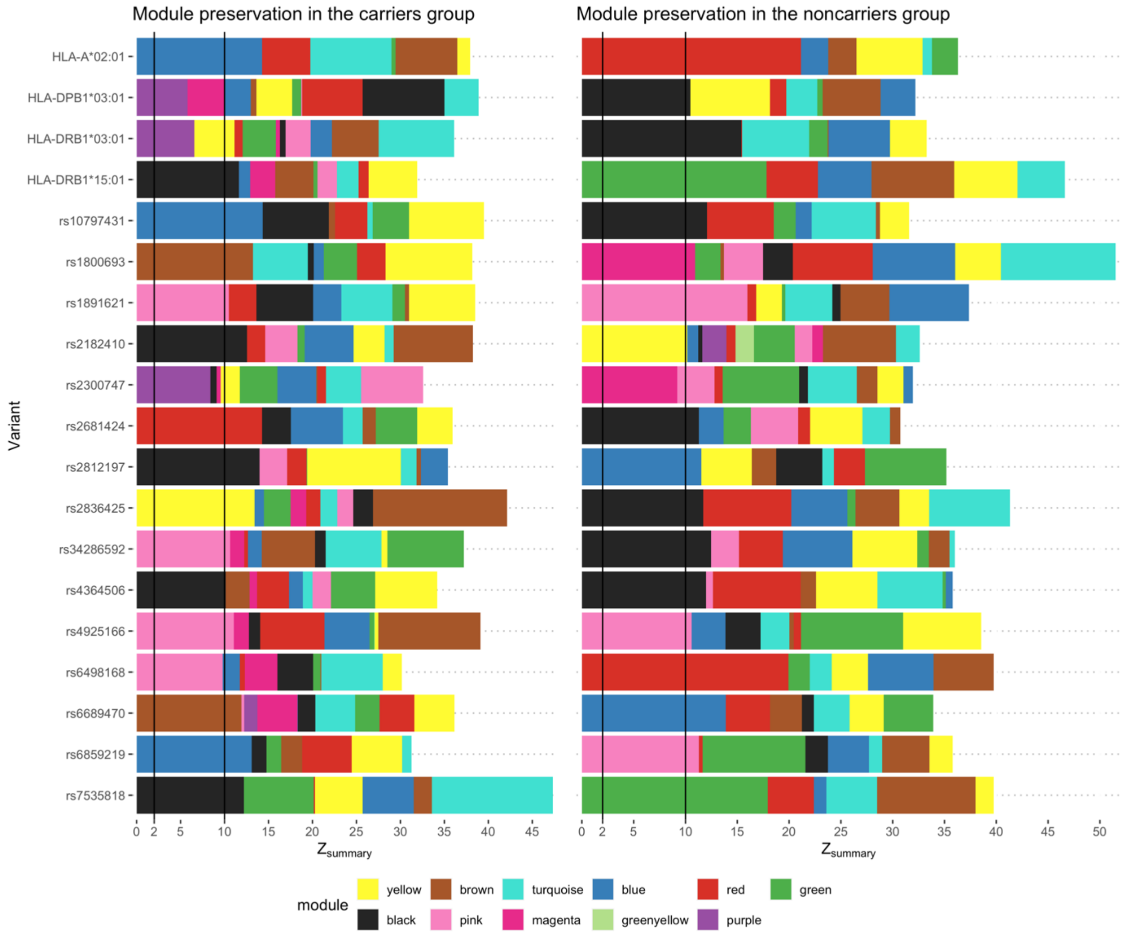

Figure 3.11 ***Module preservation statistics for each variant and the respective group, for both reference groups.*** *On the left plot, the preservation statistic Z_{summar} is presented for each SNP and its modules estimated for the MS patients carrying the listed variant. On the right plot, data from MS patients not carrying the variant was used as a reference group. $Z_{summary}$ statistics are presented in an additive scale for each SNP,* e.g., *the blue module of HLA-A*02:01 carriers has a $Z_{summary}$ of 14.3, and the red module a $Z_{summary}$ of 19.7 and so on. The two vertical lines on the plots label the region where there is low to moderate evidence of module preservation between the reference group (specified in the title of the plot) and the test group. The higher the $Z_{summary}$, the stronger the evidence for module preservation between the groups.*

For each analysis, the $Z_{summary}$ statistics was obtained. In both analyses, modules mostly exerted strong evidence of structure preservation ($Z_{summary} > 10$, Figure 3.11). There were eight modules with $Z_{summary}$ statistic lower than 10, potentially indicating the that there could exist reference group-specific structure. However, most of them still incline towards the "strong evidence of preservation" category ($Z_{summary}$ closer to 10). For one module among the eight with lower *Z* scores, the association with the GO term was still significant (Table 3.6) after correcting for multiple testing following the grouped Benjamini-Hochberg (GBH) approach (Hu et al., 2010). Genes comprising the pink module estimated from patients carrying the rs6498168 variant were associated with *neutrophil degranulation* GO term (weighted *p* value 1.2977 × 10^{-13}). The rs6498168 variant is located in the *CLEC16A* gene (C-Type Lectin Domain Containing 16A), a gene coding for the regulator of mitochondrial autophagy.

Table 3.6 ***Modules showing low to moderate evidence of preservation in the test group.***
Altogether eight different modules exerted $Z_{summary}$ statistic which reflects a possible reference group-specific module structure. Genes within the modules were tested for enrichment in the GO data base. The p *values from hypergeometric test were adjusted following the grouped FDR approach in which* p *values were weighted and then the* p *value specific threshold was estimated. Significant associations are presented in bold font. $Z_{summary}$ = preservation statistics, lower values suggest group-specific module structure; Count = number of genes from the module matching to the pathway in the data base; Size = number of genes in the GO belonging to the respective GO term;* w-p *value = weighted* p *values adjusted in a group-wise manner; threshold =* p *value threshold estimated based on p value index and a weighted alpha level of significance.*

Variant	Reference	Module	Module size	$Z_{summary}$	GO term	Count/Size	w-*p* value	threshold
rs6498168	**Carriers**	**pink**	**44**	**9.77**	***neutrophil degranulation***	**32/447**	**1.2977×10^{-13}**	**7.1725×10^{-6}**
rs2300747	Carriers	magenta	33	9.57	*regulation of proton transport*	2/3	0.0106	0.0002
rs2300747	Noncarriers	magenta	25	9.19	*nuclear division*	5/52	0.0044	0.0002
rs2300747	Carriers	purple	28	8.39	*sequestering of zinc ion*	2/2	0.0026	0.0002
rs2300747	Carriers	black	77	9.10	*actin polymerization or depolymerization*	11/71	0.0004	0.0001
HLA-DRB1*03:01	Carriers	purple	36	6.56	*sulfur compound transport*	2/4	0.0251	0.0003
HLA-DPB1*03:01	Carriers	purple	35	5.79	*translational termination*	2/2	0.0040	0.0002

3.5.7 Subtle group-specific differences in partial correlations between genes in pathway-based GGM networks

Prior to GGM network estimation, genes were divided into biological pathways using the KEGG database, as described in the Methods section. GGMs were estimated per pathway, for each variant group. The resulting models provided coefficients of pairwise partial correlations between genes, as well as corresponding *p* values and posterior probabilities of an edge being direct under the given GGM. The differential network analysis was employed per pathway, as explained in the Method section. In summary, pairwise partial correlations were subtracted between the two groups, resulting in a vector of correlation differences between groups of carriers and noncarriers of a given variant. Two measures were introduced to enable exploring the differences. First, for each variant, absolute pathway-specific differences were summed, creating the global measure of the difference, the *sum*. The second approach was to inspect the highest difference in partial correlations for a given pathway between groups of variant carriers and noncarriers, representing the *max* measure. Permutation analysis was employed to infer whether any of the described measures corresponds to an extremely rare event or if they are also highly likely to occur by chance. After 1000 permutation analysis, in which genotype group labels were shuffled 1000 times, the permutation *p* values from analyzing the *sum* measure were mostly high (*p* value minimum 0.001, median 0.494; Figure 3.12 **A**). The *sum* measure in this analysis setting did not show significant differences in pathway correlation in any of the groups, for any of the pathways after the multiple hypothesis adjusting (adjusted *p* value minimum 0.307, median 1). On the other hand, when edges with the highest difference in partial correlation for the given pathway were analyzed, the permutation *p* value distribution (*p* value minimum 0 (<0.001) , median 0.484; Figure 3.12 **B**) contained *p* values which were considered significant after the multiple testing correction. For nine pathways, the original highest difference in partial correlation (the *max*) was always bigger than the difference calculated in the permuted groups, resulting in the permutation *p* value of <0.001 (Figure 3.13, **A-H**). Such result was observed for altogether five MS-associated variants, the HLA-A*02:01, HLA-DRB1*03:01, rs7535818, rs1891621, and rs4364506 respectively (Table 3.7). On the 1000 permutation level, the *GSTM1-GSTM4* gene pair exerted a statistically significant difference in partial correlation between groups of the rs7535818 variant carriers and noncarriers in four different pathways. Protein products of *GSTM1* and *GSTM4* genes are cytoplasmic glutathione transferases of the *mu* class, important in detoxification pathways of drugs, environment toxins and similar, corresponding to the KEGG pathway annotations in Table 3.7.

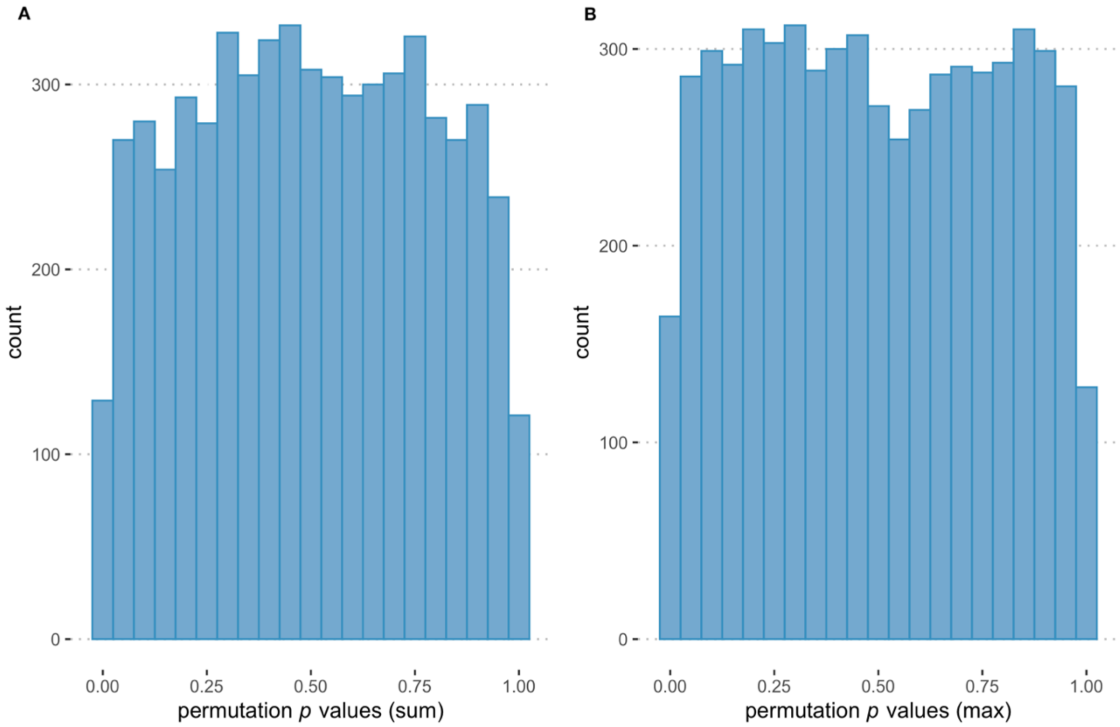

Figure 3.12 ***Permutation*** **p** ***value distribution after the analysis of 1000 permutations.***
(A) *Permutation* p *values from all variants, across all pathways testing the significance of the original* sum *measure.*
(B) *The resulting* p *values from 1000 genotype shuffling of all variants testing the statistical significance of the original* max *measure across all tested pathways.*

Carriers and noncarriers of the HLA-A*02:01 variant showed a statistically different partial correlation between *KLC2* and *MAP3K10* genes in pathway related to Huntington disease. *KLC2* gene codes for a light chain kinesin, involved in the cellular motor function. MAP3K10 is a mitogen activated protein kinase-kinase, important in signal transduction. In the acute myeloid leukemia (AML) pathway, partial correlation between transcriptional factor CEBPE and transcriptional co-repressor RUNX1T1 was significantly different between the groups of HLA-DRB1*03:01 carriers and noncarriers. When groups of rs1891621 were analyzed, the highest difference in partial correlation in T_H17 cell differentiation pathway was detected between the *IL2RG* gene, coding for a signaling component of many interleukin receptors and *MAPK10*, another mitogen activated protein kinase-kinase. This difference was always higher than the difference calculated using permuted group labels. The analysis of the groups of the rs4364506 variant suggested a different partial correlation between genes involved in activation of vitamin K, *VKORC1* and *VKORC1*, in the corresponding KEGG pathway (Table 3.7).

To get a more precise *p* value estimate for the *max* measure in nine pathways, genotypes were shuffled additional 100,000 times. After the multiple testing correction, none of the *p* values sustained the multiple testing adjustment (Table 3.7, Figure 3.13 **I-O**).

Table 3.7 ***Permutation statistics from nine pathways which passed the significance threshold after 1000 permutation runs testing the pathway-specific*** **max** ***measure.***

In nine pathways the original highest difference in partial correlation (the max*) was always higher than the permuted value after the 1000 permutation runs.* p *values below the Bonferroni adjusted 5% cut off (0.00263) were considered significant.* p *value = permutation* p *value; adj.* p *value = Holm corrected* p *value.*

Variant	**KEGG ID** KEGG pathway	Gene pair	1000 permutation		100,000 permutation	
			p value	adj. *p* value	*p* value	adj. *p* value
HLA-A*02:01	**hsa05016** Huntington disease	KLC2-MAP3K10	<0.001	<0.001	0.0030	0.9118
HLA-DRB1*03:01	**hsa05221** acute myeloid leukemia	CEBPE-RUNX1T1	<0.001	<0.001	0.0003	0.0890
rs7535818	**hsa05200** pathways in cancer	GSTM1-GSTM4	<0.001	<0.001	0.0003	0.0980
	hsa05225 hepatocellular carcinoma	GSTM1-GSTM4	<0.001	<0.001	0.0013	0.3880
	hsa05418 Fluid shear stress and atherosclerosis	GSTM1-GSTM4	<0.001	<0.001	0.0012	0.3810
	hsa00983 Drug metabolism	GSTM1-GSTM4	<0.001	<0.001	0.0018	0.5490
rs1891621	**hsa04659** T17 cell differentiation	IL2RG-MAPK10	<0.001	<0.001	0.0008	0.2579
rs4364506	**hsa00130** Metabolism of cofactors and vitamins	VKORC1-VKORC1L1	<0.001	<0.001	0.0009	0.2947

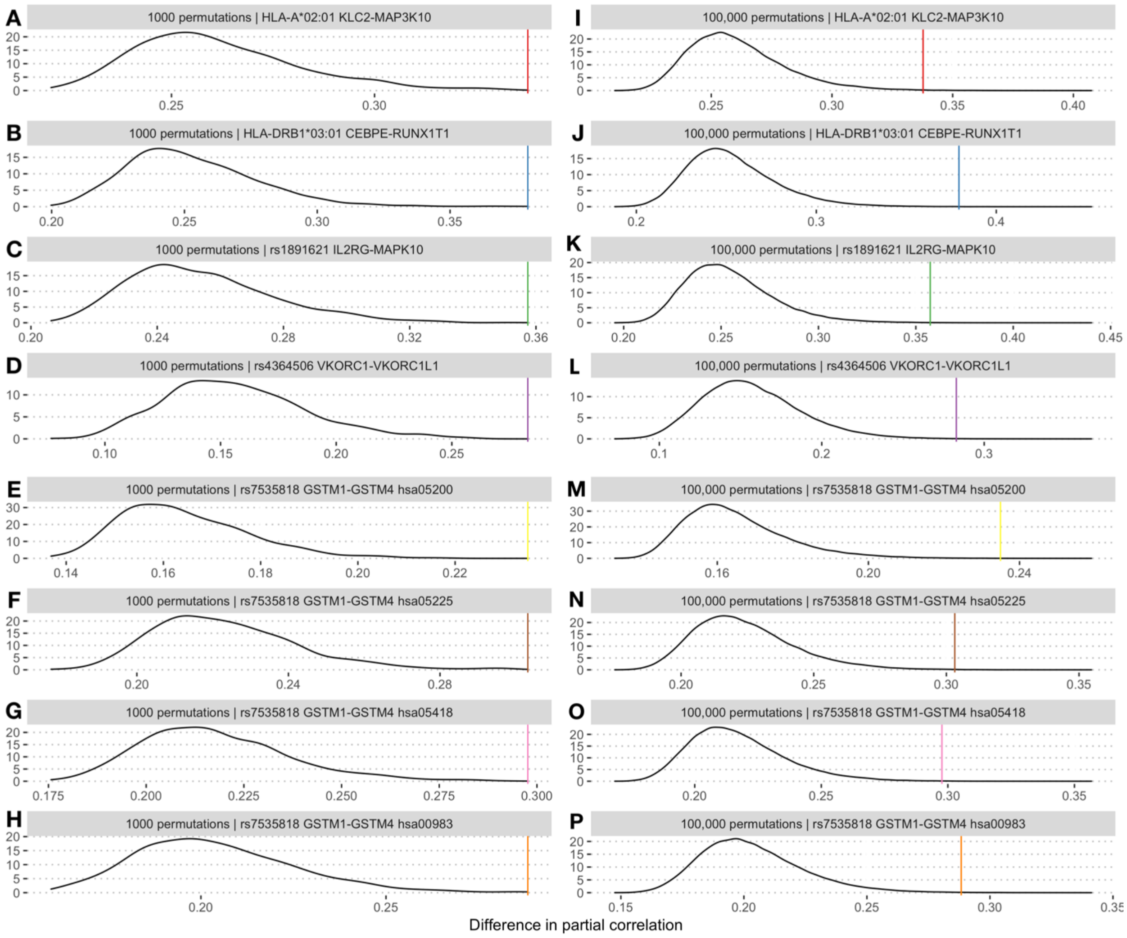

Figure 3.13 ***Posterior distributions of partial correlation differences in gene pairs which passed the multiple testing correction after analysis of 1000 permutations.***

Graphs ***A-H*** *show 1000 differences from 1000 permutation analysis. Each curve represents the posterior distribution of the partial correlation difference for the given pathway-specific edge, estimated from permutation runs. The MS-associated variant and corresponding gene pairs are given in the title of the facets. Corresponding distributions resulting from 100,000 permutations are shown on graphs* ***I-P****. Vertical lines represent the difference calculated between the original group setting and line colors indicate eight separate partial correlations, each specific for an MS-associated variant, gene pair, and the pathway.*

3.5.8 Gene connectivity analysis of pathway-based GGM networks suggests that rs6689470 variant might influence the dynamics of the hsa05416 pathway

Pathway-specific networks estimated *via* the GGM were further analyzed in the context of node connectivity. In the GGM, for each of the 307 pathways, 38 different networks were built (one for each variant group). This resulted in 843,068 different entities, which are network-specific nodes whose characteristics depend on a gene set (pathway) and a sample (group). Node connectivity profiles were compared between 19 sets of networks from carriers and 19 sets of networks from noncarriers, which are the two reference groups. In the context of conditional independence analysis, the node degree was declared zero if the entity didn't have any direct connections in that pathway.

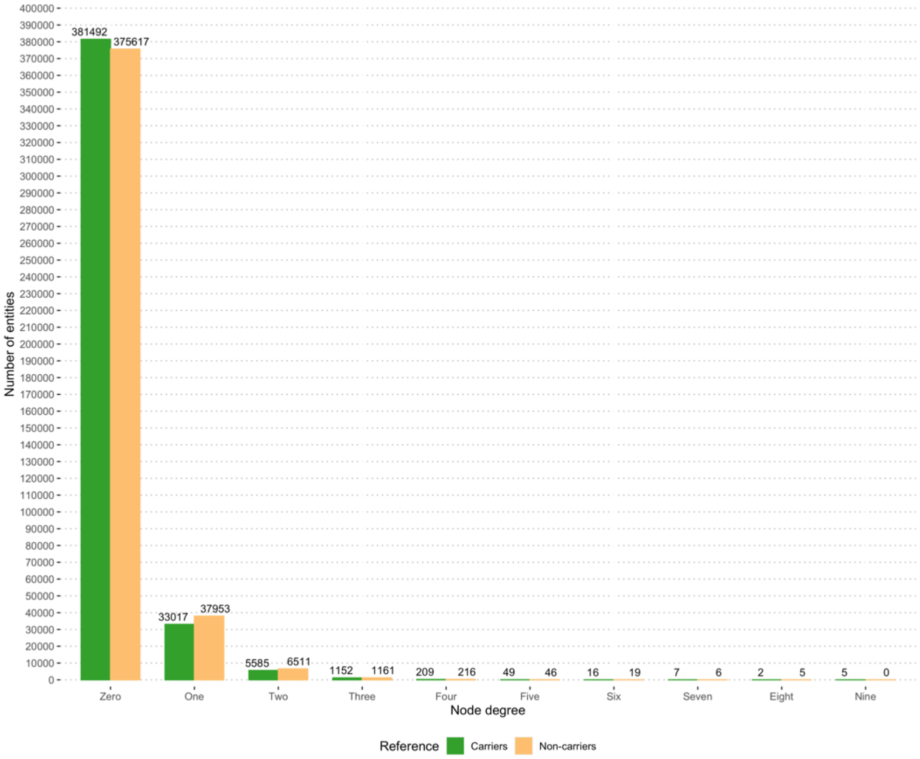

Figure 3.14 ***Node degree distribution in the data.***
For each entity, it was checked how many direct connections it established, based on GGM probabilities via Bayes factor calculation. An entity is a network-specific node whose characteristics depend on a gene set (pathway) and a sample (group). The Bayes factor was calculated using data from patients carrying the variant and the patients not carrying the variant separately, therefore introducing two reference groups for each pathway, that is, for each node. The node degree of zero describes the edges with estimated Bayes factors < 5.6, therefore, having no direct edges.

Most of the entities in networks estimated from both reference groups did not establish any direct connections (90,5% for carriers and 89,1% in noncarriers). The rest of the nodes established at least one direct connection in both carriers and noncarriers (40042 and 45917, respectively). Among nodes with direct connections, the majority formed only one or two direct connections in both reference groups (96,4% in carriers and 96,8% in noncarriers). The highest achieved node degree was nine in the carriers reference group and eight among noncarriers (Figure 3.14). It was decided to analyze nodes with at least three direct connections for two reasons. First, nodes with a larger number of connections are more likely to represent genes exerting important roles in a pathway. And second, by selecting nodes with at least three direct connections, we came the closest to meeting the top 5% criteria often used in the research, as already mentioned in the Introduction section. This left 1440 entities sharing 671 unique gene names from 141 pathways in the carriers reference group and 1453 nodes covering 559 unique gene names from 156 pathways in the noncarriers reference group, including all 19 variants. For both groups, the differential analysis was run as described in the Methods section. Based on the 1000 permutations, for 31 nodes among the noncarriers, the analysis resulted in less than 0.01% probability (permutation *p* value 0) of observing such an extreme difference in connectivity compared to the same node in the carriers group, given that the null hypothesis is true. In carriers, there were 57 such nodes (Figure 3.15 **A**). Together, for 86 nodes, spanning 45 different pathways a significant result was observed. However, for each of the 86 nodes, the estimated permutation *p* value was zero (<0.001), hence, a deeper level of permutation testing was needed to infer a more precise *p* value. The groups were shuffled for additional 10,000 times and the permutation analysis was run. The difference in connectivity for most of the nodes did not reach a statistical significance after 10,000 permutations (Figure 3.15 **B**). However, ten nodes, spanning seven different pathways still exerted statistical significance, with a permutation *p* value < 0.0001. Pathway-based networks estimated from the data from patients carrying one of the four variants still suggested potential regulation rewiring. There were six nodes in four different pathways with significantly more direct connections in carriers, based on the analysis of 10,000 permutations. A similar result was observed in pathway-based networks of four pathways in noncarriers, where the degree of four nodes was significantly higher than the degree of the same entity in carriers (Figure 3.15 **B**).

Altogether ten entities, six from the carriers' data and four from networks estimated based on the data from noncarriers, were further tested on a 100,000 permutation level in order to obtain a more precise *p* value estimate.

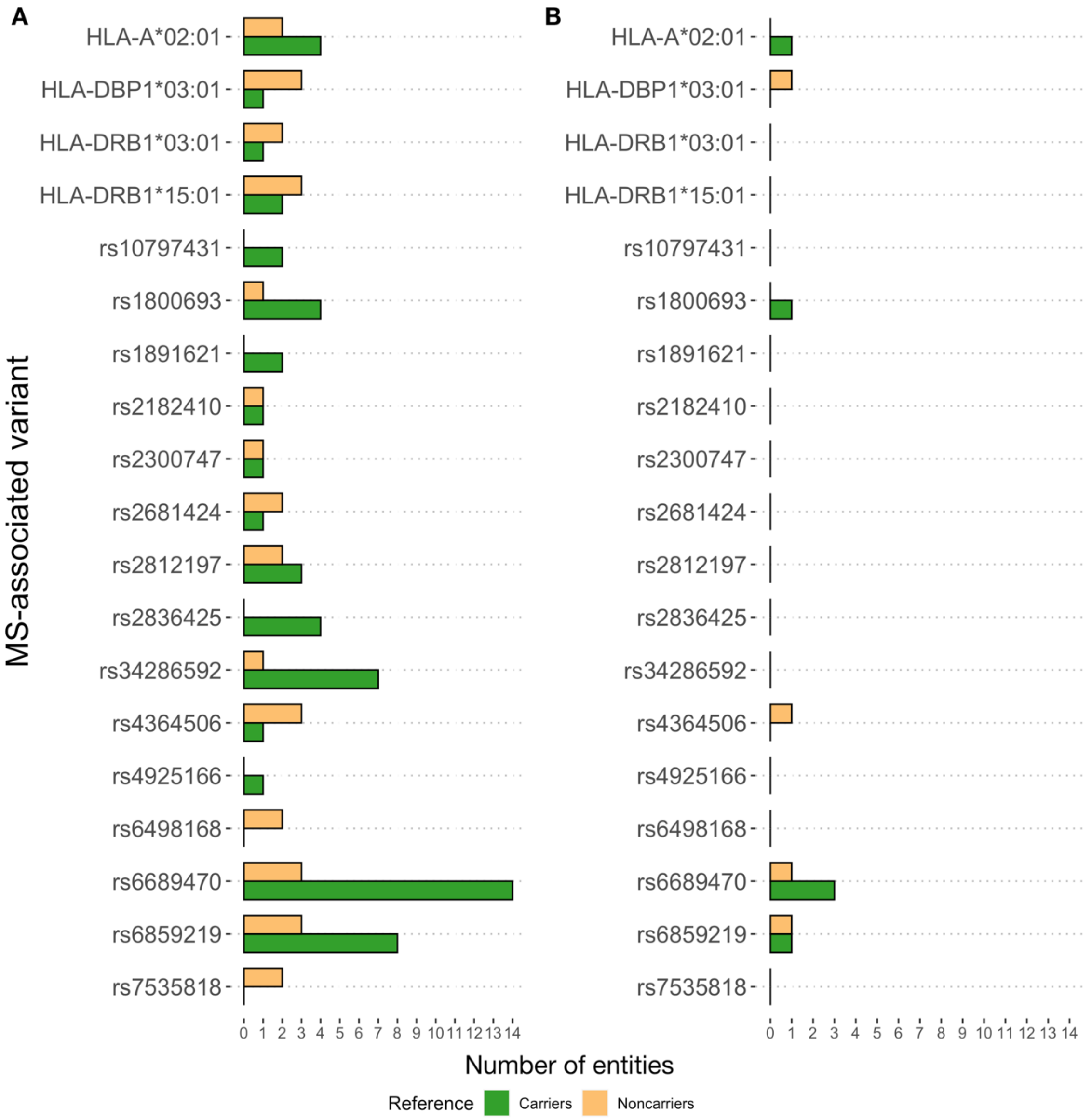

Figure 3.15 ***Analysis of differential connectivity: number of entities with statistically significant difference in connectivity between groups of a variant, depending on the reference group.***
*(**A**) Altogether 31 entities from noncarriers and 57 entities from carriers were found significant, based on 1000 permutation results. (**B**) These entities were further tested on a 10,000 permutation level, where most of the tested nodes did not reach statistical significance.*

3.5.8.1 HLA-DOB and ACTB genes exert significantly stronger connectivity profiles in rs6689470 carriers

Based on the analysis of 100,000 permutations, nodes representing transcripts mapping to the *ACTB* and *HLA-DOB* genes from the hsa05416 pathway formed significantly more direct connections in a network estimated on the data coming from patients carrying the rs6689470 variant when compared to the network of the same pathway estimated on the data from patients not carrying the variant (adj. *p* values 0.001 and 0.002, respectively; Table 3.8). The hsa05416 pathway consists of altogether 60 genes in the KEGG database. In the KKNMS data set, 48 genes were matched. The pathway has been annotated to play an important role in viral myocarditis, a cardiac disease by which the myocardium is injured and subject to inflammation. The disorder can be caused by direct cytopathic effects of a virus and also by the autoimmune response triggered by the viral infection (Esfandiarei & McManus, 2008).

In the network of the hsa05416 pathway estimated from 208 patients not carrying the variant nodes ACTB and HLA-DOB did not form direct connections with any among 48 genes comprising the pathway in the KKNMS data set. On the other hand, they achieved a markedly high degree in the network obtained from the gene expression data of 106 MS patients carrying the variant (6 and 3, respectively; Figure 3.16). The *ACTB* gene codes for one of the six highly conserved actin proteins, the building units of filaments, some of which comprise cell's cytoskeleton. In viral myocarditis emergence, actin gets rearranged upon receiving upstream signals, enabling virus particle entry into the cell (Yajima & Knowlton, 2009). HLA-DOB is the beta chain protein coded by HLA class II *HLA-DOB* gene. Together with the alpha chain (DOA), they build the HLA-DO heterodimer immersed in the membrane of intracellular vesicles in B cells and in a subset of thymic medullary epithelium, a microenvironment important for the tolerance induction of T cells (Welsh & Sadegh-Nasseri, 2020). Additionally, the HLA-DPA1, another class II MHC gene, was just above the Bonferroni adjusted threshold, also exerting stronger connectivity patterns in the viral myocarditis pathway (Table 3.8).

3.5.8.1.1 ACTB and HLA-DOB seem to exert strong connectivity profile specific to hsa05416 pathway

To explore the connectivity profiles of the two genes in other pathways estimated from the data of rs6689470 carriers and noncarriers, it was first ascertained in how many other pathways do these genes appear. ACTB is found in 28 other pathways in the data set (out of 307), while HLA-DOB appears in 23 pathways, both excluding the hsa05416. Their degree was then examined by looking into the number of direct connections they formed in the networks of those pathways. The ACTB gene formed direct connections in four pathways, one direct connection in each of the four. The HLA-DOB formed direct connections in two more pathways, one in the hsa05164 network and two direct connections in the hsa05168 (Table 3.9). None of the entities in other pathways were tested in the permutation analysis because they didn’t achieve the critical number of direct connections (3).

Table 3.8 ***Differential connectivity results after 100,000 permutation runs.***
Ten nodes exerting statistical significance after 10,000 permutations were further tested on a deeper permutation level. Two nodes from hsa05416 pathway were statistically significant after the multiple hypothesis testing correction. Results statistically significant after 100,000 permutations and the multiple testing adjustment, i.e., *adj.* p *value < 0.00263 are presented in bold font. Reference = reference group;* p *value = permutation* p *value after 100,000 permutations; adj.* p *value = Holm adjusted* p *value; Node = gene name of a node; Cnc = connectivity of a node in noncarriers; Cc = connectivity of a node in carriers; DC = difference in connectivity (reference – test).*

Variant	KEGG ID	Reference	*p* value	adj. *p* value	Node	Cnc	Cc	DC	Pathway name
rs6689470	**hsa05416**	**carrRef**	**2×10^{-5}**	**0.001**	**ACTB**	**0**	**6**	**6**	***Viral myocarditis***
rs6689470	**hsa05416**	**carrRef**	**3×10^{-5}**	**0.002**	**HLA-DOB**	**0**	**3**	**3**	***Viral myocarditis***
rs6689470	hsa05416	carrRef	5×10^{-5}	0.003	HLA-DPA1	1	5	4	*Viral myocarditis*
rs6689470	hsa05322	noncarrRef	9×10^{-5}	0.007	HIST1H3H	3	1	2	*Systemic lupus erythematosus*
HLA*A:0201	hsa05200	carrRef	0.0002	0.011	MDM2	0	3	3	*Pathways in cancer*
HLA*DPB1:0301	hsa05168	noncarrRef	3×10^{-5}	0.004	ZNF304	8	0	8	*Herpes simplex virus 1 infection*
rs1800693	hsa05168	carrRef	5×10^{-5}	0.007	ZNF254	0	5	5	*Herpes simplex virus 1 infection*
rs6859219	hsa04928	noncarrRef	5×10^{-5}	0.004	PLCB4	3	0	3	*Parathyroid hormone synthesis, secretion and action*
rs6859219	hsa04621	carrRef	9×10^{-5}	0.007	CASP8	0	4	4	*NOD-like receptor signaling pathway*
rs4364506	hsa00010	noncarrRef	0.0002	0.01	LDHA	5	0	5	*Glycolysis / Gluconeogenesis*

Table 3.9 ***Direct edges of the ACTB and HLA-DOB nodes in the pathway networks of rs6689470 carriers and noncarriers.***

The connections of the ACTB and HLA-DOB nodes exerting statistical significance after 100,000 permutations are shown in bold, and they belong to the hsa05416 pathway. Other direct connections in the rest of the analyzed pathways are presented in regular text format. BF = Bayes factor; pcor = partial correlation estimated from the GGM; p *value =* p *value from the GGM, relating to the pcor;* q *value = FDR corrected* p *value; prob = GGM-based posterior probability that the edge is direct.*

Node1	Node2	KEGG ID/KEGG name	Reference	Carriers					Noncarriers				
				BF	pcor	*p* value	*q* value	prob	BF	pcor	*p*-value	*q*-value	prob
ACTB	**HLA-E**	**hsa05416** Viral myocarditis	carriers	23.232	0.166	0.000	0.013	0.959	0.000	0.092	0.121	0.910	0.000
	ACTG1			23.232	0.157	0.000	0.020	0.959	0.000	0.087	0.143	0.923	0.000
	HLA-C			10.983	-0.138	0.001	0.049	0.917	0.000	0.055	0.353	0.967	0.000
	PRF1			10.983	0.137	0.002	0.050	0.917	0.000	0.160	0.007	0.361	0.000
	ITGB2			5.639	0.127	0.003	0.076	0.849	0.000	0.167	0.005	0.279	0.000
	CD40LG			5.639	-0.122	0.005	0.089	0.849	0.000	0.053	0.369	0.969	0.000
HLA-DOB	**CD80**			9.154	0.135	0.002	0.054	0.902	0.000	-0.003	0.961	0.988	0.000
	MYH6			5.639	0.127	0.003	0.075	0.849	0.000	-0.071	0.234	0.951	0.000
	CASP8			5.639	-0.126	0.004	0.077	0.849	0.000	0.045	0.447	0.974	0.000
ACTB	ACTG1	**hsa05164** Influenza A	carriers	119.109	0.215	0.000	0.002	0.992	1.889	0.184	0.001	0.200	0.654
ACTB	ITGB2	**hsa04390** Hippo signaling pathway	noncarriers	0.000	0.107	0.005	0.806	0.000	6.175	0.160	0.000	0.139	0.861
ACTB	ATP2A3	**hsa05410** Hypertrophic cardiomyopathy		0.000	0.054	0.164	0.934	0.000	7.890	0.140	0.000	0.085	0.888

Node1	Node2	KEGG ID/KEGG name	Reference	Carriers					Noncarriers				
				BF	pcor	*p* value	*q* value	prob	BF	pcor	*p*-value	*q*-value	prob
ACTB	ATP2A3	**hsa05412** Arrhythmogenic right ventricular cardiomyopathy	noncarriers	0.000	0.062	0.117	0.840	0.000	17.949	0.139	0.000	0.042	0.947
HLA-DOB	IL6	**hsa05164** Influenza A	carriers	11.400	-0.173	0.000	0.048	0.919	0.000	-0.055	0.306	0.974	0.000
HLA-DOB	MB21D1	**hsa05168** Herpes simplex virus 1 infection	noncarriers	0.000	-0.026	0.217	0.845	0.000	13.569	-0.122	0.000	0.033	0.931
	ZNF30			0.000	0.009	0.666	0.944	0.000	794.211	0.158	0.000	0.000	0.999

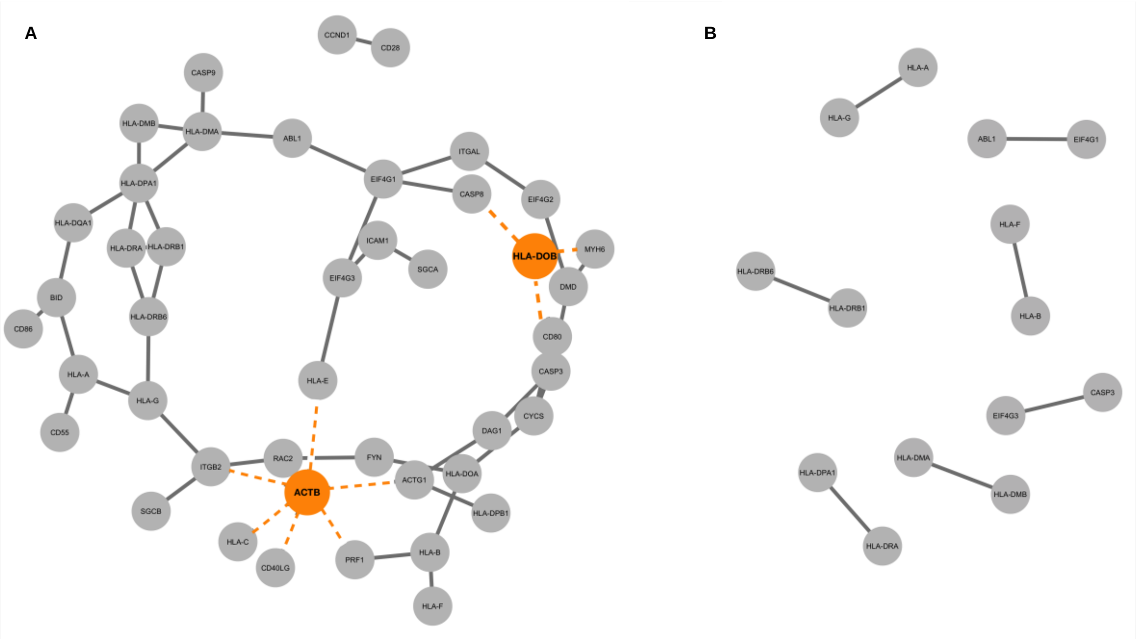

Figure 3.16 ***GGM-based sparse graphs of the hsa05416 pathway estimated from the gene expression data of MS patients carrying the rs6689470 variant (A) and patients not carrying the variant (B).***

*(**A**) In patients carrying the MS-associated variant, the network consists of nodes forming more direct connections than the (**B**) network of noncarriers. Nodes colored orange represent genes whose number of direct connections is significantly different when the two networks are contrasted, based on analysis of 100,000 permutations. Their direct connections are represented by the dashed, orange edges. Graphs were created with* network (Butts, 2008) *and* ggnet (Briatte, 2020) *packages in R. The distances between the nodes in the network and the length of edges are random.*

4 | Discussion

The analysis of a complex disease such as multiple sclerosis bares manyfold challenges. Our body systems belong within the big ecosystem and have to function in the midst of its web of influences. There are factors like genetic inheritance, environment, and epigenetics which intertwine in somewhat random ways and model our susceptibility to the disease. By finding important associations between genetic polymorphisms and the disease itself, we can solve at least one part of the equation which would then lead to a better understanding of disease pathogenetics. In this thesis, MS pathogenetics was investigated in two separate projects. Firstly, in murine models for MS, using which evidence is presented that the spontaneously induced EAE mirrors the role of human MS risk variants more faithfully than MOG-induced EAE. Secondly, the direct effect of MS-associated variants on regulation in gene co-expression networks from MS patients was examined, suggesting the potential involvement of the rs6689470 variant during early stages of MS. In this work, it is suggested that this variant might influence a more coordinate regulation of proteins involved in T cell egress from thymus into the CNS as well as a strongly synchronized activation of autoimmune response, when compared to MS patients not carrying the variant.

4.1 Human MS risk variants associated with gene expression changes in OSE

This section discusses the analysis and the results published in and related to the research paper by Faber, Kurtoic, and the colleagues (Faber et al., 2020).

Mouse models of human MS are used to explore the pathogenesis of the disease and to develop novel treatments for patients suffering from MS. But, it is still not fully clear whether currently used EAE models can mirror such complex disease as is MS, where genetics and the environment shape the disease susceptibility. It was therefore interesting and of high value to explore to which extent do OSE and MOG-induced EAE reflect the etiology of MS, as well as the involvement of genetics in disease pathogenesis.

In comparison to MOG EAE, mice developing the EAE spontaneously experience gene expression changes closely linked to immune pathways, potentially indicating more complexity in disease induction as well. Previous research has already observed B and T cell cooperation in OSE, a process important in MS pathophysiology (Lehmann Horn, Kronsbein, & Weber, 2013; Molnarfi et al., 2013). Transcripts specific for OSE were enriched for both human MS risk genes and T_H cell specific transcripts. OSE-specific transcripts showed an overrepresentation of immune-specific gene sets. Nonetheless, genes specific for both OSE and MOG EAE indicate that both models faithfully recapitulate critical functional pathways of MS, for example the role of antigen presentation and $CD4^+$ T cells in the immunopathogenesis of MS (Moutsianas et al., 2015; Patsopoulos et al., 2013). Furthermore, the transcripts mapping to the *H2-Eb1* and *H2-Ab1* genes, homologs of the *HLA-DRB5* and *HLA-DQB1* genes, were one of the mostly differentially expressed probes. The alleles *HLA-DRB*01:01* and *HLA-DQB1*06:02* are part of the DR15-DQ6 haplotype and are strongly associated with MS. These alleles are in strong linkage disequilibrium with the *HLA-DRB1*15:01*, for which the involvement in B and T cell interaction in the brain has been found, leading to the activation and growth of autoreactive $CD4^+$ T cells (Jelcic et al., 2018). The cooperation of B and T cells has also been proven important in the OSE development (Molnarfi et al., 2013).

These results show that the EAE, especially the OSE, represents a beneficial model for studying the role of genetics in MS susceptibility. Previous research findings support this claim. For example, in a humanized EAE, where the TCR recognizes the human myelin binding protein, risk variants in the MHC region, including the major risk variant, *HLA-DRB1*15:01,* were successfully replicated (Gregersen et al., 2006). Furthermore, risk loci linked to T_H cell differentiation are found to be conserved between humans and mice, and are implicated in both MS and the EAE (Blankenhorn et al., 2011). In an adoptive transfer EAE study, MS risk genes were found to be differentially regulated in pathogenic $CD4^+$ T cells, thereby underlining the role of MS risk genes in EAE pathophysiology (Hoppmann et al., 2015).

In this work, the pivotal role of lymphocyte activation in EAE induction has been highlighted, and the role of T_H1 and T_H17 cells has been further examined. Their function in CNS autoimmunity has been hotly debated (Hiltensperger & Korn, 2018). When comparing the T cells from the spleen of four OSE mice, the number of T_H1-specific transcripts was higher than that of T_H17. A high T_H1/T_H17 ratio is indicative of a lesion distribution pattern, where the spinal cord is primarily affected, a case observed for both EAE models (Domingues et al., 2010; Stromnes, Cerretti, Liggitt, Harris, & Goverman, 2008).

Genes overlapping between CDT and OSE_1ex gene sets and T_H1-specific transcripts were significantly enriched in MS risk genes (Table 2.5). T_H17-specific transcripts did not show such enrichments. In the OSE_4sp gene set, MS risk genes were enriched in T_H1 cells at only a nominal significance (unadjusted *p* value = 0.0097). The GO overrepresentation analyses revealed that immune-related biological processes like the *positive regulation of T cell proliferation* were significant for T_H1-specific genes in the OSE_4sp set. On the other hand, in the MOG_4sp gene set, T_H1-specific genes were not enriched in MS risk genes (unadjusted *p* value = 0.51), and there were no GO terms associated with T_H1-specific MOG_4sp genes. These findings suggest that, in the context of T_H1-related immune responses, the OSE model might be more closely linked to human MS risk genes than MOG EAE is. The relevance of T_H17 cells is still not completely comprehended, because T_H17 cells can shift toward the T_H1 phenotype in EAE. Therefore, the T_H1 markers analyzed in this study may, to a certain extent, actually represent the expression of former T_H17 cells.

In this work, it was observed that most genes differentially expressed in OSE_1 mice exert the same pattern in severely affected OSE_4 mice, with the same direction of regulation. Therefore, many factors active in severe EAE are also involved during the mild stage or, potentially, early disease course. As already mentioned before, the OSE develops more gradually than the MOG-induced EAE, which also makes it more similar to human MS. Studying the stages of OSE, especially the mild stage, might provide an interesting model for defining the initial triggers of MS, given that the factors determining the onset and the course of MS are largely unknown (Krishnamoorthy et al., 2006).

However, the gene expression analysis of the two EAE models suffers from several limitations. Microarrays covered only part of the murine transcriptome, therefore limiting the number of MS risk genes which could be analyzed. Next, the statistical power of the analysis is hampered by the small sample size. Lastly, the initial phases of EAE are hard to define, because the disease develops quickly. Here, mild OSE was representing the early disease stage, but it cannot be said with certainty whether these mice would develop a more severe EAE in the future.

4.2 The exclusive effect of each of the 19 analyzed MS-associated variants on gene expression in MS patients appears to be subtle

A common way of exploring the relationship between genetic polymorphisms and gene expression levels have been the eQTL analyses which test the influence of phenotype-associated variants (detected by GWAS) on gene expression and can give insight on putatively causal genes. These analyses, however, cannot provide us with insights about mechanisms underlying the association between the SNP and the disease. In the subset of MS patients in the KKNMS cohort, the eQTL analysis confirmed already known eQTL effects (Table 3.5, Andlauer et al., 2016s). The associations between the rest of the known variants outside of the MHC region were not detected in this sample. This could be due to a lower quality of probes mapping to corresponding genes, therefore excluding the probe from the analysis or simply due to a lack of power. The differential expression analysis resulted in similar signals. Groups of variants mostly showed similar levels of gene expression, suggesting that variants alone do not exert strong influence on gene expression in this sample of MS patients. For variants rs4925166 and rs10797431, the previous eQTL study found associations with the loci TOP3A and MMEL1, respectively (Andlauer et al., 2016). Furthermore, in this work, the differential expression analysis showed that patients not carrying the variant exert significantly different expression levels of those loci when compared to patients carrying the variant (Table 3.4). Four probes mapping to HLA genes were differentially expressed between groups of HLA-DRB1*15:01 and HLA-DRB1*03:01 carriers and noncarriers (Table 3.4). In previous research, where peripheral blood mononuclear cells and umbilical cord blood cells from healthy individuals were subject to RNA-Seq analysis, it has been shown that the gene expression of HLA loci indeed depends on the genetic variants occupying the locus (Yamamoto et al., 2020).

4.2.1 Direct effects of MS-associated variants on gene expression and Pearson's correlation between gene expression profiles appear to be comparatively low

MS patients not carrying the rs2836425 variant showed a statistically significant difference in Pearson's correlation between the genes *DHRS13* and *UBOX5*, with the absolute difference of 0.822 (adj. p value 3.05×10^{-4}). The rs2836425 variant is located in *ERG* gene, an erythroblast transformation-specific (ETS) related gene, which is a transcriptional regulator. The *DHRS13* gene codes for the dehydrogenase 13 and the *UBOX5* for proteins involved in the ubiquitination pathway. The ubiquitination, or the ubiquitin-proteasome system (UPS), is a post-translational modification where proteins are labelled to be degraded *via* the proteasome complex. Previous research has shown that the UPS is involved in myelin protein degradation in MS (Belogurov et al., 2014; Giordana, Richiardi, Trevisan, Boghi, & Palmucci, 2002). Furthermore, it has been shown that MS patients exert lower levels of UPS activity after treatment with IFN-beta-1b, which is correlated with better clinical status of patients after six month of IFN-beta-1b therapy (Minagar et al., 2012). Interestingly, the process of neddylation, analogous to the ubiquitination, has been linked to T cell function regulation (Mathewson et al., 2016). Very recently it has also been shown that the inhibition of neddylation in mice resulted in decrease in EAE severity (Kim et al., 2021).

However, to the best of my knowledge, the *DHRS13* gene and *UBOX5* have not been annotated to be functionally connected. The potential influence of the variant on the correlation between these two genes can either be an example of a type I error or a very interesting novel finding for which additional data sets are needed in order to either support it or confront it.

4.2.2 Immune-related modules show high preservation between the groups of MS patients

Variability in gene expression levels in modules estimated on gene expression data of 2171 immune-related genes from all 314 patients was weakly explained by MS-associated variants, with correlation coefficients centered around zero (Figure 3.9). The module preservation analysis yielded similar results, providing strong evidence of structure preservation between modules of different groups, with the lowest $Z_{summary}$ score of 5.793, which still falls into the category of having moderate evidence of module preservation, suggesting that the modules do not show group-specific characteristics (Table 3.6). However, no GO term was significantly associated with this module, after correcting for multiple hypothesis testing *via* the GBH and TST procedures. The pink module estimated from patients carrying the rs6498168 variant was the only module in the data set with the $Z_{summary}$ statistic below 10 and for which associated GO term was found. Most of the genes of the pink module (32/44) belong to the *neutrophil degranulation* GO term. The process of neutrophil degranulation is a very common process during the inflammation, whereby neutrophils excrete granules containing proinflammatory substances (Lacy, 2006). The impact of neutrophils on the immune response in MS patients is still not fully clear, but research does provide evidence of their emerging role in MS pathogenesis (Woodberry, Bouffler, Wilson, Buckland, & Brüstle, 2018). Yet, the preservation analysis of the pink module still yielded the $Z_{summary}$ score close to 10, and it is therefore hard to make any strong claims regarding the functional aspect of the rs6498168 variant in the context of MS. It could be that this effect is invoked by the random sampling error. The same effect should be further examined in an independent sample of MS patients, in order to inspect whether we can find stronger evidence that the rs6498168 risk variant influences the neutrophil degranulation pathway in MS patients.

The reason why such a low signal was observed after performing the WGCNA might be due to gene space restriction. Examining genes specifically annotated to participate in the immune response could have a downside of restricting the gene space too much. This might lead to removing highly variable genes just because they are not immune-system related, while being a differentiating factor between the two groups with respect to the MS-associated variant. Thus, instead of subsetting the genes based on their function, one could subset them based on gene expression variability between the groups, as often performed when analyzing gene expression data *via* WGCNA (Liang et al., 2018; Tang et al., 2018). This would potentially provide more power because the WGCNA would be run on the genes exerting high expression variance. Such modules would potentially also exert group specific structures yielding lower $Z_{summary}$ scores. GO enrichment analysis of such modules might reveal pathways not exclusively related to the immune-system, but nonetheless contributing to our knowledge about mechanisms in the early stage of the disease.

4.2.3 Differential network analysis shows potential in uncovering distinct patterns of partial correlation in GGM-based pathway networks

In the analysis of conditional independence, a bigger subset of genes was included, with genes exerting more diverse roles. The Gaussian graphical model (GGM) was applied to gene sets annotated to have a similar function based on the KEGG database to infer sparse gene co-expression networks. Comparison of variant-specific pathway-based sparse graphs enabled removing potentially spurious correlations between genes, reducing the background noise in the data. Such an approach takes the effect of other genes in the pathway into account, whereby removing such influences we can explore whether an association between two genes is direct, without mediators, or does it depend on other genes. The analysis revealed low variability in the partial correlation, suggesting similar pathway regulation, irrespective of genetic background. For five variants there was suggestive evidence of their influence on pathway regulation, based on 1000 permutations and after adjusting for multiple hypothesis testing. The edge between glutathione transferase genes *GSTM1* and *GSTM4* showed significant difference in partial correlation between groups of rs7535818 variant carriers and noncarriers in four different pathways, which is a potentially interesting finding. One would expect that if a variant influences the gene-gene interaction, and a certain gene pair appears in more than one pathway in the data set, the effect of a variant would be detectable in more than one pathway. However, this difference was not strong enough to sustain the deeper level of permutation, with at least thirty random differences (*p* value = 0.0003 in the hsa05200 pathway after 100,000 permutations, Table 3.7) stronger than the original difference (Figure 3.13.). It would be interesting to examine this edge in a bigger sample setting, because it is possible that this study did not have enough power to find evidence to support the alternative hypothesis.

4.2.4 Differential connectivity analysis suggests the involvement of rs6689470 MS risk variant in self-reactive T cells evasion into the CNS in the early stages of MS

The exploration of the biggest difference in partial correlation in the pathway could potentially be a too specific analysis, and might overlook many other potentially interesting edges, while examining a sum of all differences in a network might be too general, in a way that edge-specific variation is potentially lost. The network connectivity analysis would enable more edges in a pathway to be analyzed, while edge specific characteristics would be contained.

Gene connectivity patterns mostly show robustness towards the influence of MS-associated variants tested in the KKNMS sample. However, patients carrying the rs6689470 variant potentially exert patterns of higher connectivity in the network of the hsa05416 pathway (KEGG pathway name: *Viral myocarditis*) in comparison to MS patients not carrying the risk variant (Table 3.8). The connectivity of nodes mapping to the *ACTB* and *HLA-DOB* genes differed significantly between the two groups, potentially suggesting the underlying mechanism of the MS associated variant (adj. *p* values < 0.00236, Table 3.8, Table 3.9).

The rs6689470 variant is located in an intron of the *Gfi-1* gene, coding for a regulator of lymphocyte activation and development (X. Liao, Buchberg, Jenkins, & Copeland, 1995). The eQTL study by Andlauer and the colleagues (Andlauer et al., 2016) presented the association between the rs6689470 variant and gene expression changes in the *EVI5* gene. Ecotropic viral integration site 5 protein, coded by the *EVI5* gene, modulates the cell cycle progression, cytokinesis, and cellular membrane traffic (Zhou et al., 2016). The relationship of EVI5 to human MS is still unclear, and further research is needed to investigate whether the causal allele acts through the *EVI5* or another gene. The influence of the *EVI5* gene on the T cell function needs to be studied in more detail as well as its relationship to retroviral elements (Hoppenbrouwers et al., 2008).

4.2.4.1.1 The dynamic interaction between cytoskeleton and immune response in early stages of the disease might be governed by the rs6689470 variant

The two genes whose connectivity was significantly different between the groups of patients seem to play important roles in the *Viral myocarditis* pathway by participating in the viral particle entry (actin beta) and peptide presentation on antigen presenting cells (HLA-DOB) (Esfandiarei & McManus, 2008). The function which HLA-DOB exerts in viral myocarditis is not necessarily myocarditis-specific. HLA-DOB is a beta chain protein coded by the HLA class II *HLA-DOB* gene. Together with the alpha chain (DOA), they build the HLA-DO heterodimer immersed in the membrane of intracellular vesicles in B cells and in a subset of thymic medullary epithelium, a microenvironment important for tolerance induction of T cells. The HLA-DO heterodimer contributes to the selection of immunodominant epitopes. It is blocking promotion of self-reactive T cells by binding to the HLA-DM molecule (Welsh & Sadegh-Nasseri, 2020). A disbalance in self-peptide presentation to T-cells enables self-reactive CD4 T cells to escape into the periphery. Furthermore, mice with a knocked-out *DO* gene have shown to be more susceptible to EAE development, producing more CD4$^+$ T cells specifically targeting the MOG component of the myelin sheath (Welsh et al., 2020).

The HLA-DOB established three direct connections in the hsa05416 pathway network of rs6689470 carriers. It formed connections with nodes corresponding to transcripts mapping to *CD80*, *MYH6,* and *CASP8* genes (Table 3.9, and Figure 3.16). The *CD80* gene is a putative risk gene for MS (IMSGC, 2019b), coding for a ligand present on T cells. Interaction between the CD80 molecule on T cells and the CD28 molecule on B cells constitutes a costimulatory signal for T cells, thereby activating them (Menezes et al., 2014). The myosin heavy chain 6 *(*MYH6*)* gene, on the other hand, codes for a protein comprising the cardiac muscle thick filament, important in muscle contraction (Razmara & Garshasbi, 2018). Lastly, the HLA-DOB is directly connected to the node represented by the transcript mapping to the *CASP8* gene, coding for caspase-8, a protease playing a crucial role in inhibiting inflammatory cell death, *i.e.*, necroptosis. In caspase-8-deficient conditions, the cell death pathway is activated leading to the loss of oligodendrocytes and demyelination. Defective caspase-8 has been found in cortical lesions of MS patients, therefore suggesting that the deficient enzyme might play a role in MS pathogenesis (Ofengeim et al., 2015). The analysis of connectivity based on conditional independence in a pathway-specific network has found that the regulation of *HLA-DOB* expression is associated with the expression of the *CD80* gene, involved in T cell activation, the *CASP-8* gene, involved in necroptosis inhibition and the *MYH6* gene, coding for a building element of heart muscle. Even though the MYH6 does not fully fit the MS context, we should not forget that the analysis puts focus on the HLA-DOB, suggesting its differential connectivity pattern. The role of MYH6 is probably tissue-specific, but other three genes (HLA-DOB, CASP-8, and CD80) do have more general roles in terms of immune system response regulations, which are processes important for various types of inflammation, including both the myocarditis and the MS. Through the lens of the hsa05416 pathway, we might be able to learn more about associations between the rs6689470 variant and the immune response in the early stages of the disease. MS patients not carrying the rs6689470 variant might therefore be more susceptible to MS due the more strongly synchronized regulation of the three proteins, based on the stronger connectivity profile. The expression of HLA-DOB might be reduced, with the CD80 molecule being more expressed leading to a more prominent activation of self-reactive T cells and decreased expression of caspase-8, resulting in more effective necroptosis in the CNS.

Furthermore, the differential connectivity was observed for the ACTB node, forming six direct connections in the patients carrying the rs6689470 variant (Table 3.8, Table 3.9, and Figure 3.16) and zero direct connections in the other group. The two of the nodes map to HLA genes, namely the HLA-C and HLA-E. The HLA-C has been described as a mediator of NK cell and T cell activation (Blais, Dong, & Rowland-Jones, 2011), similarly to the role of the HLA-E, which is also involved in NK cell activation (Rölle, Jäger, & Momburg, 2018). Next, the CD40LG is a ligand expressed on activated CD4$^+$ T cells, and it binds to the CD40 molecule on, *e.g.*, B cells, thereby activating the immune response (Cleary, Fortune, Yellin, Chess, & Lederman, 1995). The ACTB node is further connected to gene coding for the integrin beta (ITGB2), important for cellular adhesion and the innate immune response (Arnaout, 1990). Interestingly, the ACTB forms a direct connection with pore forming protein perforin, coded by the *PRF1* gene, which is found to be associated with MS risk in Sardinian population (Sidore et al., 2020). Lastly, the ACTB is directly connected to a gene coding for gamma actin (ACTG1). Together with other actins, including the ACTB, it plays an important role in B cell cytoskeleton reorganization upon contact with antigens (Welsh & Sadegh-Nasseri, 2020). The ACTB therefore might be mediating the immune response in close connection to cytoskeleton dynamics. In MS patients carrying the rs6689470 risk variant, it was observed that the connectivity pattern of ACTB changes, potentially promoting a stronger autoimmune response.

In this work, gene expression of patients in the early stage of MS was studied. Thus, it is conceivable that the rs6689470 variant is involved very early on in MS development, by influencing the pathways related to the negative selection of T cells, closely associated with B activation. It could be that patients carrying the rs6689470 variant exert a higher risk for MS because self-reactive T cells evade the negative selection in the thymus. On the other hand, the actin beta, coded by the *ACTB* gene, is an element of the cell cytoskeleton. During B cell activation, the cytoskeleton is reorganized upon the contact of the cell with the antigen. The cytoskeleton network plays an important role in antigen entry into the cell as well as its processing (Harwood & Batista, 2011). To better discern the sequence of the events, it is important to explore the connectivity of HLA-DOB and ACTB in healthy individuals as well, in order to examine the wild-type connectivity profile of these genes.

4.2.5 Study highlights and limitations

In this thesis, it is suggested that the rs6689470 variant influences interactions of HLA-DOB and ACTB nodes in a pathway-specific network of MS patients in the early disease stage. Taking into account the sample size employed to examine subtle effects of MS-associated variants with mostly moderate influence on disease risk (ORs < 2, Table 3.3), it could be that this study is underpowered and was unable to provide stronger evidence of genetic influence on actin importance in B cell antigen internalization and activation, which can further induce the proliferation of myelin-specific T cells. However, direct calculations of power would be too complex to estimate for this project, due to ample assumptions one made by employing the linear regression, the GGM, and the WGCNA analysis. Previous research has provided proofs of concept that one indeed can obtain analyses integrating genomics, transcriptomics, (and epigenomics) on a sample of only 70 people (Pineda et al., 2015). Furthermore, a differential network analysis based on Gaussian graphical model inference was successfully performed on 58 lung adenocarcinoma samples (He et al., 2019).

The study on the sample of KKNMS patients potentially only slightly opened the doors to examining the influence of the rs6689470 variant, as well as to investigating important aspects of cytoskeleton involvement in MS pathogenesis. However, this effect needs to be studied further in an independent population sample of MS patients, where the same connectivity analysis would be conducted, then providing evidence for a population-wide effect.

A very important characteristic of this sample is its homogeneity and the treatment-naïve background of MS patients. MS patients comprising the sample are all Germans, *i.e.*, with low underlying population structure (Table 3.2). The fact that most of the patients did not take any DMTs (Table 3.2), is an advantage, because the DMT can indeed alter gene expression levels and thereby bias the results (Nickles et al., 2013). Furthermore, all patients in the sample are at either early-stage MS or suffer from the CIS. Hence, this sample provides a window into associations potentially governing the early stages of MS. It is therefore imaginable that the homogeneity of the sample could contribute to unveiling the very subtle effects, irrespective of a potentially small sample size, such as the effect of rs6689470 risk variant in the early stages of MS.

In this work gene expression data from whole blood was analyzed. Whole blood is a more convenient source to extract from individuals in comparison to, *e.g.,* brain tissue. However, the correspondence in expression levels between whole blood and the brain indicates important differences, even though certain brain transcripts do co-express in blood (Rollins, Martin, Morgan, & Vawter, 2010), and whole blood is often used as a proxy for studying the disorders of the CNS (Wittenberg et al., 2020). It indeed makes sense to explore the gene expression levels in the whole blood because according to the extrinsic model for MS emergence, the disease is triggered on the periphery, followed by the travel of self-reactive T cells into the CNS (Dendrou, Fugger, & Friese, 2015). Nonetheless, whole blood is a mixture of many different cell types, and in MS pathogenesis, certain cell types exert specific roles (IMSGC, 2019a). Even though gene expression data was adjusted for the presence of most important cell types in the sample, there still is a certain extent of background noise present in the data and the signal from cell lines important in MS pathogenesis is diluted. A study examining the associations between gene expression levels in specifically B cells or T cells and MS-associated variants might provide us with more power to unveil B- or T-cell specific events.

Furthermore, only part of the MS risk variants was analyzed in the workflow. It would be interesting to explore the nature of association between the rest of the MS risk variants and gene expression profiles of MS patients. Additionally, analyses on healthy individuals are needed to gain insight into the MS-unrelated levels of gene interaction, in order to inspect whether the connectivity in a pathway network increases due to the presence of the rs6689470 variant. Unfortunately, due to a lack of time, it was not possible to perform the analyses on the data from healthy individuals. Finally, replication analyses are needed in order to examine whether the results found in the KKNMS study are sample-specific or can we can find evidence that the association is a more general effect in the population of MS patients, therefore validating the results presented in this thesis. Analyzing one sample representing the population is not enough to enable us to draw conclusions about the population-wide effect. However, due to the specific characteristics of the KKNMS sample, it was challenging to find a corresponding sample to perform a replication analysis. The characteristics include early-stage MS patients who are mostly treatment naïve, with low population diversity, and the availability of genotype data for the rs6689470 variant. Unfortunately, to this end, it was not possible to find another sample with such specifics.

5 | Conclusion and outlook

Understanding the biology underlying the associations between genetics and human traits has been an engaging task. Analyses of complex diseases like MS would profit if a proper animal model would be found which faithfully represents the disease, enabling studying its etiology also in the context of disease pathogenetics. However, no animal model fully reflects the diversity of human diseases like MS, and each EAE model recapitulates only partial aspects of the disease. In this thesis, evidence is provided that spontaneously induced EAE is more closely linked to human MS risk genes and T_H cell biology. In conclusion, in comparison to the MOG-induced EAE, the OSE model might represent the human MS more faithfully, enabling studying human MS pathogenesis and defining specific therapeutic targets. Additional studies are needed to examine the similarity between OSE and human MS genetical background in more depth.

Furthermore, in this thesis a workflow has been presented to explore the influence of MS-associated variants on gene expression from MS patients in order to gain insight into underlying biological mechanisms of genetic influence on disease susceptibility. It was observed that the gene expression profiles of patients remain robust, irrespective of their genetic background. However, the analyses of conditional independence suggest that the rs6689470 risk variant might influence cytoskeleton re-organization, which contributes to B cell activation. These cells can further activate self-reactive T cells, supporting the inflammation process in the CNS in the early stages of the disease, corresponding to the disease stage of MS patients in the KKNMS sample. Due to the limited sample size of this study, it is important to examine these effects in an independent sample of MS patients, in order to validate results presented in this thesis.

When studying polygenic diseases such as MS, it can be difficult to capture their complexity. It is therefore important to employ an integrative approach which includes different data sources (clinical data, demographics, genetics, transcriptomics, cell proportions, environmental factors). The workflow presented in this thesis enables integration of different sources of data, coming closer to capturing the different sources of variability which in the end produce the disease. Furthermore, by analyzing the gene-gene interactions in the context of the network, we also enabled the exploration of the dynamics of the system, instead of a single-gene approaches, as in, *e.g.*, traditional eQTL analyses or differential expression analyses. Moreover, by dividing the genes into meaningful subunits, such as biological pathways, computational intensity was decreased and interpretability of the results was increased. The application of differential network approaches, as the one proposed in thesis, on the gene expression data from B or T cells might provide insights about changes in functioning of specific cell types, which are governed by genetic variants. By applying such an approach on the data from patients in the early stages of the disease we might discern the cell-specific mechanisms which happen close to the disease onset. This would potentially advance the development of the drug therapy, helping thousands of individuals suffering from the disease and might bring us one step closer to finding the cure. Furthermore, by enabling data exploration on different levels (differential gene expression, correlation analyses, and conditional independence analyses), the workflow provides powerful tools to compare groups of interest in order to study complex gene-gene regulation patterns in other complex human diseases and to provide us with the knowledge about the underlying pathological processes. Such insights can be used to enable deeper understanding of disease pathology, potentially leading to the advances in the medical treatment and to the earlier diagnosis.

6 | Bibliography

Absinta, M., Sati, P., & Reich, D. S. (2016). Advanced MRI and staging of multiple sclerosis lesions. *Nature Publishing Group*. https://doi.org/10.1038/nrneurol.2016.59

Alcina, A., de Abad-Grau, M. M., Fedetz, M., Izquierdo, G., Lucas, M., Fernández, Ó., … Matesanz, F. (2012). Multiple sclerosis risk variant HLA-DRB1*1501 associates with high expression of DRB1 gene in different human populations. *PLoS ONE*, *7*(1), 1–9. https://doi.org/10.1371/journal.pone.0029819

Andlauer, T. F. M., Buck, D., Antony, G., Bayas, A., Bechmann, L., Berthele, A., … Müller-Myhsok, B. (2016). Novel multiple sclerosis susceptibility loci implicated in epigenetic regulation. *Science Advances*, *2*(6), e1501678. https://doi.org/10.1126/sciadv.1501678

Arloth, J., Bader, D. M., Röh, S., & Altmann, A. (2015). Re-Annotator: Annotation Pipeline for Microarray Probe Sequences. *PLOS ONE*, *10*(10), e0139516. https://doi.org/10.1371/journal.pone.0139516

Arnaout, M. (1990). Structure and function of the leukocyte adhesion molecules CD11/CD18. *Blood*, *75*(5), 1037–1050. https://doi.org/10.1182/blood.v75.5.1037.1037

Ascherio, A., Munger, K. L., White, R., Köchert, K., Simon, K. C., Polman, C. H., … Pohl, C. (2014). Vitamin D as an Early Predictor of Multiple Sclerosis Activity and Progression. *JAMA Neurology*, *71*(3), 306. https://doi.org/10.1001/jamaneurol.2013.5993

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000, May). Gene ontology: Tool for the unification of biology. *Nature Genetics*. Nat Genet. https://doi.org/10.1038/75556

Axisa, P.-P., & Hafler, D. A. (2016). Multiple sclerosis. *Current Opinion in Neurology*, *29*(3), 345–353. https://doi.org/10.1097/WCO.0000000000000319

Bae, H., Perls, T., Steinberg, M., & Sebastiani, P. (2015). Bayesian polynomial regression models to fit multiple genetic models for quantitative traits. *Bayesian Analysis*, *10*(1), 53–74. https://doi.org/10.1214/14-BA880

Bar-Or, A., Pender, M. P., Khanna, R., Steinman, L., Hartung, H. P., Maniar, T., … Joshi, M. J. (2020, March 1). Epstein–Barr Virus in Multiple Sclerosis: Theory and Emerging Immunotherapies. *Trends in Molecular Medicine*. Elsevier Ltd. https://doi.org/10.1016/j.molmed.2019.11.003

Barabási, A.-L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, *5*(2), 101–113. https://doi.org/10.1038/nrg1272

Baranzini, S. E., & Oksenberg, J. R. (2017). The Genetics of Multiple Sclerosis: From 0 to 200 in 50 Years. *Trends in Genetics*, *33*(12), 960–970. https://doi.org/10.1016/j.tig.2017.09.004

Belogurov, A., Kudriaeva, A., Kuzina, E., Smirnov, I., Bobik, T., Ponomarenko, N., … Gabibov, A. (2014). Multiple sclerosis autoantigen myelin basic protein escapes control by ubiquitination during proteasomal degradation. *Journal of Biological Chemistry*, *289*(25), 17758–17766. https://doi.org/10.1074/jbc.M113.544247

Ben-nun, A., Kaushansky, N., Kawakami, N., Krishnamoorthy, G., Berer, K., Liblau, R., & Hohlfeld, R. (2014). From classic to spontaneous and humanized models of multiple sclerosis : Impact on understanding pathogenesis and drug development. *Journal of Autoimmunity*. https://doi.org/10.1016/j.jaut.2014.06.004

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, *93*(3), 491–507. https://doi.org/10.1093/biomet/93.3.491

Blais, M. E., Dong, T., & Rowland-Jones, S. (2011, May). HLA-C as a mediator of natural killer and T-cell activation: Spectator or key player? *Immunology*. Wiley-Blackwell. https://doi.org/10.1111/j.1365-2567.2011.03422.x

Blankenhorn, E. P., Butterfield, R., Case, L. K., Wall, E. H., Del Rio, R., Diehl, S. A., … Teuscher, C. (2011). Genetics of experimental allergic encephalomyelitis supports the role of T helper cells in multiple sclerosis pathogenesis. *Annals of Neurology*, *70*(6), 887–896. https://doi.org/10.1002/ana.22642

Bos, S. D., Berge, T., Celius, E. G., & Harbo, H. F. (2016). From genetic associations to functional studies in multiple sclerosis. *European Journal of Neurology*, *23*(5), 847–853. https://doi.org/10.1111/ene.12981

Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., & Sherlock, G. (2004). GO::TermFinder - Open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, *20*(18), 3710–3715. https://doi.org/10.1093/bioinformatics/bth456

Briatte, F. (2020). ggnet: Functions to plot networks with ggplot2. R package version 0.1.0. Retrieved from https://github.com/briatte/ggnet

Butts, C. T. (2008). network: A package for managing relational data in R. *Journal of Statistical Software*, *24*(2), 1–36. https://doi.org/10.18637/jss.v024.i02

Cai, C., Langfelder, P., Fuller, T. F., Oldham, M. C., Luo, R., van den Berg, L. H., … Horvath, S. (2010). Is human blood a good surrogate for brain tissue in transcriptional studies? *BMC Genomics*, *11*(1). https://doi.org/10.1186/1471-2164-11-589

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., … Gaudet, P. (2009). AmiGO: Online access to ontology and annotation data. *Bioinformatics*, *25*(2), 288–289. https://doi.org/10.1093/bioinformatics/btn615

Chase Huizar, C., Raphael, I., & Forsthuber, T. G. (2020). Genomic, proteomic, and systems biology approaches in biomarker discovery for multiple sclerosis. *Cellular Immunology*, *358*, 104219. https://doi.org/10.1016/j.cellimm.2020.104219

Chikina, M., Zaslavsky, E., & Sealfon, S. C. (2015). CellCODE: A robust latent variable approach to differential expression analysis for heterogeneous cell populations. *Bioinformatics*, *31*(10), 1584–1591. https://doi.org/10.1093/bioinformatics/btv015

Cleary, A. M., Fortune, S. M., Yellin, M. J., Chess, L., & Lederman, S. (1995). Opposing roles of CD95 (Fas/APO-1) and CD40 in the death and rescue of human low density tonsillar B cells. *Journal of Immunology*, *155*(7), 3329–3337. Retrieved from http://www.jimmunol.org/content/155/7/3329.abstract

Damsker, J. M., Hansen, A. M., & Caspi, R. R. (2010). Th1 and Th17 cells: Adversaries and collaborators. *Annals of the New York Academy of Sciences*, *1183*, 211–221. https://doi.org/10.1111/j.1749-6632.2009.05133.x

de la Fuente, A. (2010). From “differential expression” to “differential networking” - identification of dysfunctional regulatory networks in diseases. *Trends in Genetics*. https://doi.org/10.1016/j.tig.2010.05.001

de la Fuente, A., Bing, N., Hoeschele, I., & Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, *20*(18), 3565–3574. https://doi.org/10.1093/bioinformatics/bth445

Dendrou, C. A., Fugger, L., & Friese, M. A. (2015). Immunopathology of multiple sclerosis. *Nature Publishing Group*, *15*(9), 545–558. https://doi.org/10.1038/nri3871

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., & West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, *90*(1 SPEC. ISS.), 196–212. https://doi.org/10.1016/j.jmva.2004.02.009

Domingues, H. S., Mues, M., Lassmann, H., Wekerle, H., & Krishnamoorthy, G. (2010). Functional and Pathogenic Differences of Th1 and Th17 Cells in Experimental Autoimmune Encephalomyelitis. *Plos O*, *5*(11). https://doi.org/10.1371/journal.pone.0015531

Du, P., Kibbe, W. A., & Lin, S. M. (2008). lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, *24*(13), 1547–1548. https://doi.org/10.1093/bioinformatics/btn224

Dunn, O. J. (1961). Multiple Comparisons among Means. *Journal of the American Statistical Association*, *56*(293), 52–64. https://doi.org/10.1080/01621459.1961.10482090

Dunning, M. J., Smith, M. L., Ritchie, M. E., & Tavare, S. (2007). beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics*, *23*(16), 2183–2184. https://doi.org/10.1093/bioinformatics/btm311

Dyment, D. A., Dessa Sadnovich, A., & Ebers, G. C. (1997). Genetics of multiple sclerosis. *Human Molecular Genetics*, *6*(10 REV. ISS.), 1693–1698. https://doi.org/10.1093/hmg/6.10.1693

Ebers, G. C. (1985). Optic Neuritis and Multiple Sclerosis. *Archives of Neurology*, *42*(7), 702–704. https://doi.org/10.1001/archneur.1985.04060070096025

Efron, B. (2008). Simultaneous inference: When should hypothesis testing problems be combined? *The Annals of Applied Statistics*, *2*(1), 197–223. https://doi.org/10.1214/07-AOAS141

Enz, L. S., Zeis, T., Schmid, D., Geier, F., van der Meer, F., Steiner, G., … Schaeren-Wiemers, N. (2020). Increased HLA-DR expression and cortical demyelination in MS links with HLA-DR15. *Neurology(R) Neuroimmunology & Neuroinflammation*, *7*(2), 656. https://doi.org/10.1212/NXI.0000000000000656

Esfandiarei, M., & McManus, B. M. (2008). Molecular biology and pathogenesis of viral myocarditis. *Annual Review of Pathology: Mechanisms of Disease*, *3*, 127–155. https://doi.org/10.1146/annurev.pathmechdis.3.121806.151534

Faber, H., Kurtoic, D., Krishnamoorthy, G., Weber, P., Pütz, B., Müller-Myhsok, B., … Andlauer, T. F. M. (2020). Gene Expression in Spontaneous Experimental Autoimmune Encephalomyelitis Is Linked to Human Multiple Sclerosis Risk Genes. *Frontiers in Immunology*, *11*(September), 1–12. https://doi.org/10.3389/fimmu.2020.02165

Ferber, I. A., Brocke, S., Taylor-Edwards, C., Ridgway, W., Dinisco, C., Steinman, L., … Fathman, C. G. (1996). Mice with a disrupted IFN-gamma gene are susceptible to the induction of experimental autoimmune encephalomyelitis (EAE). *The Journal of Immunology*, *156*(1).

Fisher, R. A. (1921). 014: On the “Probable Error” of a Coefficient of Correlation Deduced from a Small Sample. *Metron*, 3–32.

Fox, J., & Weisberg, S. (2020). *An R Companion to Applied Regression*. Retrieved from https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Fuller, T. F., Ghazalpour, Æ. A., Aten, Æ. J. E., Drake, Æ. T. A., Lusis, A. J., & Horvath, Æ. S. (2007). Weighted gene coexpression network analysis strategies applied to mouse weight, 463–472. https://doi.org/10.1007/s00335-007-9043-3

Gill, R., Datta, S., & Datta, S. (2010). A statistical framework for differential network analysis from microarray data. *BMC Bioinformatics*, *11*(1), 95.

Giordana, M. T., Richiardi, P., Trevisan, E., Boghi, A., & Palmucci, L. (2002). Abnormal ubiquitination of axons in normally myelinated white matter in multiple sclerosis brain. *Neuropathology and Applied Neurobiology*, *28*(1), 35–41. https://doi.org/10.1046/j.1365-2990.2002.00372.x

Glatigny, S., & Bettelli, E. (2018, November 1). Experimental Autoimmune Encephalomyelitis (EAE) as Animal Models of Multiple Sclerosis (MS). *Cold Spring Harbor Perspectives in Medicine*. NLM (Medline). https://doi.org/10.1101/cshperspect.a028977

Gregersen, J. W., Kranc, K. R., Ke, X., Svendsen, P., Madsen, L. S., Thomsen, A. R., … Fugger, L. (2006). Functional epistasis on a common MHC haplotype associated with multiple sclerosis. *Nature*, *443*(7111), 574–577. https://doi.org/10.1038/nature05133

Grifka-Walk, H. M., Lalor, S. J., & Segal, B. M. (2013). Highly polarized Th17 cells induce EAE via a T-bet independent mechanism. *European Journal of Immunology*, *43*(11), 2824–2831. https://doi.org/10.1002/eji.201343723

Grimes, T., Potter, S. S., & Datta, S. (2019). Integrating gene regulatory pathways into differential network analysis of gene expression data. *Scientific Reports*, *9*(1), 1–12. https://doi.org/10.1038/s41598-019-41918-3

Harwood, N. E., & Batista, F. D. (2011). The cytoskeleton coordinates the early events of B-cell activation. *Cold Spring Harbor Perspectives in Biology*, *3*(2), 1–15. https://doi.org/10.1101/cshperspect.a002360

Hauser, S. L., & Cree, B. A. C. (2020). Treatment of Multiple Sclerosis: A Review. *American Journal of Medicine*, *133*(12), 1380-1390.e2. https://doi.org/10.1016/j.amjmed.2020.05.049

He, H., Cao, S., Zhang, J., Shen, H., Wang, Y.-P., & Deng, H. (2019). A Statistical Test for Differential Network Analysis Based on Inference of Gaussian Graphical Model. *Scientific Reports*, *9*(1), 10863. https://doi.org/10.1038/s41598-019-47362-7

Hedström, A. K., Bomfim, I. L., Barcellos, L., Gianfrancesco, M., Schaefer, C., Kockum, I., … Alfredsson, L. (2014). Interaction between adolescent obesity and HLA risk genes in the etiology of multiple sclerosis. *Neurology*, *82*(10), 865–872. https://doi.org/10.1212/WNL.0000000000000203

Hedström, A. K., Sundqvist, E., Bäärnhielm, M., Nordin, N., Hillert, J., Kockum, I., … Alfredsson, L. (2011). Smoking and two human leukocyte antigen genes interact to increase the risk for multiple sclerosis. *Brain*, *134*(3), 653–664. https://doi.org/10.1093/brain/awq371

Hiltensperger, M., & Korn, T. (2018, January 1). The interleukin (IL)-23/T helper (Th)17 axis in experimental autoimmune encephalomyelitis and multiple sclerosis. *Cold Spring Harbor Perspectives in Medicine*. Cold Spring Harbor Laboratory Press. https://doi.org/10.1101/cshperspect.a029637

Hollenbach, J.A., & Oksenberg, J. R. (2016). The Immunogenetics of Multiple Sclerosis: A Comprehensive Review Jill. *Physiology & Behavior*, *176*(1), 139–148. https://doi.org/10.1016/j.jaut.2015.06.010.The

Hollenbach, Jill A., & Oksenberg, J. R. (2015, November 1). The immunogenetics of multiple sclerosis: A comprehensive review. *Journal of Autoimmunity*. Academic Press. https://doi.org/10.1016/j.jaut.2015.06.010

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure A Simple Sequentially Rejective Multiple Test Procedure. *Source: Scandinavian Journal of*

Statistics, *6*(2), 65–70. https://doi.org/10.2307/4615733

Hoppenbrouwers, I. A., Aulchenko, Y. S., Ebers, G. C., Ramagopalan, S. V., Oostra, B. A., van Duijn, C. M., & Hintzen, R. Q. (2008). EVI5 is a risk gene for multiple sclerosis. *Genes and Immunity*, *9*(4), 334–337. https://doi.org/10.1038/gene.2008.22

Hoppmann, N., Graetz, C., Paterka, M., Poisa-Beiro, L., Larochelle, C., Hasan, M., … Siffrin, V. (2015). New candidates for CD4 T cell pathogenicity in experimental neuroinflammation and multiple sclerosis. *Brain*, *138*(4), 902–917. https://doi.org/10.1093/brain/awu408

Hotelling, H. (1953). New Light on the Correlation Coefficient and its Transforms. *Source Journal of the Royal Statistical Society. Series B (Methodological) Journal of the Royal Statistical Society. Series B*, *15*(2), 193–232. Retrieved from http://www.jstor.org/stable/2983768

Hu, J. X., Zhao, H., & Zhou, H. H. (2010). False discovery rate control with groups. *Journal of the American Statistical Association*, *105*(491), 1215–1227. https://doi.org/10.1198/jasa.2010.tm09329

Huber, W., Von Heydebreck, A., Sültmann, H., Poustka, A., & Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. In *Bioinformatics* (Vol. 18). Oxford University Press. https://doi.org/10.1093/bioinformatics/18.suppl_1.S96

IMSGC. (2007). Risk Alleles for Multiple Sclerosis Identified by a Genomewide Study. *New England Journal of Medicine*, *357*(9), 851–862. https://doi.org/10.1056/nejmoa073493

IMSGC. (2013). Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature Genetics*, *45*(11), 1353–1362. https://doi.org/10.1038/ng.2770

IMSGC. (2019a). A systems biology approach uncovers cell-specific gene regulatory effects of genetic associations in multiple sclerosis. *Nature Communications*, *10*(1), 2236. https://doi.org/10.1038/s41467-019-09773-y

IMSGC. (2019b). Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science*, *365*(6460). https://doi.org/10.1126/science.aav7188

IMSGC, & WTCCC. (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, *476*(7359), 214–219. https://doi.org/10.1038/nature10251

International Multiple Sclerosis Genetics Consortium (IMSGC), I. M. S. G. C., Beecham, A. H., Patsopoulos, N. A., Xifara, D. K., Davis, M. F., Kemppinen, A., … McCauley, J. L. (2013). Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature Genetics*, *45*(11), 1353–1360. https://doi.org/10.1038/ng.2770

Jäger, A., Dardalhon, V., Sobel, R. A., Bettelli, E., & Kuchroo, V. K. (2009). Th1, Th17, and Th9 Effector Cells Induce Experimental Autoimmune Encephalomyelitis with Different Pathological Phenotypes. *The Journal of Immunology*, *183*(11), 7169–7177. https://doi.org/10.4049/jimmunol.0901906

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2000). *An introduction to Statistical Learning*. *Springer*. https://doi.org/10.1007/978-1-4614-7138-7

Jarius, S., Paul, F., Weinshenker, B. G., Levy, M., Kim, H. J., & Wildemann, B. (2020). Neuromyelitis optica. *Nature Reviews Disease Primers*, *6*(1). https://doi.org/10.1038/s41572-020-0214-9

Jeffreys, H. (1961). *Theory of Probability*. *Oxford University Press* (Vol. 15). https://doi.org/10.1063/1.3057804

Jelcic, I., Nimer, F. Al, Wang, J., Piehl, F., Sospedra, M., Martin, R., … Madjovski, A. (2018). Memory B Cells Activate Brain-Homing , Autoreactive Article Memory B Cells Activate Brain-Homing , Autoreactive CD4 + T Cells in Multiple Sclerosis, 1–16. https://doi.org/10.1016/j.cell.2018.08.011

Johnen, A., Bürkner, P. C., Landmeyer, N. C., Ambrosius, B., Calabrese, P., Motte, J., … Ziemann, U. (2019). Can we predict cognitive decline after initial diagnosis of multiple sclerosis? Results from the German National early MS cohort (KKNMS). *Journal of Neurology*, *266*(2), 386–397. https://doi.org/10.1007/s00415-018-9142-y

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118–127. https://doi.org/10.1093/biostatistics/kxj037

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., & Tanabe, M. (2020). KEGG: integrating viruses and cellular organisms. *Nucleic Acids Research*, *49*(D1), D545–D551. https://doi.org/10.1093/nar/gkaa970

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kim, K., Pröbstel, A. K., Baumann, R., Dyckow, J., Landefeld, J., Kogl, E., … Baranzini, S. E. (2021). Cell type-specific transcriptomics identifies neddylation as a novel therapeutic target in multiple sclerosis. *Brain : A Journal of Neurology*, *144*(2), 450–461. https://doi.org/10.1093/brain/awaa421

Kimura, K. (2020). Regulatory T cells in multiple sclerosis. *Clinical and Experimental Neuroimmunology*, *11*(3), 148–155. https://doi.org/10.1111/cen3.12591

Kingwell, E., Marriott, J. J., Jetté, N., Pringsheim, T., Makhani, N., Morrow, S. A., … Marrie, R. A. (2013). Incidence and prevalence of multiple sclerosis in Europe: A systematic review. *BMC Neurology*, *13*. https://doi.org/10.1186/1471-2377-13-128

Krishnamoorthy, G., Holz, A., & Wekerle, H. (2007, November 14). Experimental models of spontaneous autoimmune disease in the central nervous system. *Journal of Molecular Medicine*. Springer. https://doi.org/10.1007/s00109-007-0218-x

Krishnamoorthy, G., Lassmann, H., Wekerle, H., & Holz, A. (2006). Spontaneous opticospinal encephalomyelitis in a double-transgenic mouse model of autoimmune T cell/B cell cooperation. *Journal of Clinical Investigation*, *116*(9), 2385–2392. https://doi.org/10.1172/JCI28330

Krumsiek, J., Suhre, K., Illig, T., Adamski, J., & Theis, F. J. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Systems Biology*, *5*(1), 21. https://doi.org/10.1186/1752-0509-5-21

Lacy, P. (2006). Mechanisms of Degranulation in Neutrophils. *Allergy, Asthma & Clinical Immunology*, *2*(3), 98. https://doi.org/10.1186/1710-1492-2-3-98

Lam, T. H., Shen, M., Tay, M. Z., & Ren, E. C. (2017). Unique allelic eQTL clusters in human MHC haplotypes. *G3: Genes, Genomes, Genetics*, *7*(8), 2595–2604. https://doi.org/10.1534/g3.117.043828

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 559. https://doi.org/10.1186/1471-2105-9-559

Langfelder, P., & Horvath, S. (2012). Fast R functions for robust correlations and hierarchical clustering. *Journal of Statistical Software*, *46*(11), 1–17. https://doi.org/10.18637/jss.v046.i11

Langfelder, P., & Horvath, S. (2017). WGCNA package: Frequently Asked Questions. Retrieved May 17, 2021, from https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/faq.html

Langfelder, P., Luo, R., Oldham, M. C., & Horvath, S. (2011). Is My Network Module Preserved and Reproducible? *PLoS Computational Biology*, *7*(1), e1001057. https://doi.org/10.1371/journal.pcbi.1001057

Langfelder, P., Zhang, B., & Horvath, S. (2009). Dynamic Tree Cut: in-depth description, tests and applications 1 Why Dynamic Tree Cut? Retrieved from http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/BranchCutting.

Lassmann, H., Van Horssen, J., & Mahad, D. (2012). Progressive multiple sclerosis: Pathology and pathogenesis. *Nature Reviews Neurology*. Nature Publishing Group. https://doi.org/10.1038/nrneurol.2012.168

Ledoit, O., & Wolf, M. (2003). Honey, I Shrunk the Sample Covariance Matrix.

Lehmann Horn, K., Kronsbein, H. C., & Weber, M. S. (2013). Targeting B cells in the treatment of multiple sclerosis: Recent advances and remaining challenges. *Therapeutic Advances in Neurological Disorders*, *6*(3), 161–173. https://doi.org/10.1177/1756285612474333

Leonardson, A. S., Zhu, J., Chen, Y., Wang, K., Lamb, J. R., Reitman, M., … Schadt, E. E. (2009). The effect of food intake on gene expression in human peripheral blood. *Human Molecular Genetics*, *19*(1), 159–169. https://doi.org/10.1093/hmg/ddp476

Levin, L. I., Munger, K. L., Rubertone, M. V., Peck, C. A., Lennette, E. T., Spiegelman, D., & Ascherio, A. (2005). Temporal relationship between elevation of Epstein-Barr virus antibody titers and initial onset of neurological symptoms in multiple sclerosis. *Journal of the American Medical Association*, *293*(20), 2496–2500. https://doi.org/10.1001/jama.293.20.2496

Liang, J. W., Fang, Z. Y., Huang, Y., Liuyang, Z. Y., Zhang, X. L., Wang, J. L., … Liu, R. (2018). Application of Weighted Gene Co-Expression Network Analysis to Explore the Key Genes in Alzheimer's Disease. *Journal of Alzheimer's Disease*, *65*(4), 1353–1364. https://doi.org/10.3233/JAD-180400

Liao, X., Buchberg, A. M., Jenkins, N. A., & Copeland, N. G. (1995). Evi-5, a common site of retroviral integration in AKXD T-cell lymphomas, maps near Gfi-1 on mouse chromosome 5. *Journal of Virology*, *69*(11), 7132–7137. https://doi.org/10.1128/jvi.69.11.7132-7137.1995

Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., & Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Research*, *47*(W1), W199–W205. https://doi.org/10.1093/nar/gkz401

Macqueen, J. (1967). *SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS. Proc. Fifth Berkeley Symp. on Math. Statist. and Prob*. University of California Press.

Maier, L. M., Lowe, C. E., Cooper, J., Downes, K., & Anderson, D. E. (2009). IL2RA Genetic Heterogeneity in Multiple Sclerosis and Type 1 Diabetes Susceptibility and Soluble Interleukin-2 Receptor Production. *PLoS Genet*, *5*(1), 1000322. https://doi.org/10.1371/journal.pgen.1000322

Mathewson, N. D., Fujiwara, H., Wu, S. R., Toubai, T., Oravecz-Wilson, K., Sun, Y., … Reddy, P. (2016). SAG/Rbx2-Dependent Neddylation Regulates T-Cell Responses. *American Journal of Pathology*, *186*(10), 2679–2691. https://doi.org/10.1016/j.ajpath.2016.06.014

Menezes, S. M., Decanine, D., Brassat, D., Khouri, R., Schnitman, S. V., Kruschewsky, R., … Weyenbergh, J. V. (2014). CD80+ and CD86+ B cells as biomarkers and possible therapeutic targets in HTLV-1 associated myelopathy/tropical spastic paraparesis and multiple sclerosis. *Journal of Neuroinflammation*, *11*(1), 1–15. https://doi.org/10.1186/1742-2094-11-18

Merrill, J. E., Kong, D. H., Clayton, J., Ando, D. G., Hinton, D. R., & Hofman, F. M.

(1992). Inflammatory leukocytes and cytokines in the peptide-induced disease of experimental allergic encephalomyelitis in SJL and B10.PL mice. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(2), 574–578. https://doi.org/10.1073/pnas.89.2.574

Miller, D., Barkhof, F., Montalban, X., Thompson, A., & Filippi, M. (2005). Clinically isolated syndromes suggestive of multiple sclerosis, part I: Natural history, pathogenesis, diagnosis, and prognosis. *Lancet Neurology*, *4*(5), 281–288. https://doi.org/10.1016/S1474-4422(05)70071-5

Miller, J. A., Horvath, S., & Geschwind, D. H. (2010). Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(28), 12698–12703. https://doi.org/10.1073/pnas.0914257107

Miller, S. D., Karpus, W. J., & Davidson, T. S. (2010). Experimental autoimmune encephalomyelitis in the mouse. *Current Protocols in Immunology*. NIH Public Access. https://doi.org/10.1002/0471142735.im1501s77

Minagar, A., Ma, W., Zhang, X., Wang, X., Zhang, K., Steven Alexander, J., … Albitar, M. (2012). Plasma ubiquitin-proteasome system profile in patients with multiple sclerosis: Correlation with clinical features, neuroimaging, and treatment with interferon-beta-1b. *Neurological Research*, *34*(6), 611–618. https://doi.org/10.1179/1743132812Y.0000000055

Mokry, L. E., Ross, S., Ahmad, O. S., Forgetta, V., Smith, G. D., Leong, A., … Richards, J. B. (2015). Vitamin D and Risk of Multiple Sclerosis: A Mendelian Randomization Study. *PLoS Medicine*, *12*(8), 1–20. https://doi.org/10.1371/journal.pmed.1001866

Molnarfi, N., Schulze-Topphoff, U., Weber, M. S., Patarroyo, J. C., Prod’homme, T., Varrin-Doyer, M., … Zamvil, S. S. (2013). MHC class II-dependent B cell APC function is required for induction of CNS autoimmunity independent of myelin-specific antibodies. *Journal of Experimental Medicine*, *210*(13), 2921–2937. https://doi.org/10.1084/jem.20130699

Moutsianas, L., Jostins, L., Beecham, A. H., Dilthey, A. T., Xifara, D. K., Ban, M., … McVean, G. (2015). Class II HLA interactions modulate genetic risk for multiple sclerosis. *Nature Genetics*, *47*(10), 1107–1113. https://doi.org/10.1038/ng.3395

Myers, L., & Sirois, M. J. (2014). Spearman Correlation Coefficients, Differences between. *Wiley StatsRef: Statistics Reference Online*, 1–2. https://doi.org/10.1002/9781118445112.stat02802

Nickles, D., Chen, H. P., Li, M. M., Khankhanian, P., Madireddy, L., Caillier, S. J., … Baranzini, S. E. (2013). Blood RNA profiling in a large cohort of multiple sclerosis patients and healthy controls. *Human Molecular Genetics*, *22*(20), 4194–4205. https://doi.org/10.1093/hmg/ddt267

O’Gorman, C., Lin, R., Stankovich, J., & Broadley, S. A. (2013). Modelling Genetic Susceptibility to Multiple Sclerosis with Family Data. *Neuroepidemiology*, *40*(1), 1–12. https://doi.org/10.1159/000341902

O’Riordan, J. I., Gallagher, H. L., Thompson, A. J., Howard, R. S., Kingsley, D. P. E., Thompson, E. J., … Miller, D. H. (1996). Clinical, CSF, and MRI findings in Devic’s neuromyelitis optica. *Journal of Neurology Neurosurgery and Psychiatry*, *60*(4), 382–387. https://doi.org/10.1136/jnnp.60.4.382

Ofengeim, D., Ito, Y., Najafov, A., Zhang, Y., Shan, B., DeWitt, J. P., … Yuan, J. (2015). Activation of necroptosis in multiple sclerosis. *Cell Reports*, *10*(11), 1836–1849. https://doi.org/10.1016/j.celrep.2015.02.051

Paolillo, A., Coles, A. J., Molyneux, P. D., Gawne-Cain, M., MacManus, D., Barker, G. J., … Miller, D. H. (1999). Quantitative MRI in patients with secondary progressive

MS treated with monoclonal antibody Campath 1H. *Neurology*, *53*(4), 751–757. https://doi.org/10.1212/wnl.53.4.751

Patsopoulos, N. A. (2018). Genetics of multiple sclerosis: An overview and new directions. *Cold Spring Harbor Perspectives in Medicine*, *8*(7), 1–12. https://doi.org/10.1101/cshperspect.a028951

Patsopoulos, N. A., Barcellos, L. F., Hintzen, R. Q., Schaefer, C., van Duijn, C. M., Noble, J. A., … W de Bakker, P. I. (2013). Fine-Mapping the Genetic Association of the Major Histocompatibility Complex in Multiple Sclerosis: HLA and Non-HLA Effects. *PLOS Genetics | Www.Plosgenetics.Org 1*, *9*(11). https://doi.org/10.1371/journal.pgen.1003926

Patsopoulos, N. A., & De Bakker, I. W. (2011). Genomewide meta-analysis identifies novel multiple sclerosis susceptibility loci. *Ann Neurol*, *70*(6), 897–912. https://doi.org/10.1002/ana.22609

Pineda, S., Gomez-Rubio, P., Picornell, A., Bessonov, K., Márquez, M., Kogevinas, M., … Malats, N. (2015). Framework for the Integration of Genomics, Epigenomics and Transcriptomics in Complex Diseases. *Human Heredity*, *79*(3–4), 124–136. https://doi.org/10.1159/000381184

R Core Team. (2020). R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabási, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, *297*(5586), 1551–1555. https://doi.org/10.1126/science.1073374

Razmara, E., & Garshasbi, M. (2018). Whole-exome sequencing identifies R1279X of MYH6 gene to be associated with congenital heart disease. *BMC Cardiovascular Disorders*, *18*(1), 137. https://doi.org/10.1186/s12872-018-0867-4

Reich, D. S., Lucchinetti, C. F., & Calabresi, P. A. (2018). Multiple Sclerosis. *New England Journal of Medicine*, *378*(2), 169–180. https://doi.org/10.1056/NEJMra1401483

Reverter, A., Ingham, A., Lehnert, S. A., Tan, S.-H., Wang, Y., Ratnakumar, A., & Dalrymple, B. P. (2006). Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer, *22*(19), 2396–2404. https://doi.org/10.1093/bioinformatics/btl392

Rhead, B., Bäärnhielm, M., Gianfrancesco, M., Mok, A., Shao, X., Quach, H., … Barcellos, L. F. (2016). Mendelian randomization shows a causal effect of low Vitamin D on multiple sclerosis risk. *Neurology: Genetics*, *2*(5). https://doi.org/10.1212/NXG.0000000000000097

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Rivers, T. M., Sprunt, D. H., & Berry, G. P. (1933). Observations on attempts to produce acute disseminated encephalomyelitis in monkeys. *Journal of Experimental Medicine*, *58*(1), 39–52. https://doi.org/10.1084/jem.58.1.39

Rölle, A., Jäger, D., & Momburg, F. (2018, October 17). HLA-E peptide repertoire and dimorphism - Centerpieces in the adaptive NK cell puzzle? *Frontiers in Immunology*. Frontiers Media S.A. https://doi.org/10.3389/fimmu.2018.02410

Rollins, B., Martin, M. V., Morgan, L., & Vawter, M. P. (2010). Analysis of whole genome biomarker expression in blood and brain. *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, *153*(4), 919–936. https://doi.org/10.1002/ajmg.b.31062

Sabatino, J. J., Pröbstel, A. K., & Zamvil, S. S. (2019). B cells in autoimmune and neurodegenerative central nervous system diseases. *Nature Reviews Neuroscience*, *20*(12), 728–745. https://doi.org/10.1038/s41583-019-0233-2

Schäfer, J., & Strimmer, K. (2005a). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, *4*, Article32. https://doi.org/10.2202/1544-6115.1175

Schäfer, J., & Strimmer, K. (2005b). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, *21*(6), 754–764. https://doi.org/10.1093/bioinformatics/bti062

Schäfer, J., & Strimmer, K. (2005c). Learning large-scale graphical Gaussian models from genomic data. *AIP Conference Proceedings*, *776*(2005), 263–276. https://doi.org/10.1063/1.1985393

Seder, R. A., & Ahmed, R. (2003, September 1). Similarities and differences in CD4+ and CD8+ effector and memory T cell generation. *Nature Immunology*. Nat Immunol. https://doi.org/10.1038/ni969

Sidore, C., Orrù, V., Cocco, E., Steri, M., Inshaw, J. R., Pitzalis, M., … Zoledziewska, M. (2020). PRF1 mutation alters immune system activation, inflammation, and risk of autoimmunity. *Multiple Sclerosis Journal*, 135245852096393. https://doi.org/10.1177/1352458520963937

Smyth, G. K. (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, *3*(1).

Smyth, G. K., Michaud, J., & Scott, H. S. (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, *21*(9), 2067–2075. https://doi.org/10.1093/bioinformatics/bti270

Stromnes, I. M., Cerretti, L. M., Liggitt, D., Harris, R. A., & Goverman, J. M. (2008). Differential regulation of central nervous system autoimmunity by T H1 and TH17 cells. *Nature Medicine*, *14*(3), 337–342. https://doi.org/10.1038/nm1715

Takaba, H., & Takayanagi, H. (2017). The Mechanisms of T Cell Selection in the Thymus. *Trends in Immunology*, *38*(11), 805–816. https://doi.org/10.1016/j.it.2017.07.010

Tang, J., Kong, D., Cui, Q., Wang, K., Zhang, D., Gong, Y., & Wu, G. (2018). Prognostic genes of breast cancer identified by gene co-expression network analysis. *Frontiers in Oncology*, *8*(SEP), 374. https://doi.org/10.3389/fonc.2018.00374

The Gene Ontology Consortium. (2020). The Gene Ontology resource: enriching a GOld mine. Retrieved January 11, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7779012/

Thompson, A. J., Baranzini, S. E., Geurts, J., Hemmer, B., & Ciccarelli, O. (2018). Multiple sclerosis. *The Lancet*, *391*(10130), 1622–1636. https://doi.org/10.1016/S0140-6736(18)30481-1

Trinh Do, K., Kastenmu, G., Mook-Kanamori, D. O., Yousri, N. A., Theis, F. J., Suhre, K., & Krumsiek, J. (2014). Network-Based Approach for Analyzing Intra-and Interfluid Metabolite Associations in Human Blood, Urine, and Saliva. https://doi.org/10.1021/pr501130a

Van Oosten, B. W., Lai, M., Hodgkinson, S., Barkhof, F., Miller, D. H., Moseley, I. F., … Polman, C. H. (1997). Treatment of multiple sclerosis with the monoclonal anti-CD4 antibody cM-T412: Results of a randomized, double-blind, placebo-controlled, MR- monitored phase II trial. *Neurology*, *49*(2), 351–357. https://doi.org/10.1212/WNL.49.2.351

von Bismarck, O., Dankowski, T., Ambrosius, B., Hessler, N., Antony, G., Ziegler, A., … Salmen, A. (2018). Treatment choices and neuropsychological symptoms of a large cohort of early MS. *Neurology - Neuroimmunology Neuroinflammation*, *5*(3), e446. https://doi.org/10.1212/nxi.0000000000000446

Wei, T., & Simko, V. (2021). R package "corrplot": Visualization of a Correlation Matrix.

Welsh, R. A., & Sadegh-Nasseri, S. (2020). The love and hate relationship of HLA-DM/DO in the selection of immunodominant epitopes. *Current Opinion in Immunology*, *64*(Mhc Ii), 117–123. https://doi.org/10.1016/j.coi.2020.05.007

Welsh, R. A., Song, N., Foss, C. A., Boronina, T., Cole, R. N., & Sadegh-Nasseri, S. (2020). Lack of the MHC class II chaperone H2-O causes susceptibility to autoimmune diseases. *PLoS Biology*, *18*(2), e3000590. https://doi.org/10.1371/journal.pbio.3000590

Whittaker, J. (1990). Graphical Models in Applied Multivariate Statistics | Wiley. Retrieved December 18, 2020, from https://www.wiley.com/en-gb/Graphical+Models+in+Applied+Multivariate+Statistics-p-9780471917502

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York*, (978-3-319-24277–4).

Wittenberg, G. M., Greene, J., Vértes, P. E., Drevets, W. C., & Bullmore, E. T. (2020). Archival Report Major Depressive Disorder Is Associated With Differential Expression of Innate Immune and Neutrophil-Related Gene Networks in Peripheral Blood: A Quantitative Review of Whole-Genome Transcriptional Data From Case-Control Studies. https://doi.org/10.1016/j.biopsych.2020.05.006

Wolfe, C. J., Kohane, I. S., & Butte, A. J. (2005). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, *6*, 1–10. https://doi.org/10.1186/1471-2105-6-227

Wong, P., & Pamer, E. G. (2003). CD8 T cell responses to infectious pathogens. *Annual Review of Immunology*. Annu Rev Immunol. https://doi.org/10.1146/annurev.immunol.21.120601.141114

Woodberry, T., Bouffler, S., Wilson, A., Buckland, R., & Brüstle, A. (2018). The Emerging Role of Neutrophil Granulocytes in Multiple Sclerosis. *Journal of Clinical Medicine*, *7*(12), 511. https://doi.org/10.3390/jcm7120511

Wucherpfennig, K. W., & Sethi, D. (2011, April). T cell receptor recognition of self and foreign antigens in the induction of autoimmunity. *Seminars in Immunology*. NIH Public Access. https://doi.org/10.1016/j.smim.2011.01.007

Yajima, T., & Knowlton, K. U. (2009). Viral myocarditis from the perspective of the virus. *Circulation*, *119*(19), 2615–2624. https://doi.org/10.1161/CIRCULATIONAHA.108.766022

Yamamoto, F., Suzuki, S., Mizutani, A., Shigenari, A., Ito, S., Kametani, Y., … Shiina, T. (2020). Capturing Differential Allele-Level Expression and Genotypes of All Classical HLA Loci and Haplotypes by a New Capture RNA-Seq Method. *Frontiers in Immunology*, *11*(May), 1–14. https://doi.org/10.3389/fimmu.2020.00941

Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N., & Liang, H. (2014). Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature Communications*, *5*(1), 1–9. https://doi.org/10.1038/ncomms4231

Zannas, A. S., Jia, M., Hafner, K., Baumert, J., Wiechmann, T., Pape, J. C., … Binder, E. B. (2019). Epigenetic upregulation of FKBP5 by aging and stress contributes to NF-κB-driven inflammation and cardiovascular risk. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(23), 11370–11379.

https://doi.org/10.1073/pnas.1816847116

Zhang, B., & Horvath, S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*(1). https://doi.org/10.2202/1544-6115.1128

Zhang, J., Zheng, H., Li, Y., Li, H., Liu, X., Qin, H., … Wang, D. (2016). Coexpression network analysis of the genes regulated by two types of resistance responses to powdery mildew in wheat. *Scientific Reports*, *6*(1), 1–15. https://doi.org/10.1038/srep23805

Zhou, Y., Zhu, G., Charlesworth, J. C., Simpson, S., Rubicz, R., Göring, H. H. H., … Taylor, B. V. (2016). Genetic loci for Epstein-Barr virus nuclear antigen-1 are associated with risk of multiple sclerosis. *Multiple Sclerosis*, *22*(13), 1655–1664. https://doi.org/10.1177/1352458515626598