Understanding the Genetic Nature of Multiple Sclerosis Using Next-Generation Sequencing Genomic Analysis Methods

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Fahad Moqbel Almsned
Master of Advanced Studies
University of California – San Diego, 2013
Bachelor of Medicine, Bachelor of Surgery
King Faisal University, 2007

Director: M. Saleet Jafri, Professor
School of Systems Biology – Department of Bioinformatics and Computational Biology

Fall Semester 2020
George Mason University
Fairfax, VA

ACKNOWLEDGEMENTS

This dissertation presents my research results during the time I worked within making my advisor, Professor M. Saleet Jafri, and the end of my journey in obtaining my Ph.D. So many people have supported me in making this possible and an unforgettable experience. First and foremost, I want to express my sincere thanks to Saleet for his continuous support since my first days at Mason. I have learned a lot from working with him. Saleet has given me opportunities to develop my skills and be involved in multiple projects.

I want to expand my thanks to Professors Robert Lipsky, Iosif Vaisman, Doctors Nadine Kabbani, Aman Ullah, and Fayaz Seifuddin for their helpful comments dissertation and my defense. I also want to thank Diane St. Germain, Andrea Nikoi, and Kimberly L Harris for providing information and assistance during the years I was a student at the Department of Bioinformatics and Computational Biology. My five years at George Mason has been a great time thanks to my friends Aman Ullah, Roshan, Nasrin, Lamya, Bader, Raquel, and Leena.

Finally, I would like to dedicate this dissertation to my family, who have always been with me, educating me and giving me encouragement and inspiration.

Fahad Moqbel Almsned
George Mason University
November 2020

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Multiple Sclerosis ..... MS
Central Nervous System ..... CNS
Ribonucleic Acid Sequencing ..... RNA-Seq
Healthy Controls ..... HC
Gene Set Analysis ..... GSA
Major Histocompatibility Complex ..... MHC
Human Leukocyte Antigen ..... HLA
Interleukins IL
Over-Representation Analysis ..... ORA
Functional Class Scoring ..... FCS
Pathway Topology ..... PT
Kyoto Encyclopedia of Genes and Genomes ..... KEGG
Geno Ontology ..... GO
Principal Component Analysis ..... PCA
Differentially Expressed Genes ..... DEGs
Transcription Factor Binding Sites TFBS ..... TFBS
Passing Attributes between Networks for Data Assimilation ..... PANDA
Protein-Protein Interaction ..... PPI
Pearson correlation coefficients ..... PCCs
Experimental Autoimmune Encephalomyelitis ..... EAE
Peripheral Blood Mononuclear Cells ..... PBMCs
Transcription Factor ..... TF
Genome-Wide Association Studies ..... GWAS
Single Nucleotide Polymorphisms ..... SNPs
Expression Quantitative Trait Loci ..... eQTL

ABSTRACT

UNDERSTANDING THE GENETIC NATURE OF MULTIPLE SCLEROSIS USING NEXT-GENERATION SEQUENCING GENOMIC ANALYSIS METHODS

Fahad Moqbel Almsned, Ph.D.

George Mason University, 2020

Dissertation Director: Dr. M. Saleet Jafri, Professor

Multiple Sclerosis (MS) is an incapacitating neurological illness, where changes in gene expression play a crucial role. Affecting nearly two million people worldwide, MS is the most common acquired neurological disorder of young adults just after physical trauma. Up to now, an understanding of the complex molecular mechanism of MS, which is vital to develop effective therapies, has remained elusive. Most of the studies that have been conducted to address this problem have used microarray technology, which does not reflect the high variability of protein expression. The primary goal of this work was to analyze the molecular interactions and possible sequence variants underlying the pathogenesis of Multiple Sclerosis (MS) by utilizing RNA-Seq expression data, which is also capable of catching the high variability in protein expression associated with MS pathology. Results from this study will deliver a better understanding of the complex molecular mechanisms underlying MS and, hopefully, provide a groundwork for effective therapeutics. At the end of the study I ended up with a list of candidate genes, among them Transcriptions Factors, and Single Nucleotide Polymorphisms with potential implications in MS. Future studies will need to incorporate

more metadata and biological replicates in the analysis. Experimental validation will also required.

CHAPTER ONE: INTRODUCTION

Multiple Sclerosis

Multiple Sclerosis (MS) is an immune-mediated inflammatory disorder damaging fatty myelin sheaths around the axons of the Central Nervous System (CNS) leading to a broad spectrum of clinical signs and symptoms.[1] MS is the second most common acquired neurological disorder of young adults, with physical trauma being the most common. The illness demonstrations a range of severity, fluctuating from an asymptomatic pathological process to severe disabling illness. The clinical presentation involves two forms, relapsing disease in which distinct attacks with clinical stability in between, or progressive condition in which gradual worsening of neurological deficits.

Many factors are believed to contribute to the source of MS, including genetic susceptibility and environmental factors. MS affects mainly young people between the ages of 15 and 50 years, with a peak onset at about age 30. There is a substantial gender preference; most MS patients (70-75%) are women.[2]

The incidence and prevalence of MS vary throughout the world. MS affects nearly two million people worldwide with evident variability in geographic distribution.[3] Near the equator, typically in tropical regions, there is low risk, while MS risk north and south of the equator increases with higher latitudes, in both northern and southern hemispheres.[4]

Though the pathogenesis of MS is ill-understood, evidence suggests that both genetic and environmental components play essential roles in disease development, both independently and interactively.[4] The role of genetics in MS and its interaction with environmental causes have been extensively studied. Environmental factors have historically been thought to be necessary to disease risk. The geographical distribution and familial aggregation of MS have often been credited to the rule of infectious agents, but there is no consensus regarding this theory.[5] A Canadian study examined a population-based sample of 15,000 individuals with MS using standardized, personally administered questionnaires to identify adoptees and those who had adopted relatives. The rate of MS among first-degree, non-biological relatives living with the index case was no higher than the expected rate from the Canadian population prevalence data and was significantly less than the rate for biological relatives. These findings support the hypothesis that the familial aggregation of MS is genetically determined rather than environmentally determined.[5,6] A significant contributor to the genetic risk is the major histocompatibility complex (MHC) antigen.[7]

Several reports showing familial aggregation of the disorder, high concordance rates between twins, and more significant risk among relatives of patients with MS are supporting the contribution of genetics to MS. People with MS have a 5–26% chance of having one or more affected relatives,[5,6] which is a much higher chance than one would expect for a disease with no genetic component. Furthermore, the relative risk of MS for identical twins, if one is affected, is approximately 200 to 300 times greater than that of

the general population.[6,8] Lastly, the first-degree relatives of MS patients have a 2–5% risk of also developing the disease.[8]

Evidence suggests that transcription factors play a role in the pathogenesis of MS and other autoimmune diseases.[9] For example, it has previously been observed that members of the NF-kappaB, STAT, AP-1, and E2F families,[10] IRF-1,[11] IRF-2,[12] IRF-5,[13] IRF-8, [11] CREB,[14] PPARgamma and PPARalpha,[15,16] SP1,[13] SP3,[17] RORC,[18] NR4A2, TCF2,[19] ETS-1,[20] and FOXP3[21] may be implicated in MS and its disease subtypes.

Previous studies identified several alleles in immune function as heritable risk factors for MS. Genetic complexity, primarily related to Human Leukocyte Antigens (HLA) of the MHC and, to a lesser extent, non-MHC-related genes, plays a significant role in influencing disease susceptibility, phenotypic expression CD4 T cells, which have leading rule in MS pathogenesis[22], experience profound changes in gene expression during the initial hours after activation. Co-stimulation via the CD28 receptor is essential for the effective activation of naive T cells.[23] Pre-clinical studies showed that the transcription factor is highly induced in a CD28-dependent manner upon T cell activation and is involved in essential CD4 effector T cell functions, participating in the regulation of several T cell activation pathways, together with a large group of CD28-regulated genes.[24] Furthermore, Levels of blood monocytes secreting IL-6 and IL-12 were higher in patients with untreated MS and other neurological diseases compared to healthy controls.[25] MS patients' blood monocytes also displayed elevated mean fluorescence intensity for the co-stimulatory molecule CD86, and MS patients with longer disease

duration (>10 years) and higher disease severity had higher percentages of CD80 expressing monocytes compared to patients with short duration or lower severity.[25]

Moleuclar Approaches applied to study MS: Microarrays and RNA-Seq Technology

Since its launch, RNA-Seq has been compared to microarray technology as a means of generating transcriptome information. Both follow a parallel path to answering a biological question. Nevertheless, there are a few key advantages of RNA-Seq technology.[26] First, using microarray technology limits the researcher in spotting transcripts that linked to existing genomic sequencing information. RNA-Seq experiments, instead, work well for examining both known transcripts and explore new ones.[26] Second, RNA-seq delivers a low background signal because DNA sequences can be unambiguously mapped to unique regions of the genome. As a result, noise in the experiment is effortlessly eliminated during analysis. Hybridization issues seen with microarrays, such as cross-hybridization or non-ideal hybridization kinetics, are also removed in RNA-Seq experiments, which offers another signal-to-noise advantage.[26] Finally, RNA-seq can quantify a broad dynamic range of expression levels, with absolute rather than relative values.[26]

CHAPTER TWO: TRANSCRIPTOMIC ANALYSIS OF MULTIPLE SCLEROSIS MONOCYTES BY RNA-SEQUENCING

Abstract

Multiple Sclerosis (MS) is an inflammatory disorder associated with immune abnormalities in the central nervous system, including the presence of many monocytes in MS lesions. Despite contributing to morbidity in neurological disorders, the molecular mechanisms of MS continue to remain poorly understood. This study aimed to investigate specific transcriptome changes occurring in monocytes of patients with MS compared to Healthy Controls (HC) patients to improve diagnosis and possible treatment of affected subjects. Unlike other studies, which use microarray technology, the transcriptome of all participants was studied by Ribonucleic Acid sequencing (RNA-Seq). The advantage of RNA-Seq is that it does not report high variability of protein expression, as seen with microarray studies. Data analysis revealed that 6120 genes were significantly altered between the two groups (16% up-regulated and 17% down-regulated in MS group compared to healthy controls). The KEGG hsa04062 Chemokine signaling pathway was the most significant up-regulated pathway in the functional scoring analysis. We offered candidate genes and pathways with potential implications in MS. Results from this study will provide the groundwork for the new therapy development of MS.

Background

Multiple Sclerosis

Multiple Sclerosis (MS) is an immune-mediated inflammatory disorder damaging fatty myelin sheaths around the axons of the Central Nervous System (CNS) leading to a broad spectrum of clinical signs and symptoms.[1] MS is the second most common acquired neurological disorder of young adults, with physical trauma being the most common. The illness demonstrations a range of severity, fluctuating from an asymptomatic pathological process to severe disabling illness. The clinical presentation involves two forms, relapsing disease in which distinct attacks with clinical stability in between, or progressive condition in which gradual worsening of neurological deficits.

Many factors are believed to contribute to the source of MS, including genetic susceptibility and environmental factors. MS affects mainly young people between the ages of 15 and 50 years, with a peak onset at about age 30. There is a substantial gender preference; most MS patients (70-75%) are women.[2]

The incidence and prevalence of MS vary throughout the world. MS affects nearly two million people worldwide with evident variability in geographic distribution.[3] Near the equator, typically in tropical regions, there is low risk, while MS risk north and south of the equator increases with higher latitudes, in both northern and southern hemispheres.[4]

Though the pathogenesis of MS is ill-understood, evidence suggests that both genetic and environmental components play essential roles in disease development, both independently and interactively.[4] The rule of genetics in MS and its interaction with

environmental causes are presently extensively studied. MS is a disease with evident geographic variability in both prevalence and incidence. The role of environmental factors has historically been thought to be necessary. The geographical distribution and familial aggregation of MS have often been credited to the rule of infectious agents, but there is no consensus regarding this theory.[5] A Canadian study examined a population-based sample of 15,000 individuals with MS using standardized, personally administered questionnaires to identify adoptees and those who had adopted relatives. The rate of MS among first-degree, non-biological relatives living with the index case was no higher than the expected rate from the Canadian population prevalence data and was significantly less than the rate for biological relatives. These findings support the hypothesis that the familial aggregation of MS is genetically determined rather than environmentally determined.[5,6] A significant contributor to the genetic risk is the major histocompatibility complex (MHC) antigen.[7]

Several reports showing familial aggregation of the disorder, high concordance rates between twins, and more significant risk among relatives of patients with MS are supporting the contribution of genetics to MS. People with MS have a 5–26% chance of having one or more affected relatives,[5,6] which is a much higher chance than one would expect for a disease with no genetic component. Furthermore, the relative risk of MS for identical twins, if one is affected, is approximately 200 to 300 times greater than that of the general population.[6,8] Lastly, the first-degree relatives of MS patients have a 2–5% risk of also developing the disease.[8]

Previous studies identified several alleles as heritable risk factors for MS. Genetic complexity, primarily related to Human Leukocyte Antigens (HLA) of the MHC and, to a lesser extent, non-MHC-related genes, plays a significant role in influencing disease susceptibility, phenotypic expression CD4 T cells, which have leading rule in MS pathogenesis[22], experience profound changes in gene expression during the initial hours after activation. Co-stimulation via the CD28 receptor is essential for the effective activation of naive T cells.[23] Pre-clinical studies showed that the transcription factor is highly induced in a CD28-dependent manner upon T cell activation and is involved in essential CD4 effector T cell functions, participating in the regulation of several T cell activation pathways, together with a large group of CD28-regulated genes.[24] Furthermore, Levels of blood monocytes secreting IL-6 and IL-12 were higher in patients with untreated MS and other neurological diseases compared to healthy controls.[25] MS patients blood monocytes also displayed elevated mean fluorescence intensity for the co-stimulatory molecule CD86, and MS patients with longer disease duration (>10 years) and higher disease severity had higher percentages of CD80 expressing monocytes compared to patients with short duration or lower severity.[25]

RNA-Seq Technology

Since its launch, RNA-Seq has been compared to microarray technology as a means of generating transcriptome information. Both follow a parallel path to answering a biological question. Nevertheless, there are a few key advantages of RNA-Seq technology.[26] First, using microarray technology limits the researcher in spotting transcripts that linked to existing genomic sequencing information. RNA-Seq

experiments, instead, work well for examining both known transcripts and explore new ones.[26] Second, RNA-seq delivers a low background signal because DNA sequences can be unambiguously mapped to unique regions of the genome. As a result, noise in the experiment is effortlessly eliminated during analysis. Hybridization issues seen with microarrays, such as cross-hybridization or non-ideal hybridization kinetics, are also removed in RNA-Seq experiments, which offers another signal-to-noise advantage.[26] Finally, RNA-seq can quantify a broad dynamic range of expression levels, with absolute rather than relative values.[26]

Gene Set Analysis

Methods such as high-throughput sequencing and gene/protein profiling have altered research by permitting wide-ranging monitoring of a biological system. [27] Regardless of the technology used, analysis of high-throughput data typically yields a list of differentially expressed genes or proteins, which fails to provide mechanistic insights into the underlying biology of the studied condition. One approach to reduces the complexity of analysis has been to simplify interpretation by long grouping lists of individual genes into smaller sets of related genes or proteins using a large number of knowledge bases to help with this task.[27] Investigating high-throughput molecular measurements at the functional level is very appealing for two reasons. First, grouping thousands of genes, proteins, and other biological molecules by the pathways they are involved in reducing the complexity to just several hundred pathways for the experiment. Second, identifying active pathways that differ between the two conditions can have more explanatory power than a simple list of different genes or proteins.[28] We can divide

knowledge base-driven pathway analysis into three classes: over-representation analysis (ORA), functional class scoring (FCS), and Pathway topology (PT)-based methods.[29] Comparatively, topological methods have shown better performance in the simulation scenarios with non-overlapping pathways. However, they were not conclusively better in other situations suggesting that a simple gene set approach might be enough to detect an enriched pathway under realistic circumstances.[29] Out of popular gene set analysis methods, Applicable Gene-set Enrichment (GAGE) showed significantly improved results when compared to the two other commonly used GSA methods of GSEA and PAGE, in terms of consistency across repeated studies/experiments, sensitivity and specificity, when applied on two lung cancer data sets.[30]

Rationale

Though several advances have been made in the treatment of MS, there is still no known cure. Understanding the complex molecular mechanism of MS is crucial to develop effective therapies, and many studies have been conducted to address this problem, but most of them used microarray expression analysis, which does not consider the high variability of protein expression and de novo transcriptome discovery. Also, most of the studies have been conducted on CD4 and CD8 adaptive immunity cells, but not so much to discover the role of innate immunity role, especially monocytes. In this study, RNA-Seq expression data and state of the art analysis tools will be used to conduct the analysis between treatment naïve Multiple Sclerosis patients and then identify the change in comparison with the healthy controls using peripheral blood monocytes RNA-Seq expression level. RNA-Seq is superior in detecting low abundance transcripts,

differentiating biologically critical isoforms, and allowing the identification of genetic variants compared to the microarray platform. Results from this study will provide the groundwork for the new therapy development of MS.

Methods

Raw RNA-Seq expression data from human monocytes for both treatment naïve MS patients and healthy controls (HC) were aligned to a reference human genome sedquence, counted, and then analyzed statistically to look for the highest expressed genes list in MS across cell types. After that, the biochemical pathways involved in MS were investigated by functional pathway enrichment analysis using the differential expression data and knowledge-based databases.

RNA-Seq data

We obtained the raw expression files (FASTQ format) for both treatment naïve MS patients and HC from the ArrayExpress[31] database (Accession code: E-GEOD-77598). These datasets are based on the Illumina HiScanSQ platform. Samples are obtained from 7 biological replicates (3 HC and 4 MS patients). Each biological replicate is divided into three technical replicates taken from the same sample.According to the source study, the total RNA in the samples was isolated from purified monocytes using the Qiagen RNeasy minikit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. Then enriched the samples for mRNA using the Ambion polyA purist kit (ThermoFisher Scientific, Waltham, MA) according to the manufacturer's instructions. Libraries for sequencing were prepared by the Australian Genome Research Facility (AGRF) from 200 ng mRNA, with the mRNA from each ligated with a unique multiplex

tag. Libraries were then pooled and divided across three lanes of the Illumina HiSeq sequencer (Illumina San Diego, CA), sequenced with 100 bp single-end reads.[32]

Table 1. RNA-seq samples information, which includes the condition, RNA-seq sample run ID (SRR#), and Gene expression omnibus[33] sample ID (GSM).

ID	Run	Condition	GSM[1]
1	SRR3146470	N	GSM2054988
2	SRR3146469	N	GSM2054988
3	SRR3146468	N	GSM2054988
4	SRR3146473	N	GSM2054989
5	SRR3146472	N	GSM2054989
6	SRR3146471	N	GSM2054989
7	SRR3146476	N	GSM2054990
8	SRR3146475	N	GSM2054990
9	SRR3146474	N	GSM2054990
10	SRR3146479	MS	GSM2054991
11	SRR3146478	MS	GSM2054991
12	SRR3146477	MS	GSM2054991
13	SRR3146482	MS	GSM2054992
14	SRR3146481	MS	GSM2054992
15	SRR3146480	MS	GSM2054992
16	SRR3146485	MS	GSM2054993
17	SRR3146484	MS	GSM2054993
18	SRR3146483	MS	GSM2054993
19	SRR3146488	MS	GSM2054994
20	SRR3146487	MS	GSM2054994
21	SRR3146486	MS	GSM2054994

22	SRR3146491	MS	GSM2054995
23	SRR3146490	MS	GSM2054995
24	SRR3146489	MS	GSM2054995

[1] Gene expression omnibus

RNA-Seq data analysis

We performed the quality control on raw reads using the FastQC[34] tool (Version 0.11.7) for each sample. Raw reads refinement, and clipping have been performed using Trimmomatic[35] (Version 0.36). We mapped high-quality reads to a reference human genome (GRCh38/hg38) and a human reference transcriptome (Ensembl v70) from the Ensembl[36] genome database using STAR[37] Aligner (Version 2.4.0.1). Unique mapped reads have used to quantify gene expression in each sample. We estimated gene expression as reads counts after filtering, and normalization of raw reads counts using Rsubread[38] and DEseq2[39] R packages as well as differential expression analysis between MS and HC samples. Multiple testing correction has been performed using Benjamini-Hochberg[40]. Genes with adjusted p-values less than alpha of 0.05 were considered differentially expressed. Clustering, principal component analysis of the significant gene list will be conducted using the same package.

Gene ontology and pathway analysis

Functional pathway enrichment analysis was conducted to recognize the differentially expressed genes enriched biochemical pathways that were performed using edgeR,[41,42] GAGE,[30], and goseq[43] R packages. We used Gene sets obtained from the

Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database and Geno Ontology (GO) database.

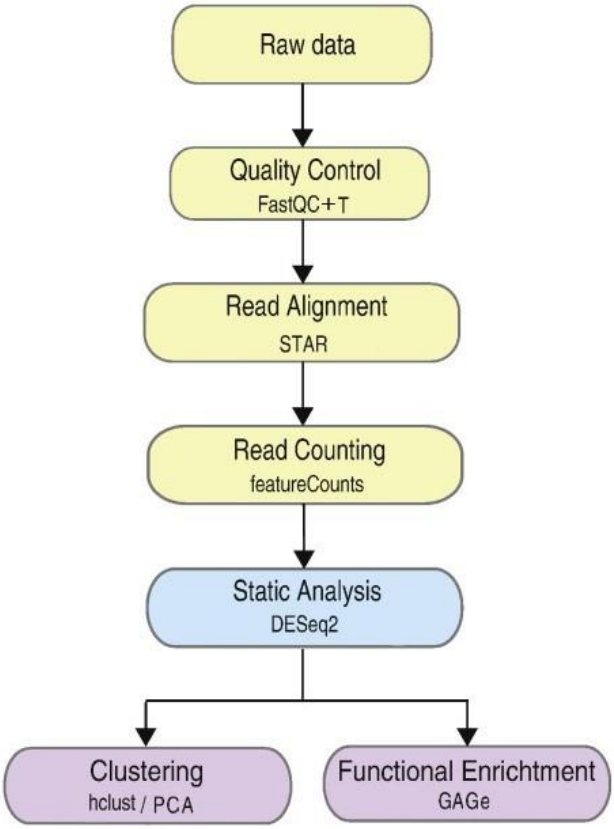


Figure 1. RNA-Seq time course and downstream analysis.[44] The analysis started with assessing the quality of the raw FASTAQ files that have been generated by the sequencing machine using FastQC[34] tool; the data are then processed using Trimmomatic[35] tool. We aligned the processed data to the reference genome (GRCh38/hg38)) and then, the specific features have been counted using Rsubread[38] and DEseq2[39] R packages as differential expression analysis between MS and HC samples. Clustering, principal component analysis of significant genes have been conducted using the same R package. Functional pathway enrichment analysis was performed using the GAGE[30] R package.

Results

Transcriptome analysis of monocytes

Changes occurring in the transcriptome of monocytes in MS patients were compared to HC by RNA-Seq technology. Globally, RNA-Seq produced paired-end reads with enough quality and read coverage per sample to perform reliable gene expression analysis.[45] We analyzed the expression levels of protein-coding and non-coding genes for all samples. Count sum for all samples is showing a strong correlation with sample size factor (Fig. 2.A). The quality analysis confirmed consistency in reads quality between all samples. Multidimensional scaling analysis confirmed high correlation and reproducibility among individual samples of each group (Fig 2.B). Principal component analysis (PCA) and Hierarchical Clustering revealed that the two different groups, MS patients and healthy controls, significantly differed after data normalization using three different normalization methods, log transformation, rlog transformation (RLD), and Variance Stabilizing Transformation (VSD) (Fig. 3 and Fig. 4). As a general observation, there were significant differences between the two groups' transcriptomes; an attempt was made to identify genes with differential expression that were potentially associated with disease etiology. The results of rlog and VST are showing a clearer separation between healthy controls and MS patients in both HC and PCA.

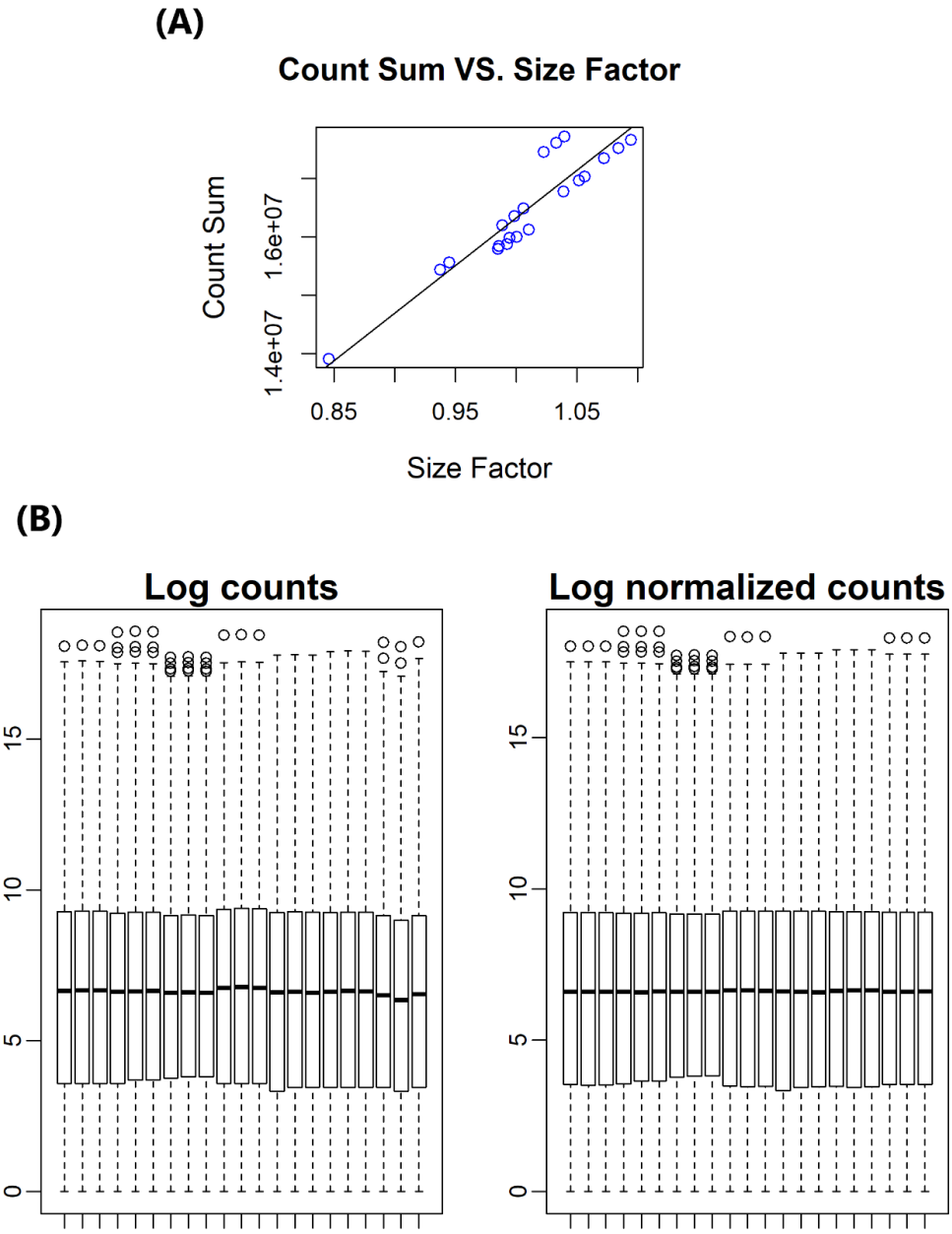


Figure2 (A) Correlation between count sum and sample size factor. For each sample, the size factor of the sample was calculated using the median of the ratios of observed counts and then plotted against the total sum count for the same sample.[46] (B) Box plots of non-normalized raw reads (log2(count+1)) per sample vs. log2 normalized reads count per sample.

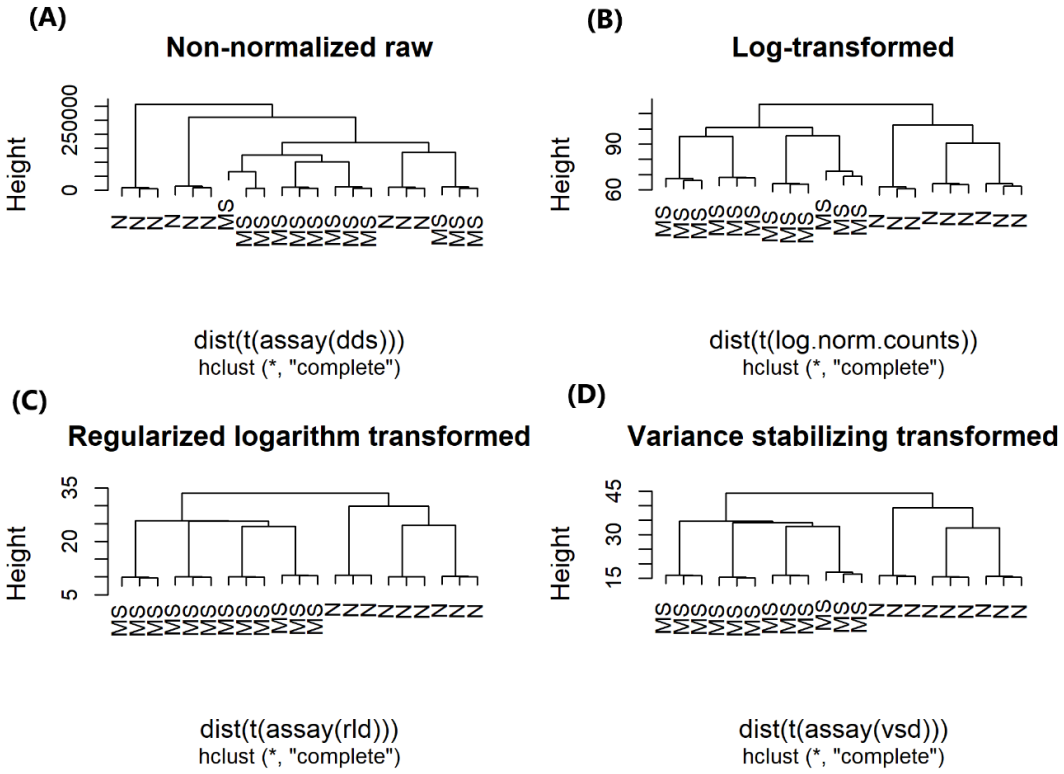


Figure 3. Hierarchal clustering of all samples showing the difference in clustering between the different methods. The methods are the non-normalized raw count(A), the log-transformed count (B), the regularized logarithm transformation (rlog) (C), and variance stabilizing transformations (VST) (D). Both transformations produce transformed data on the normalized log2 scale to library size. The results of the rlog and VST are showing a clear separation between healthy controls and MS patients.

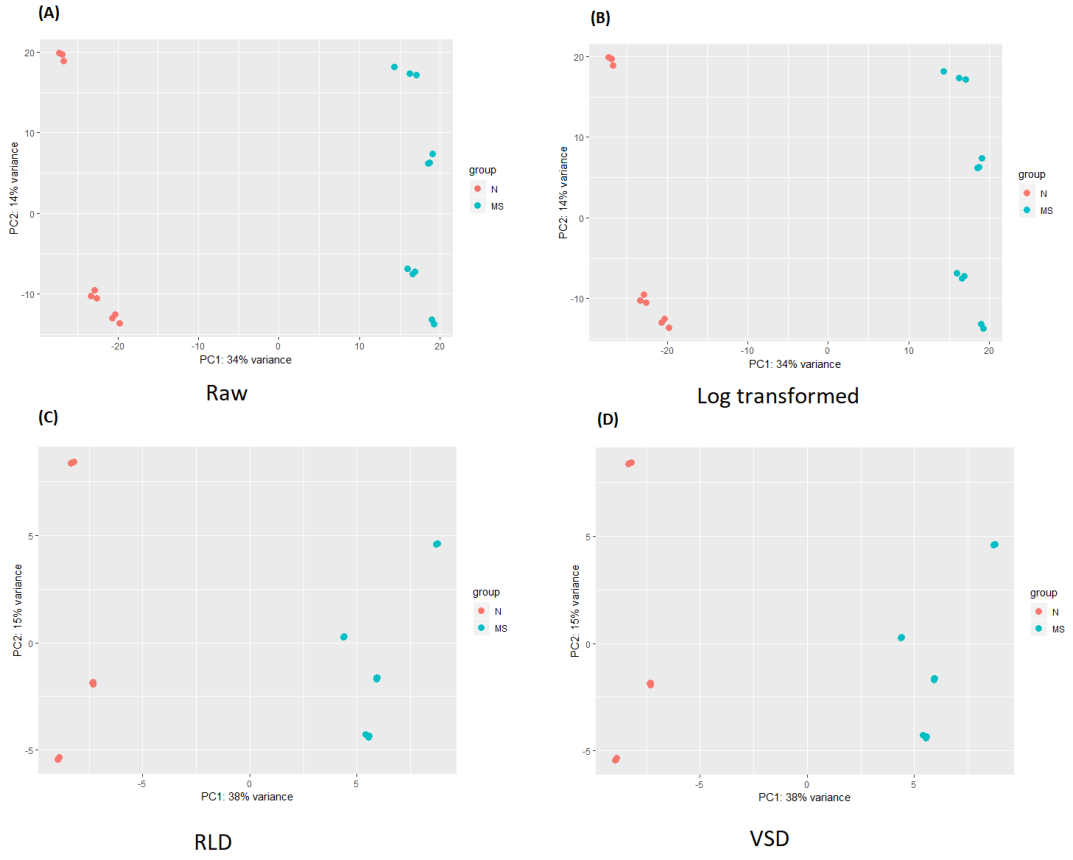


Figure 4. Principal Component Analysis (PCA) plot of first and second principal components of all samples showing the difference in clustering between the different methods. The methods are the non-normalized raw count (A), the log-transformed count (B), the regularized logarithm transformation (rlog) (C), and variance stabilizing transformations (VST) (D). Both transformations produce transformed data on the normalized log2 scale to library size. The results of rlog and VST are showing better segregation of different batches of the same patient.

Significant alteration in the protein-coding transcriptome between the two groups

The initial determination for each sample was the expression level of all loci (Ensembl Homo sapiens.GRCh38.89), then the focus turned to the identification of Differentially Expressed Genes (DEGs) between the two groups. Data analysis revealed 6120 DEGs in the two groups (16% up-regulated and 17% down-regulated in MS group compared to healthy controls). We showed the DEGs expression value as log2 fold

changes (calculated as diseased/healthy samples). Over the mean of normalized counts in the MA plot (Fig. 5. A) and as y-axis 1-posterior probability in log10 scale and on the x-axis log10FC (fold change calculated as diseased/healthy samples) in volcano plots (Fig. 5. B).

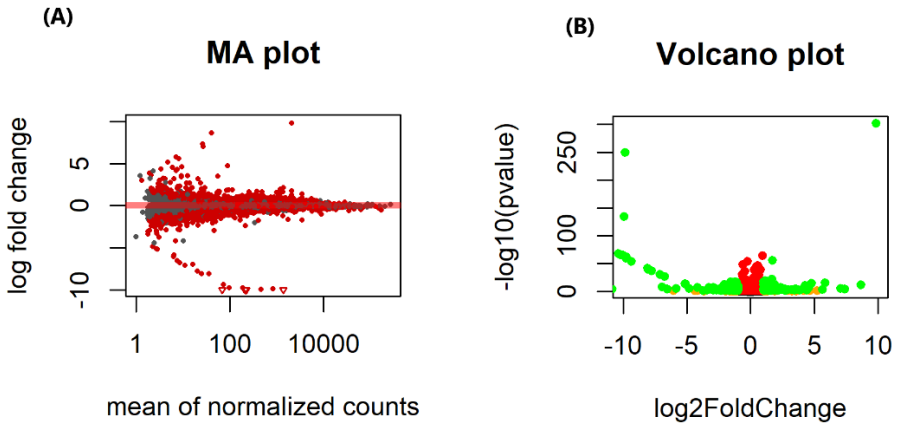


Figure 5. (A) MA plot shows the log2 fold changes over the mean of normalized counts for all the samples, which visualizes the differences between measurements taken in two samples. We plotted the 6120 genes that were Differentially Expressed (DEGs) between the two groups in red. The plots show the log2 fold changes from the treatment over the mean of normalized counts, i.e., the average of counts normalized by size factors. (B) Volcano plot reporting on the y-axis 1-P (posterior probability) in log10 scale and on the x-axis log10FC (fold change calculated as disease/healthy samples). We showed genes identified as significantly differentially expressed (PP > 0.95) as red dots, orange of log2FC>1, green if both.

Tables 2-4 show the top 10 Differentially Expressed Genes (DEGs) based on p. Adjusted value, top 10 down-regulated genes based on expression fold change, and top 10 up-regulated genes based on expression fold change .the plot of the highest significant gene (RPS4Y1) is showing a different clear pattern of expression between MS and healthy controls for all samples (Fig. 6).

Table 2. Top 10 Differentially Expressed Genes (DEGs) based on p. Adjusted value.

Gene	Estimate	p.adjusted	Symbol
ENSG00000129824	-10.0254	0	RPS4Y1
ENSG00000229807	9.822781	1.20E-298	XIST
ENSG00000067048	-9.87465	7.09E-247	DDX3Y
ENSG00000012817	-9.94745	1.07E-131	KDM5D
ENSG00000198692	-10.4063	2.71E-65	EIF1AY
ENSG00000131002	-10.2602	7.45E-64	TXLNGY
ENSG00000183878	-10.0315	7.63E-63	UTY
ENSG00000147050	0.929966	1.01E-61	KDM6A
ENSG00000099725	-9.80276	1.68E-59	PRKY
ENSG00000067646	-9.78947	6.66E-57	ZFY

Table 3. Top 10 down-regulated genes based on expression fold change.

Gene	Estimate	p.adjusted	Symbol
ENSG00000078114	-10.9123	0.000669	NEBL
ENSG00000198692	-10.4063	2.71E-65	EIF1AY
ENSG00000131002	-10.2602	7.45E-64	TXLNGY
ENSG00000183878	-10.0315	7.63E-63	UTY
ENSG00000129824	-10.0254	0	RPS4Y1
ENSG00000012817	-9.94745	1.07E-131	KDM5D
ENSG00000067048	-9.87465	7.09E-247	DDX3Y
ENSG00000099725	-9.80276	1.68E-59	PRKY
ENSG00000067646	-9.78947	6.66E-57	ZFY
ENSG00000114374	-9.40127	9.34E-52	USP9Y

Table 4. Top 10 up-regulated genes based on expression fold change.

Gene	Estimate	p.adjusted	Symbol
ENSG00000229807	9.822781	1.20E-298	XIST

ENSG00000204644	8.655077	1.28E-10	ZFP57
ENSG00000184292	7.38083	0.000631	TACSTD2
ENSG00000213058	7.071241	0.000108	AL365357.1
ENSG00000270641	5.826036	4.06E-14	TSIX
ENSG00000283445	5.644727	0.0043	AL136985.3
ENSG00000004939	5.21146	0.053432	SLC4A1
ENSG00000254521	4.761434	7.51E-13	SIGLEC12
ENSG00000198010	4.63521	0.060525	DLGAP2
ENSG00000237604	4.541081	5.08E-06	AP001056.1

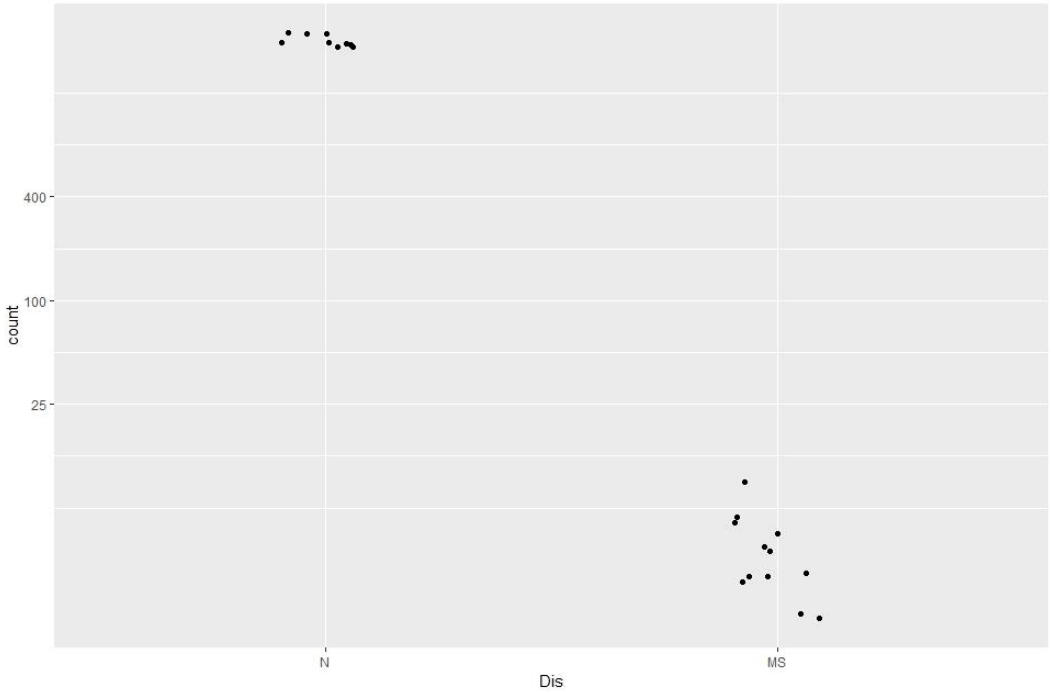


Figure 6. Plot for the normalized count for the top gene based on p.adjusted value (RPS4Y1). Each dot represents a sample HC or MS pool of participants. The plot is clearly showing the upregulated pattern in the healthy controls and downregulated pattern in MS of the gene expression.

Gene ontology and KEGG pathway analysis of DEGs

Functional pathway enrichment analysis has been performed using the GAGE[30] R package using information obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database and Geno Ontology (GO) database. We first analyzed the top KEGG pathways. Pathways "NOD-like receptor signaling pathway," "Chemokine signaling pathway," "Jak-STAT signaling pathway," Toll-like receptor signaling pathway," and "Endocytosis" were up-regulated. Pathways "Ribosome," "Taste transduction," "Steroid biosynthesis," "Oxidative phosphorylation," "Metabolism of xenobiotics by cytochrome P450", and "Calcium signaling pathway" were down-regulated in MS samples compared to healthy controls (Table 5-6). Then, the Gene

Ontology database has been analyzed for three structured networks: biological process (BP), molecular function (MF), and cellular component (CC) (Table 7-8). All the top 10 hits of gene ontology analysis of the biological process are involved in body immunity. The top process based on the percentage of hits is "response to the virus," which could support that a viral infection may cause MS or It could be merely an antiviral innate immune response that happened for common viruses (Fig 8). The top percentage of hits in the Gene ontology analysis of the cellular component domain is in the T-cell receptor complex (Fig 9). This finding supports the role of T-cell involvement in the pathogenesis of MS.

Table 5. KEGG up-regulated pathways.

Pathway	p.geomean	stat.mean	p.val	q.val	set.size
hsa04062 Chemokine signaling pathway	0.001935	2.915741	0.001935	0.299956	138
hsa04621 NOD-like receptor signaling pathway	0.006438	2.550422	0.006438	0.331475	54
hsa04630 Jak-STAT signaling pathway	0.007987	2.441364	0.007987	0.331475	93
hsa04620 Toll-like receptor signaling pathway	0.008554	2.422561	0.008554	0.331475	81
hsa04380 Osteoclast differentiation	0.010767	2.318791	0.010767	0.333775	115
hsa04144 Endocytosis	0.013813	2.215594	0.013813	0.356834	169

Table 6. KEGG down-regulated pathways.

Pathway	p.geomean	stat.mean	p.val	q.val	set.size
hsa03010 Ribosome	0.0312	-1.87988	0.0312	0.998065	85
hsa04742 Taste transduction	0.29141	-0.55355	0.29141	0.998065	22

hsa00100 Steroid biosynthesis	0.297133	-0.54069	0.297133	0.998065	17
hsa00190 Oxidative phosphorylation	0.307036	-0.50543	0.307036	0.998065	119
hsa00980 Metabolism of xenobiotics by cytochrome P450	0.329629	-0.44433	0.329629	0.998065	21
hsa04020 Calcium signaling pathway	0.344394	-0.40108	0.344394	0.998065	101

Table 7. Gene Ontology up-regulated components.

Pathway	p.geomean	stat.mean	p.val	q.val	set.size
GO:0045087 innate immune response	9.95E-12	6.80052	9.95E-12	3.60E-08	432
GO:0051707 response to other organism	1.17E-08	5.638913	1.17E-08	1.88E-05	425
GO:0009607 response to biotic stimulus	1.56E-08	5.584864	1.56E-08	1.88E-05	444
GO:0019221 cytokine-mediated signaling pathway	6.32E-08	5.359285	6.32E-08	5.71E-05	275
GO:0071345 cellular response to cytokine stimulus	1.46E-07	5.184146	1.46E-07	9.74E-05	330
GO:0002252 immune effector process	1.62E-07	5.159235	1.62E-07	9.74E-05	365

Table 8. Gene Ontology down-regulated components.

Pathway	p.geomean	stat.mean	p.val	q.val	set.size
GO:0006613 cotranslational protein targeting to membrane	0.037396	-1.79273	0.037396	1	106
GO:0006415 translational termination	0.037718	-1.79137	0.037718	1	90
GO:0045047 protein targeting to ER	0.037901	-1.78649	0.037901	1	106
GO:0072599 establishment of protein localization to endoplasmic reticulum	0.037901	-1.78649	0.037901	1	106
GO:0006614 SRP-dependent cotranslational protein targeting to membrane	0.038243	-1.78256	0.038243	1	104
GO:0006414 translational elongation	0.039528	-1.76709	0.039528	1	104

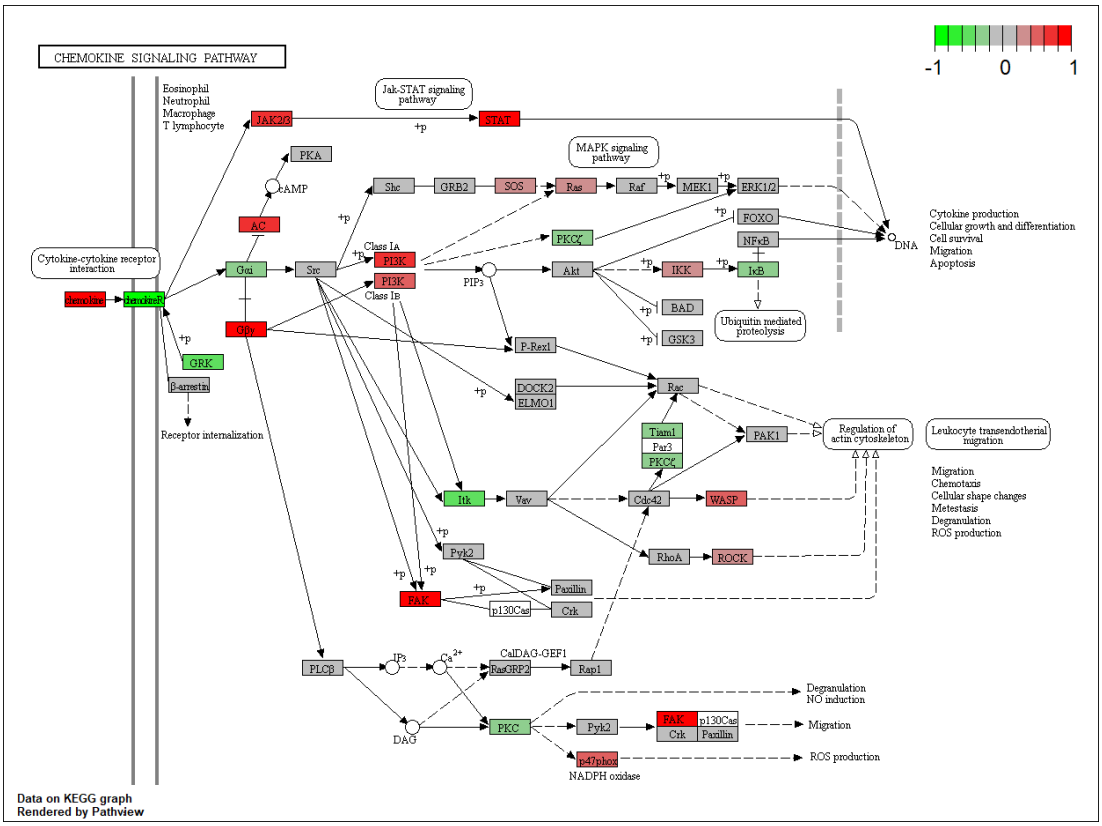


Figure 7. The enriched hsa04062 Chemokine signaling pathway. The red color means up-regulated expression, green means down-regulated expression; the gray means no expression information in the list.

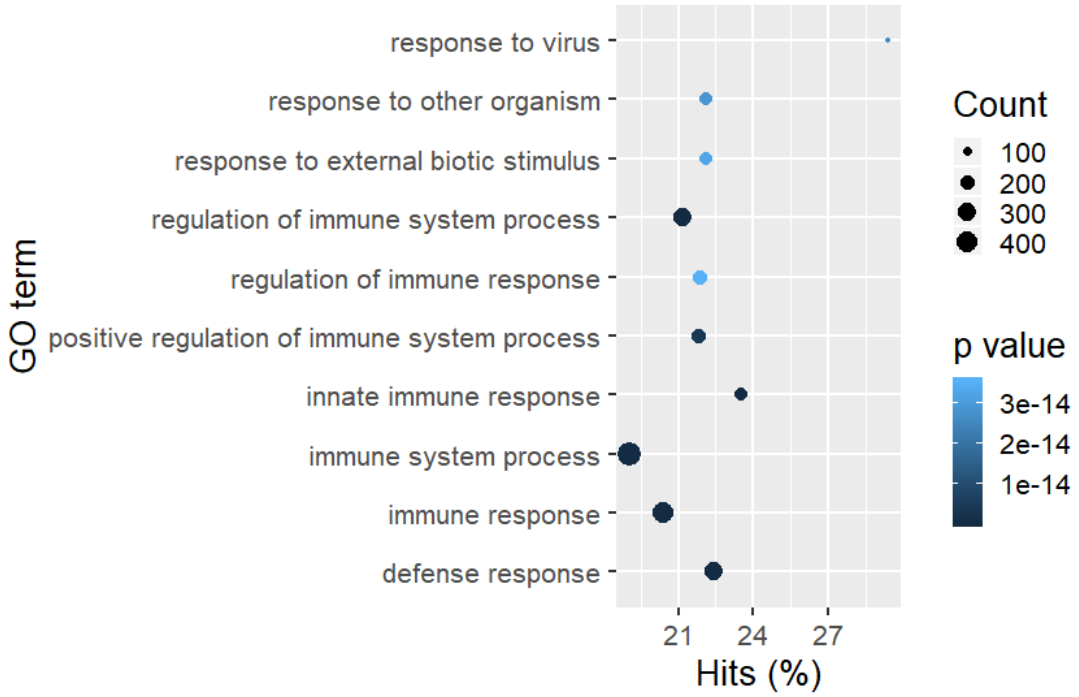


Figure 8. Gene ontology analysis of the biological process, operations, or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms. All the top 10 hits are involved in processes involved in body immunity. The top process based on the percentage of hits is "response to the virus," which could support that a viral infection may cause MS.[47]

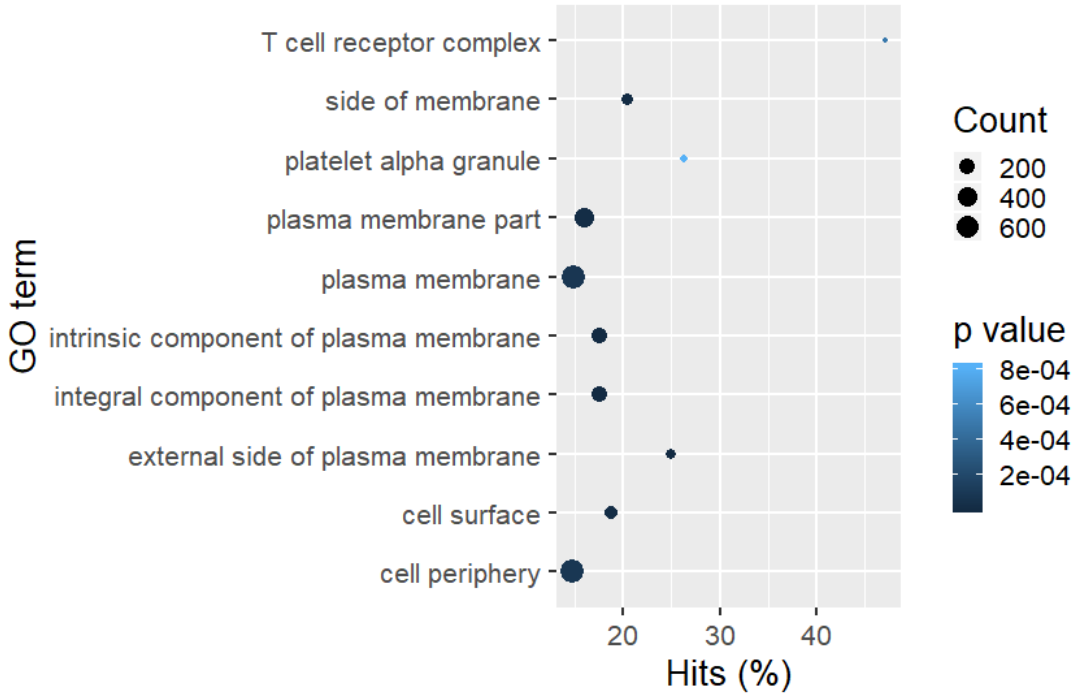


Figure 9. Gene ontology analysis of the cellular component domain, including parts of a cell or its extracellular environment.

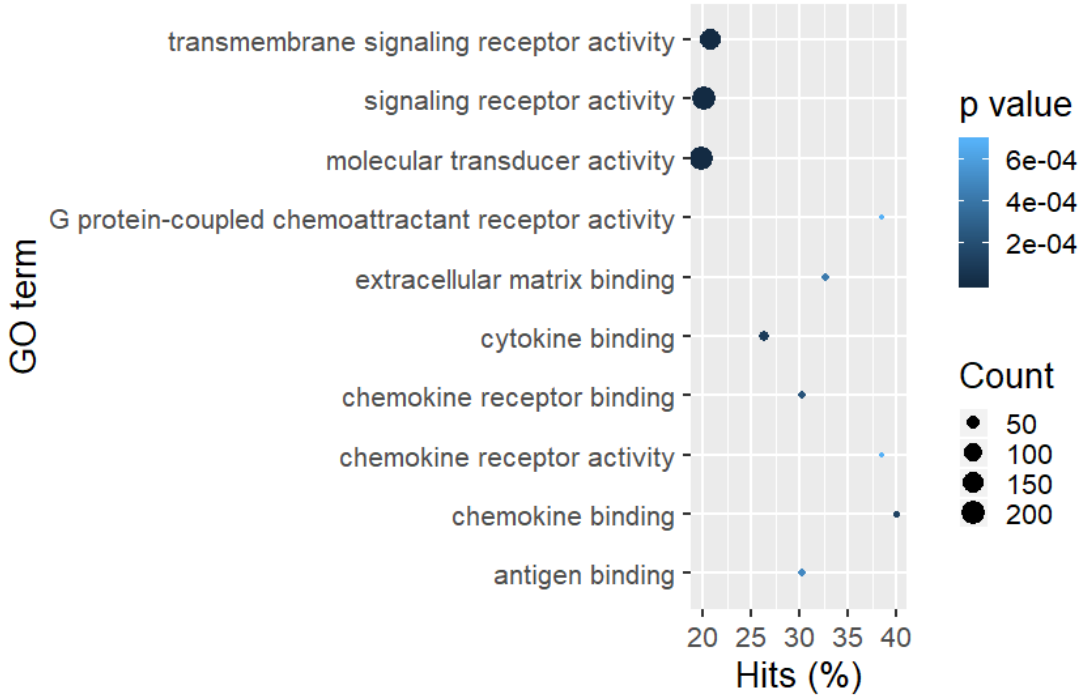


Figure 10. Gene ontology analysis of molecular function domain, the elemental activities of a gene product at the molecular level, such as binding or catalysis.

Discussion

The main findings of the present study were: 1) The identification of specific changes occurring in the transcriptome of MS patients compared to healthy subjects. In particular, the bioinformatic analysis revealed significant alteration of the Chemokine signaling pathway, Jak-STAT signaling pathway, Toll-like receptor signaling pathway, NOD-like receptor signaling pathway. 2) The identification of specific changes occurring in the transcriptome of MS patients compared to healthy subjects not previously linked to MS. In particular, Osteoclast differentiation and Endocytosis. 3) The identification of some genes, which previous studies have not reported them, had an association with MS (RPS4Y1, XIST, DDX3Y, KDM5D, EIF1AY, and TXLNGY). 4) The number of the top

ten DEGs based on p.adjusted value is first to be linked directly to MS (RPS4Y1, XIST, KDM5D, KDM6A, TXLNGY, UTY, PRKY). 5) the number of the top ten DEGs based on p.adjusted value that are involved in gene regulation (RPS4Y1, XIST, DDX3Y, KDM5D, KDM6A, EIF1AY, ZFY).

Monocytes are a conserved subset of white blood cells that originate from myeloid progenitors in the bone marrow, which represents 10% of all white blood cells in humans[48]. Monocytes are rapidly recruited to tissues during infection and inflammation, where they differentiate into macrophages or dendritic cells (DC).[49] They also play a crucial role in the maintenance of homeostasis. While monocytes are essential for removing invading bacteria, viruses, fungi, and protozoans, they can also have adverse effects on the pathogenesis of inflammatory and degenerative diseases.

RPS4Y1 gene codes for 40S ribosomal protein S4, Y isoform 1, which join the larger 60 S subunit to catalyze protein synthesis. RPS4Y, a Y-linked gene in the human, encodes ribosomal protein S4. [50] A homologous locus on the human X chromosome, RPS4X, lies close to the X-inactivation center but fails to undergo X inactivation. [50] The downregulation RPS4Y1 gene expression, but not RPS4X gene expression, could be explained by escaping from activation by RPS4X.[51] On the Y chromosome, RPS4Y maps to a 90-kb segment that has been implicated in Turner syndrome,[50] but it has to be that the low RPS4Y1 gene expression affecting the function or the expression of other effective protein forms is the case since the number of monocytes are within normal ranges for girls affected by Turner Syndrome.[52] XIST, first discovered by searching

cDNA libraries for clones in the 1980s and 1990s, is dysregulated in some cancers and is correlated with tumor progression and poor prognosis. Recently, XIST is reported to be up-regulated in rat spinal cord injury (a neurological disease) model, and XIST knockdown has a noticeable protective effect on the recovery of spinal cord injury by suppressing apoptosis.[53] XIST, Up-regulated in MS patients compared to healthy controls, is an RNA gene on the X chromosome of placental mammals that acts as a major effector of the X inactivation process. Similar results have been found in the SJL mouse strain (used to model the sexual dimorphism observed in MS). [54] Both males and females in the study had up-regulated XIST expression *(* LogFC-female= 3.837, LogFC-male= 3.544*).*[54] ATP-dependent RNA helicase DDX3Y is an enzyme that in humans is encoded by the DDX3Y gene.[55] DDX3Y encodes a class I MHC-restricted H-Y antigen.[56] Rosinski et al. found that an HLAB*2705 HY antigen encoded by DDX3Y was recognized by a CD8 -positive cytotoxic lymphocyte (CTL) clone isolated from a male who had received a hematopoietic cell graft from his HLA-identical sister. [56] This may imply affected antigen presentation in MS patient presentation since the DDX3Y is down-regulated. In a study conducted by Lutterotti et al., DDX3Y found to be down-regulated in MS patient comparing to healthy controls (LogFC= –2.454, p-value= 0.14760), [57] which is consistent with our finding (LogFC= -9.87465, p.adjusted= 7.09E-247). Like XIST, DDX3Y found to be affected in the SJL mouse strain. [54] Both males and females in the study had down-regulated DDX3Y expression (LogFC-female= –11.042, LogFC-male= –9.345). Both KDM5D and KDM6A, which are involved in histone modification via lysine demethylation, are showing activity suggesting

transcriptional activation. [58] KDM6A, activator, is showing a slight upregulation (LogFC=0.929966, p. adjusted=1.01E-61) while KDM5D, repressor, is showing an almost 10 folds downregulation (LogFC=-9.94745, p. adjusted=1.07E-131). Target gene regulation by histone Lysine methylation is a dynamic process that modulates inflammatory responses in the development of a variety of autoimmune diseases, including MS. [59] EIF1AY is an EIF1AX Y-linked homolog, which is an essential translation initiation factor and may function in stabilizing the binding of the initiator Met-tRNA to 40S ribosomal subunits. [55] EIF1AY gene is down-regulated in MS (LogFC=-10.4063, p. adjusted=2.71E-65), which could be one of the factors affecting the RPS4Y1 gene expression level. RPS4Y1 is found to be down-regulated in MS patients comparing to healthy controls in the same study conducted by Lutterotti et al., (LogFC= –1.695, p-value= 0.09977). [57] UTY gene encodes Histone demethylase UTY, which is a protein containing tetratricopeptide repeats, which are thought to be involved in protein-protein interactions. [60] This protein is a minor histocompatibility antigen which may induce graft rejection of male stem cell grafts. [60] UTY involvement may explain the contribution of the major histocompatibility complex (MHC) antigen in MS genetic risk .[7] ZFY gene is a zinc finger-containing protein that may function as a transcription factor. [61] ZFY is down-regulated in microarray gene expression, and B-lymphocytes of siblings with multiple sclerosis (MS) were compared to healthy controls. [19]

Two of the top six KEGG up-regulated pathways (“Jak-STAT signaling pathway,” and “Toll-like receptor signaling pathway”), and one of the bottom six down-regulated pathways (“Calcium signaling pathway”) found in the previous study. [62] (“Jak-

STAT signaling pathway," "Toll-like receptor signaling pathway," "Ribosome," and "Taste transduction") Found to be enriched in another study. [63] The top up-regulated pathway in the KEGG enrichment analysis is supported by many experimental studies, as many members of the CCL and CXCL families of chemokines are found to play a role in the pathogenesis of MS. [64–68] The nucleotide-binding oligomerization domain-like receptors, represented as hsa04621 NOD-like receptor signaling pathway in KEGG, are playing a vital role in the regulation of innate immune response, which explains the up-regulated pattern in monocytes (as part of the innate immune response in the body) and clear role on innate immunity in MS pathogenesis. [69] NOD-like receptors are intracellular sensors of pathogen-associated molecular pattern that enter the cell via phagocytosis and Damage-associated molecular pattern that is associated with cell stress. [70] The up-regulation of the hsa04144 Endocytosis pathway may emphasize the role of innate immunity in the pathogenesis since both hsa04144 Endocytosis, and hsa04621 NOD-like receptor signaling pathway is up-regulated at the same time. The JAK/STAT pathway is one of the most critical signal transduction systems utilized by cells of the innate and adaptive immune systems to initiate and regulate immune responses. [71] Abnormal activation of this pathway promotes dysregulation of innate and adaptive immunity, including activation of pathogenic Th1 and Th17 cells, activation of macrophages, neutrophils, and DCs, and excessive production of proinflammatory cytokines, all of which contribute to the pathogenesis of MS. [72–74] The up-regulated hsa04630 Jak-STAT signaling pathway is consistent with the experimental findings on this matter. There is even a remarkable advance in the development of specific JAK inhibitors that show great

promise in the treatment of autoimmune diseases. [75] Studies to date suggest that Toll-like receptors (TLRs), which activate MyD88-dependent signaling, contribute to the development of MS, whereas MyD88-independent pathways may mitigate disease severity. [76]

Conclusions

The present findings revealed a specific expression pattern of protein-coding in MS. The knowledge of an expression network signature may offer valuable insights into the complex pathogenesis of MS; it may also provide potential targets for therapeutic intervention. All the novel changes in DEGs could be responsible for altered immune response in MS. However, many of these factors play critical roles in protein synthesis, genome methylation, and cell regulation, as well as in other human diseases. We offered an original list of novel candidate genes with potential implications in MS.

For future studies, incorporating more samples, including detailed metadata (age, sex, ethnicity…etc.) to control for the variation of gene expression, is needed. More biological replicates are required to find DEGs with the minuscule difference between the groups.

Acknowledgments

This work was supported by George Mason University - 2018 SSB Bioinformatics Summer Fellowship; George Mason University - 2019 Dissertation Completion Grant; George Mason University - 2019 Provest Office Summer Research Fellowship; and King Fahad Specialist Hospital - Dammam scholarship.

CHAPTER THREE: REGULATORY NETWORK OF BOTH INNATE AND ADAPTIVE IMMUNITY IN MULTIPLE SCLEROSIS

Abstract

Despite Multiple Sclerosis (MS) contributing to morbidity in neurological disorders, the molecular mechanisms of MS continue to remain poorly understood, and biomarkers have yet to be identified. Although there are recognized important cellular and transcriptomic differences between MS samples and healthy controls samples, a systematic overview of the differences between the regulatory processes has not been conducted. This study aimed to investigate Transcription Factors (TF) that play an important role in MS complex molecular pathogenesis by a message-passing network model and specific expression change using co-expression and differential networks for both adaptive and innate immunity cells. Unlike previous studies, we used Ribonucleic Acid Sequencing (RNA-Seq) instead of microarray technology, which does not reflect the high variability of protein expression, to generate transcriptome information. We found that both innate and adaptive immunity share nine out of the top ten upregulated transcription factors, which indicates similar epigenomic control over both systems.

Background

Multiple Sclerosis

Multiple Sclerosis (MS) is an immune-mediated inflammatory disease in which the fatty myelin sheaths around the axons of the Central Nervous System (CNS) are damaged, leading to demyelination and scarring as well as a broad spectrum of signs and

symptoms.[1] MS is the second most common acquired neurological disorder of young adults, with physical trauma being the most common. The disease shows a spectrum of severity, ranging from an asymptomatic pathological process to severe disabling illness. The clinical presentation involves two forms, relapsing disorder in which distinct attacks with clinical stability in between, or progressive condition in which gradual worsening of neurological deficits.

Numerous factors are thought to contribute to the cause of MS, including genetic susceptibility and environmental factors. MS affects mainly young people between the ages of 15 and 50 years, with a peak onset at about age 30. There is a substantial gender preference; most MS patients (70-75%) are women.[2]

The incidence and prevalence of MS vary throughout the world. MS affects nearly two million people worldwide with evident variability in geographic distribution.[3] Recognized low, medium, and high-risk zones have been identified. Near the equator, typically in tropical regions, there is low risk, while MS risk north and south of the equator increases with higher latitudes, in both northern and southern hemispheres.[4]

Although the pathogenesis of MS is poorly understood, evidence suggests that both genetic and environmental components play essential roles in disease development, both independently and interactively.[4] The rule of genetics in MS and its interaction with environmental triggers are currently extensively studied. MS is a disease with evident geographic variability in both prevalence and incidence. The role of environmental factors has historically been thought to be necessary. The geographical distribution and familial aggregation of MS have often been ascribed to the rule of infectious agents, but

there is no consensus regarding this theory,[5] A Canadian study examined a population-based sample of 15,000 individuals with MS using standardized, personally administered questionnaires to identify adoptees or those who had adopted relatives. The frequency of MS among first-degree, non-biological relatives living with the index case was no higher than the expected rate from the Canadian population prevalence data and was significantly less than the rate for biological relatives. These findings support the hypothesis that the familial aggregation of MS is genetically determined rather than environmentally determined.[5,6] A significant contributor to the genetic risk is the major histocompatibility complex (MHC) antigen.[7]

The contribution of genetics to MS is supported by many reports showing familial aggregation of the disease, high concordance rates among twins, and increased risk among relatives of patients with MS. People with MS have a 5–26% chance of having one or more affected relatives, which is a much higher chance than one would expect for a disease with no genetic component. [5,6] Furthermore, the relative risk of MS for identical twins, if one is affected, is approximately 200 to 300 times greater than that of the general population.[6,8] Finally, the first-degree relatives of MS patients have a 2–5% risk of also developing the disease.[8]

Several alleles have been identified as heritable risk factors for MS. Genetic complexity, primarily related to human leukocyte antigens (HLA) of the MHC and, to a lesser extent, non-MHC-related genes, plays a significant role in influencing disease susceptibility, phenotypic expression CD4 T cells, which have central role in MS pathogenesis, [22] experience profound changes in gene expression during the initial hours

after activation. Co-stimulation via the CD28 receptor is required for the effective activation of naive T cells. [23] From pre-clinical studies, it is known that the transcription factor is highly induced in a CD28-dependent manner upon T cell activation and is involved in essential CD4 effector T cell functions, participating in the transcriptional regulation of several T cell activation pathways, including a large group of CD28-regulated genes. [24] Furthermore, Levels of blood monocytes secreting IL-6 and IL-12 were higher in patients with untreated MS and other neurological diseases compared to healthy controls, and MS patients blood monocytes also displayed elevated mean fluorescence intensity for the co-stimulatory molecule CD86, and MS patients with longer disease duration (>10 years) and higher disease severity had higher percentages of CD80 expressing monocytes compared to patients with short duration or lower severity.[25]

Evidence suggests that transcription factors are playing a role in the pathogenesis of MS and other autoimmune diseases.[9] For example, it has previously been observed that members of the NF-kappaB, STAT, AP-1, and E2F families,[10] IRF-1,[11] IRF-2,[12] IRF-5,[13] IRF-8, [11] CREB,[14] PPARgamma and PPARalpha,[15,16] SP1,[13] SP3,[17] RORC,[18] NR4A2, TCF2,[19] ETS-1,[20] and FOXP3[21] may be implicated in MS and its disease subtypes.

RNA-Seq Technology

Since its launch, RNA-Seq has been compared to microarray technology as a means of generating transcriptome information.[26] Both follow a parallel path to answering a biological question. Nevertheless, there are a few key advantages of RNA-Seq technology. First, using microarray technology limits the researcher in detecting

transcripts that correspond to existing genomic sequencing information. RNA-Seq experiments, on the other hand, work well for investigating both known transcripts and explore new ones.[26] Second, RNA-seq delivers a low background signal, which is because DNA sequences can be unambiguously mapped to unique regions of the genome. As a result, noise in the experiment is effortlessly eliminated during analysis. Hybridization issues seen with microarrays, such as cross-hybridization or non-ideal hybridization kinetics, are also removed in RNA-Seq experiments that offer another signal-to-noise advantage.[26] Lastly, RNA-seq can quantify a broad dynamic range of expression levels, with absolute rather than relative values.[26]

Passing attributes between Network for Data Assimilation

Understanding the structure of gene regulation in both healthy and diseased states in different cell types has the potential to help interpret the differential expression results and to reveal critical regulatory differences. Transcriptional regulation involves several distinct mechanisms that must work together to respond to internal or external stimuli.[77] Though the existence of transcription factor binding sites (TFBS) in the promoter or enhancer regions can suggest how that gene is controlled, not all TFBS are functionally relevant or active. Likewise, the binding of a single TF alone may not be enough to recruit RNA polymerase, and several TFs may interact to promote or diminish regulatory potential.[77]

PANDA (Passing Attributes between Networks for Data Assimilation) is an approach that integrates multiple types of genomic data to infer the network of interactions between TFs and their target genes.[78] In contrast to other network

reconstruction approaches, PANDA searches for consistency across multiple sources of information to build a holistic regulatory model. The core of the PANDA algorithm is a message-passing approach in which regulatory processes are modeled as a communication process between "transmitters" (TFs) and "receivers" (target genes). For communication to occur, both transmitters and receivers play an active role: TFs are responsible for regulating genes, and the target genes must be available to be regulated. PANDA starts with a TF/target gene prior regulatory network consisting of potential routes for communication, which is built by mapping TFs motifs to the genome. PANDA integrates this prior network with protein-protein interaction (PPI) and gene expression data, using it to model TF cooperativity and gene co-expression, respectively. Based on this information, it then iteratively estimates the most likely routes of communication through the regulatory network.

Co-expression and differential networks

Co-expression networks are reassembled from gene expression data by pairwise correlation metrics.[79] Changed co-expression patterns of genes between two conditions (healthy vs. diseased) are called differential co-expression, which represent significant potential to identify gene clusters affected by condition change.[80] The creation of differential co-expression networks and their topological analysis, provide us valuable information on the alterations in biological systems in response to environmental and biological perturbations, such as disease formation and gene mutation.[81,82] In many studies, differential co-expression networks were used to identify disease-associated gene

modules in human diseases, including obesity,[83] tumor-associated macrophages,[84] and breast cancer.[85]

Rationale

A significant research effort has been made to understand the mechanisms of MS pathogenesis and to identify diagnostic and prognostic targets. However, disease-specific and effective biomarkers were still not available. Understanding the complex molecular mechanism of MS, including the role of TFs and the gene clusters, is crucial to develop effective therapies. This work aims to uncover TFs that potentially dysregulate many genes and the altered co-expression pattern in MS using both innate and adaptive immune cells. Ultimately, our goal is to piece together relationships and infer a network of TFs and gene modules that are implicated in MS as inferred from the differential expression and co-expression of several hundreds of genes. The results could be a ground for further investigation of involved epigenomics systems. The integration of biological data (gene and transcript information) provides valuable insights on the clarification of the disease mechanisms and identification of molecular signatures of human diseases.

Methods

We focused on pointing out TFs and the gene clusters that play an important role in MS complex molecular pathogenesis from cells that represent both of human's innate immunity (Monocytes cells) and adaptive immunity (CD4, CD8, and B cells).

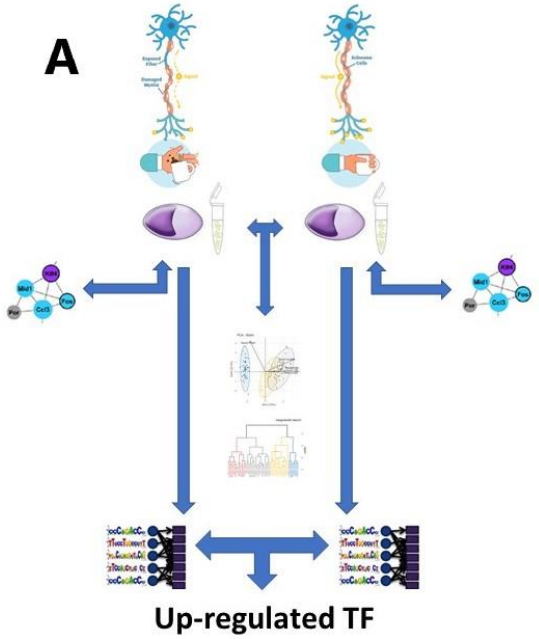

A
Up-regulated TF
B
Up-regulated TF

Figure11 . Analysis workflow for monocytes cells (A) and lymphoblast cells (CD4, CD8, and B cells) (B). in both workflows, RNA-seq data from both healthy controls and treatment naïve MS patients have been analyzed to produce two differential co-expression networks and two TF bipartite networks. For each workflow, bipartite TF networks have been compared to point out up-regulated TFs. We conducted further analysis of gene expression data in each workflow (PCA, Hierarchical clustering).

RNA-Seq data

We obtained the raw expression files (FASTQ format) of Monocytes cells (representing innate immunity) for both treatment naïve MS patients, from the ArrayExpress[31] database (Accession code: E-GEOD-77598). These datasets are based on the Illumina HiScanSQ platform. A total of 24 samples (9 for HC and 15 for MS patients) was used in the analysis (Table 1). According to the source study, the total RNA in the samples was isolated from purified monocytes using the Qiagen RNeasy minikit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. Then enriched the samples for mRNA using the Ambion polyA purist kit (ThermoFisher Scientific, Waltham, MA) according to the manufacturer's instructions. Libraries for sequencing were prepared by the Australian Genome Research Facility (AGRF) from 200ng mRNA, with the mRNA from each ligated with a unique multiplex tag. Libraries were then pooled and divided across three lanes of the Illumina HiSeq sequencer (Illumina San Diego, CA), sequenced with 100bp single-end reads.[32]

We obtained the raw expression files (FASTQ format) of CD4, CD8, and B cells (representing adaptive immunity), for both treatment naïve MS patients and HC, from ArrayExpress[31] database (Accession code: E-GEOD-60424). A total of 16 samples (8 for HC and 8 for MS patients) was used in the analysis (Table 9-10). CD4, CD8, and B cells were obtained from the same source blood sample in both patient and healthy controls. At

the time of blood draw, an aliquot of whole blood was collected into a Tempus tube (Invitrogen), while the remainder of the primary fresh blood sample was processed to highly pure populations of neutrophils, monocytes, B cells, CD4 T cells, and CD8 T cells. RNA was extracted from each of these cell subsets and processed into RNA sequencing libraries (Illumina TruSeq). Sequencing libraries were analyzed on an Illumina HiScan.[86] Further analysis will be conducted, including monocytes and neutrophils from the same patient and HC.

Table 9. RNA-seq samples information for MS patients, which includes the condition, RNA-seq sample run ID (SRR#), and cell type.

ID	Run	Cell type
1	SRR1551097	B
2	SRR1551035	B
3	SRR1550994	B
4	SRR1551098	CD4
5	SRR1551036	CD4
6	SRR1550995	CD4
7	SRR1551099	CD8
8	SRR1551037	CD8

Table 10. RNA-seq samples information for healthy controls, which includes the condition, RNA-seq sample run ID (SRR#), and cell type.

ID	Run	Cell type
9	SRR1551097	B
10	SRR1551035	B
11	SRR1550994	B
12	SRR1551098	CD4
13	SRR1551036	CD4
14	SRR1550995	CD4
15	SRR1551099	CD8

16	SRR1551037	CD8

RNA-Seq data analysis

The quality control on raw reads has been performed using the FastQC[34] tool (Version 0.11.7) for each sample. Raw reads refinement, and clipping have been performed using Trimmomatic[35] (Version 0.36). High-quality reads have been mapped to the reference human genome (GRCh38/hg38) and the human reference transcriptome (Ensembl v70) from the Ensembl[36] genome database using STAR[37] Aligner (Version 2.4.0.1). Unique mapped reads have been used to quantify gene expression in each sample. Gene expression, as reads counts, have been estimated after filtering and normalization of raw reads counts using Rsubread[38] and DEseq2[39] R packages as well as differential expression analysis between MS and HC samples. Multiple testing correction has been performed using Benjamini-Hochberg[40]. Genes with adjusted p-values less than alpha of 0.05 were considered differentially expressed. Clustering, Principal Component Analysis (PCA) of the significant gene list were conducted using the same package.

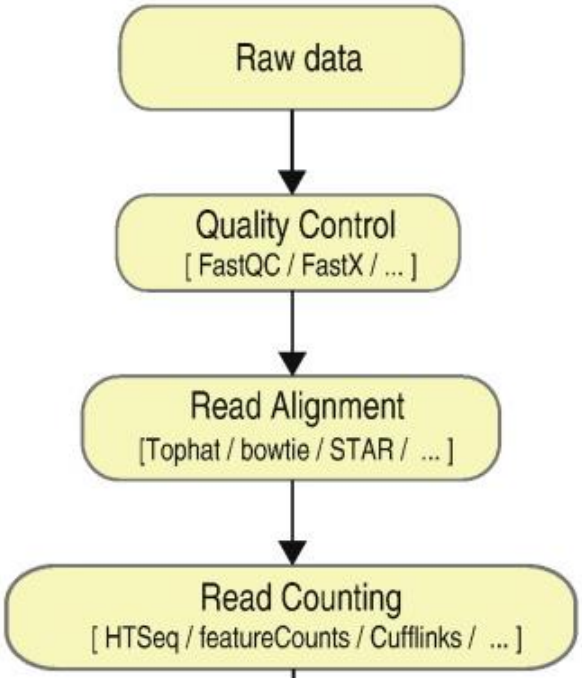


Figure 12. RNA-Seq time course and downstream analysis.[44]

PANDA analysis

We constructed gene regulatory networks using the PANDA[78] R package (pandaR) to point out Transcription Factors (TFs) of interest. PANDA starts with a prior regulatory network inferred by mapping TF binding sites to the genome, integrates Protein-Protein Interaction (PPI), and gene expression data to refine the network structure iteratively and assumes a final consensus regulatory network (Fig. 13). In the regulatory networks estimated by PANDA, each edge connects a TF to a target gene, and the edge weight indicates the strength of the inferred regulatory relationship. The idea here is to start with a prior list of human transcription factors list. Then, the TF has been integrated with the differential expression data that have been generated. We ended up with four

genome-wide, condition-specific regulatory networks for innate (Monocytes) and adaptive (CD4. CD8, and B cells).

We generated one PANDA regulatory network for each group (innate and adaptive immunity) using default parameters (a hamming distance of 0.001 and the update parameter α of 0.1): HC and MS patients. For each network, the same TF/target gene prior regulatory network has been used. To generate the TF/target gene regulatory prior, Transcriptional Regulatory Relationships Unraveled by Sentence-based Text mining (TRRUST v2) transcription factors list will be used.[87] TRRUST is a manually curated database of human and mouse transcriptional regulatory networks. Version 2 of TRRUST contains 8,444 and 6,552 TF-target regulatory relationships of 800 human TFs. They have been derived from 11,237 PubMed articles, which describe small-scale experimental studies of transcriptional regulations.

For each group, the TF/target gene prior and the sample group gene expression data have been used. The TF/target gene edge weights emerging from PANDA were then used to compare networks between the two conditions. For pairs of networks, the TF out-degree, defined as the sum of edge weights from that TF, and the gene in-degree, defined as the sum of all incoming edge weights, a gene received from all expressed TFs in the network have been compared.

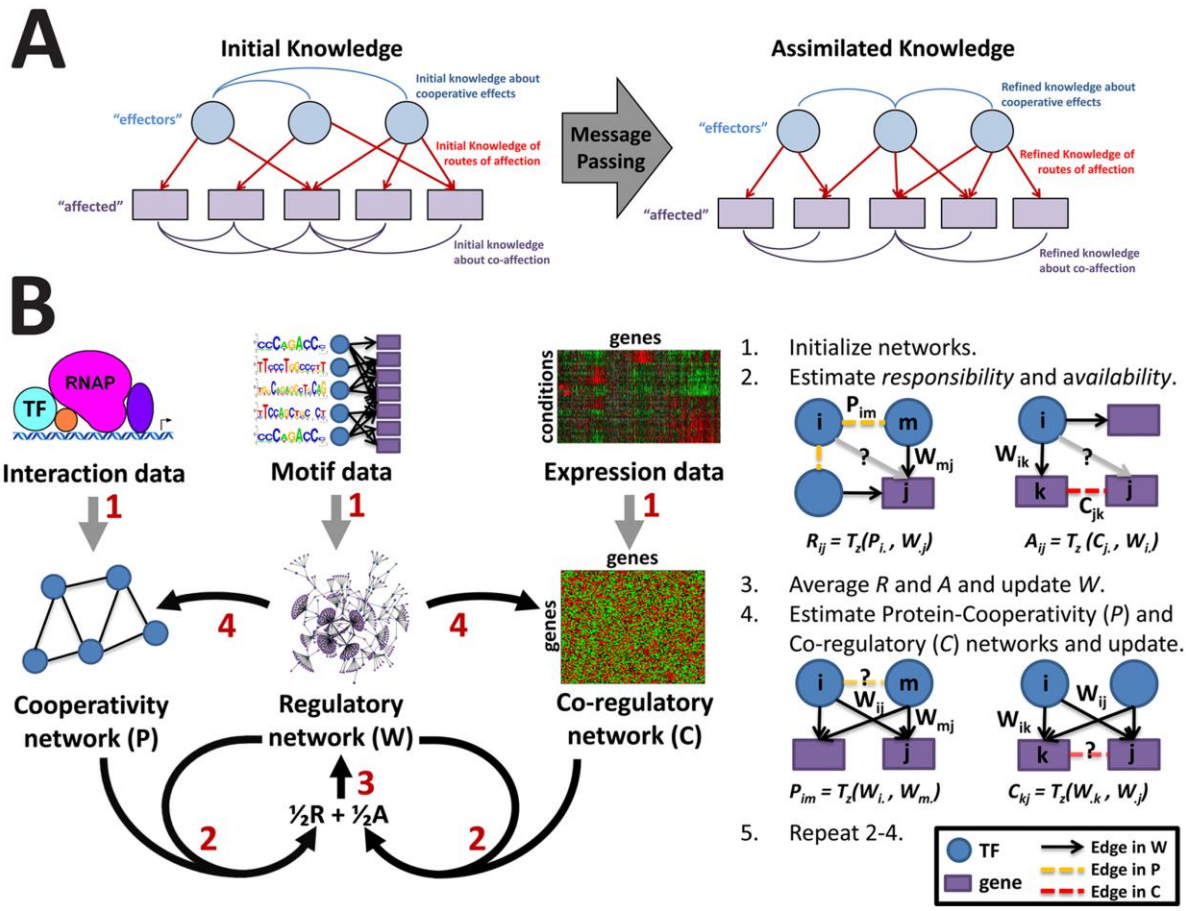


Figure 13. Outline of the PANDA approach for regulatory network inference integrating three data types.[77] (A) A conceptual illustration showing the generalized framework for the message-passing procedure. (B) An illustration of how the message-passing procedure is applied in assimilating data that represents several various components of biological regulation. The networks are initialized from sequence motif data, physical protein interactions, and co-expression, respectively. The method iteratively passes messages within and among networks to emphasize agreement regarding the TF-gene regulatory relationships occurring within a system. At each time step regulatory (*W*), co-regulatory (*C*), and protein-cooperativity (*P*) networks are updated by passing information between the regulatory network, that reflects potential paths for regulation in the biological system, and the data-specific networks, that reflect "static" pair-wise information shared between gene products and TFs. At convergence, the method provides harmonized expression and interaction modules specific to a biological condition of interest, as well as the output regulatory network controlling those modules in each condition.

Construction of co-expression and differential networks in diseased and healthy states

We constructed four co-expression networks based on gene expression profiles of both innate and adaptive immunity (as represented monocytes cells and CD4, CD8, and B

cells respectively) in HC and MS patients. The expression data were log-transformed before conducting the analysis using the limma[88] R package. We computed the Pearson correlation coefficient (PCC) of the expression profiles between every pair of genes in the same group.

We created two differential networks for both innate and adaptive immunity (as represented monocytes cells and CD4, CD8, and B cells). In each network, we compared the HC vs. MS networks (healthy network used as reference). The idea is to cancel out common shared nodes and edges between HC and MS networks in both innate and adaptive immunity. Differential network analysis has been conducted using the Cytoscape[89] Diffany[90] plugin.

Local and global topological features of networks and their modules were represented by several metrics, including degree, betweenness connectivity, network density, and clustering coefficient, and were determined via NetworkAnalyzer[91] and Cytohubba[92] plugins of Cytoscape (Version 3.7.1).

The differential networks were analyzed using the Cytoscape MCODE[93] plugin to find network modules. Ranking of modules was based on MCODE scores (i.e., average connectivity).

Cytoscape CytoMCS[94] plugin has been used to compute the maximum common edge subgraph between the two differential networks (innate and adaptive immunity) to point out conserved nodes. The networks have been treated as undirected networks with 20% perturbation and no edge exception.

The list of the conserved nodes has been used to conduct pathway enrichment analyses of gene sets through the Cytoscape ClueGo[95] plugin using KEGG[96] and Reactome[97] as the data sources. P-values were determined through a 2-sided hypergeometric test and adjusted via Bonferroni's method. A threshold of adjusted p-value < 0.05 was used to determine the statistical significance of enrichment results.

Results

Transcriptome analysis of Monocytes cells

RNA-Seq produced paired-end reads with enough quality and read coverage per sample to perform reliable gene expression analysis. The expression levels of protein-coding and non-coding genes have been analyzed for all samples. Count sum for all samples is showing a strong correlation with sample size factor (Fig. 2A). The quality analysis confirmed consistency in reads quality between all samples. Multidimensional scaling analysis confirmed high correlation and reproducibility among individual samples of each group (Fig. 2B).

PCA and Hierarchical Clustering revealed that the two different groups, MS patients and healthy controls, significantly differed after data normalization using three different normalization methods, log transformation, rlog transformation (RLD), and Variance Stabilizing Transformation (VSD) (Fig. 3 and Fig. 4).

Transcriptome analysis of CD4, CD8, B, Monocytes, and Neutrophils cells

The expression levels of protein-coding and non-coding genes have been analyzed for all samples. Log-transformed counts density plots are showing a clear different distribution pattern for Neutrophils and Monocytes cells compared to CD4, CD8, B cells (Fig. 14A). The distribution of CD4, CD8, and B cells are almost identical after removing neutrophils and monocytes cells samples (Fig. 14B, 14C).

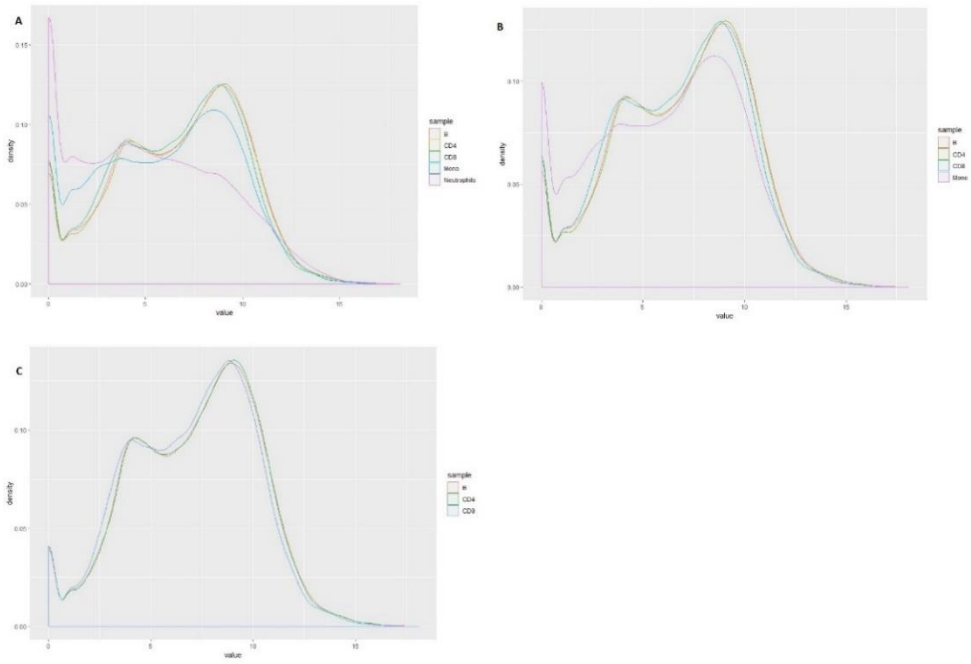


Figure14 . Log-transformed counts density plots for CD4, CD8, B, monocytes, and neutrophils cells (A), CD4, CD8, B, and monocytes cells (B), CD4, CD8, and B cells only (C). The plots are showing the almost equal distribution for CD4, CD8, and B cells.

PCA and Hierarchical Clustering revealed that the cell lines, in both MS patients and HC, significantly differed after data normalization using two different normalization methods, rlog transformation (RLD), and Variance Stabilizing Transformation (VSD). We found that CD4, CD8, and B cells are grouped in one cluster, and monocytes and neutrophils are grouped in another cluster (Fig. 15 and Fig. 17). PCA analysis revealed a very close expression pattern of CD4 and CD8 cells. CD4, CD8, and B cells tend to group tightly after RLD and VSD transformation (Fig. 16 and Fig. 18). Monocytes and Neutrophils cells tend to group, which is consistent with the clustering results. PCA and Hierarchal clustering of all MS patients and HC samples are showing that cell lines from the two different conditions and the same cell line tend to cluster together, which indicate a slight change in the expression pattern between MS patients and HC samples (Fig. 19 and Fig. 20).

Differential expression between MS patients and HC samples revealed 18, 0, 40, 10, 18 DEGs for CD4, CD8, B, Neutrophils, and Monocytes cells, respectively. SLC2A14 gene is the only gene shared by all cell lines (Fig. 21).

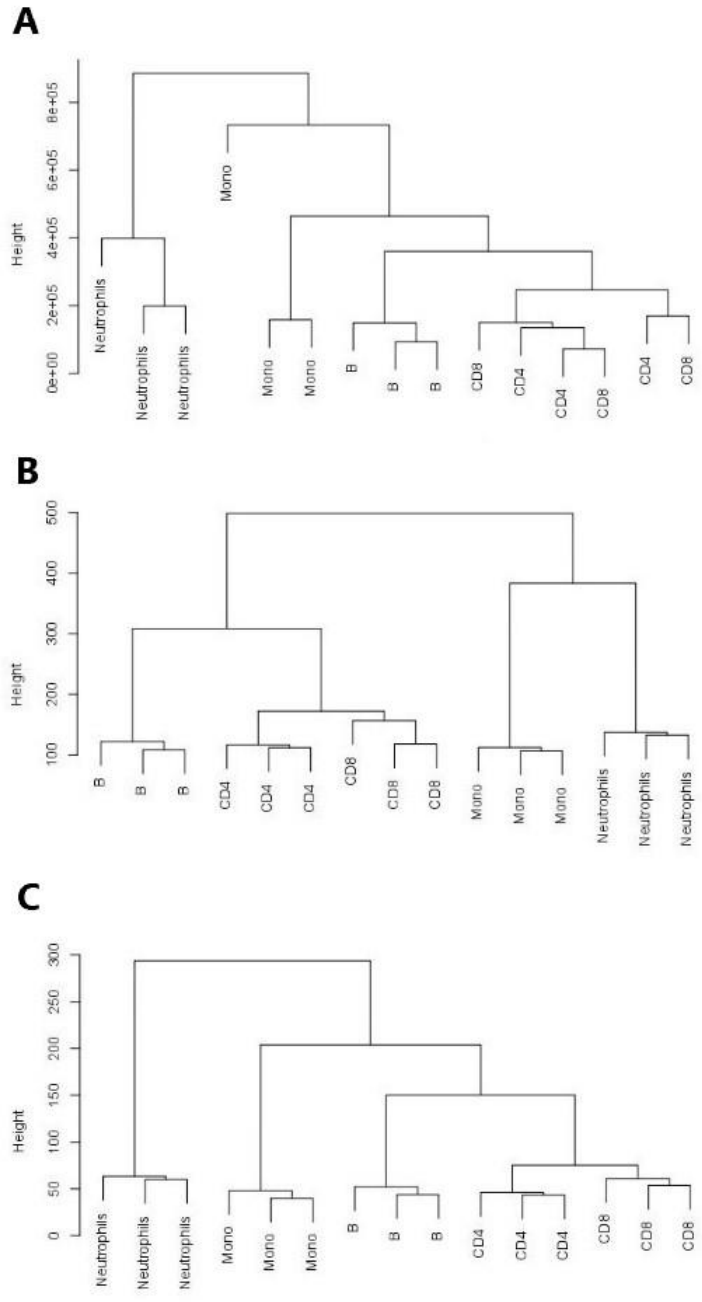


Figure 15. Hierarchal clustering of all MS patients' samples showing the difference in clustering between the different methods. The methods are the non-normalized raw count(A), the regularized logarithm transformation (rlog) (B), and variance stabilizing transformations (VST) (C). Both transformations produce transformed data on the log2 scale normalized to library size. The results of rlog and VST, showing the clustering of the same cell line.

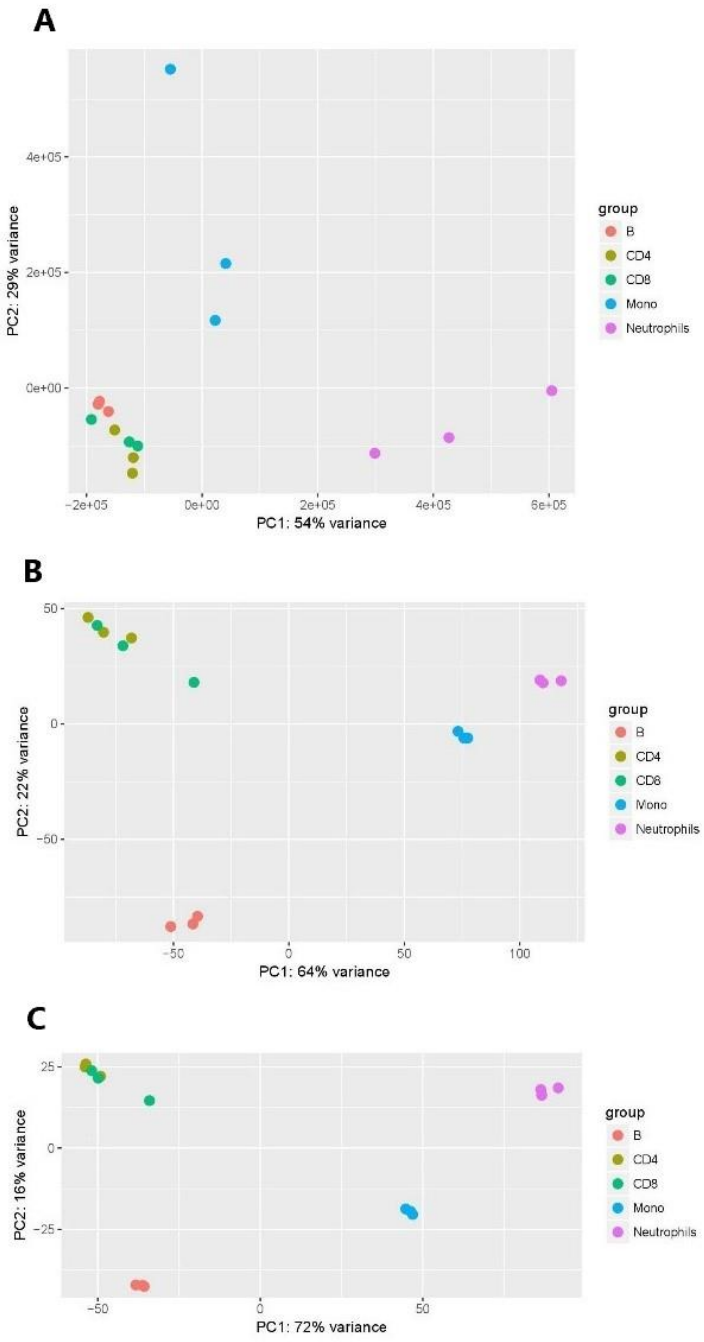


Figure 16. Principal Component Analysis (PCA) of all MS patients' samples showing the difference in clustering between the different methods. The methods are the non-normalized raw count (A), the regularized logarithm transformation (rlog) (B), and variance stabilizing transformations (VST) (C). Both transformations produce transformed data on the log2 scale normalized o library size. The results of the rlog and VST are showing better segregation of different cell lines. Also, it is showing a similar expression pattern of CD4, CD8 cells.

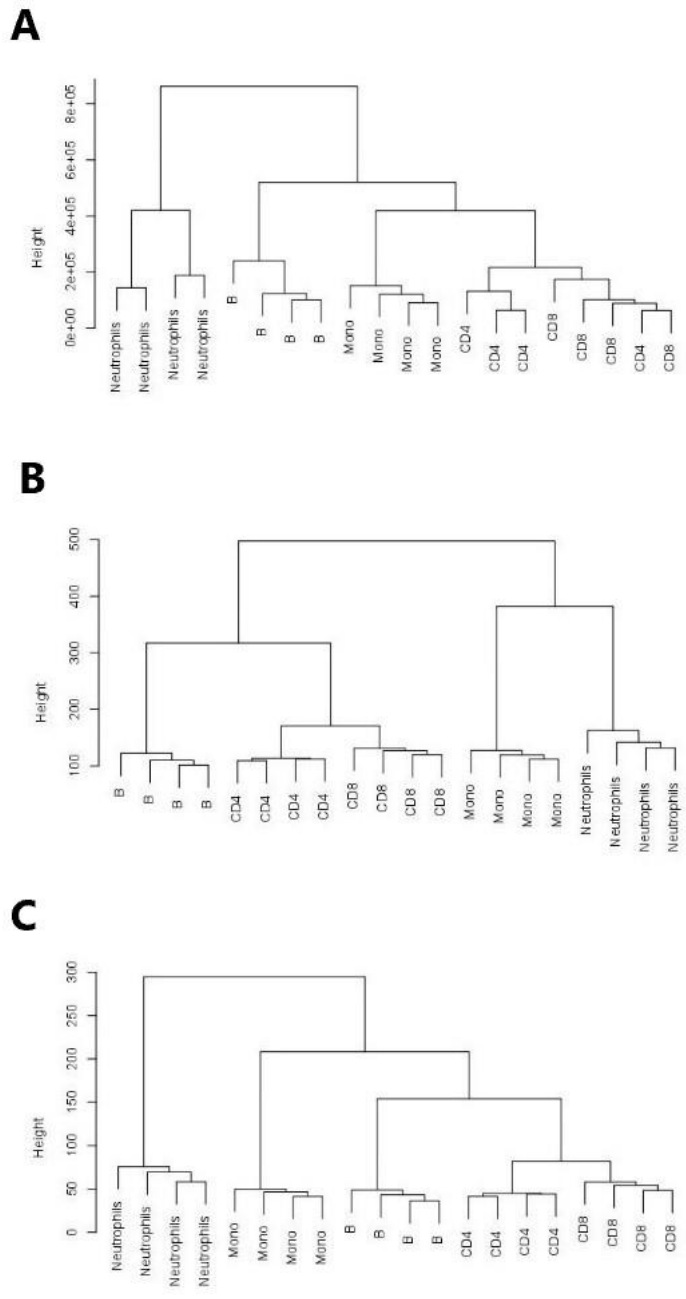


Figure 17. Hierarchal clustering of all healthy controls' samples showing the difference in clustering between the different methods. The methods are the non-normalized raw count(A), the regularized logarithm transformation (rlog) (B), and variance stabilizing transformations (VST) (C). Both transformations produce transformed data on the log2 scale normalized to library size. The results of rlog and VST showing the clustering of monocytes and neutrophils cells in one group (representing innate immunity), and CD4, CD8, and B cells in another group (representing adaptive immunity).

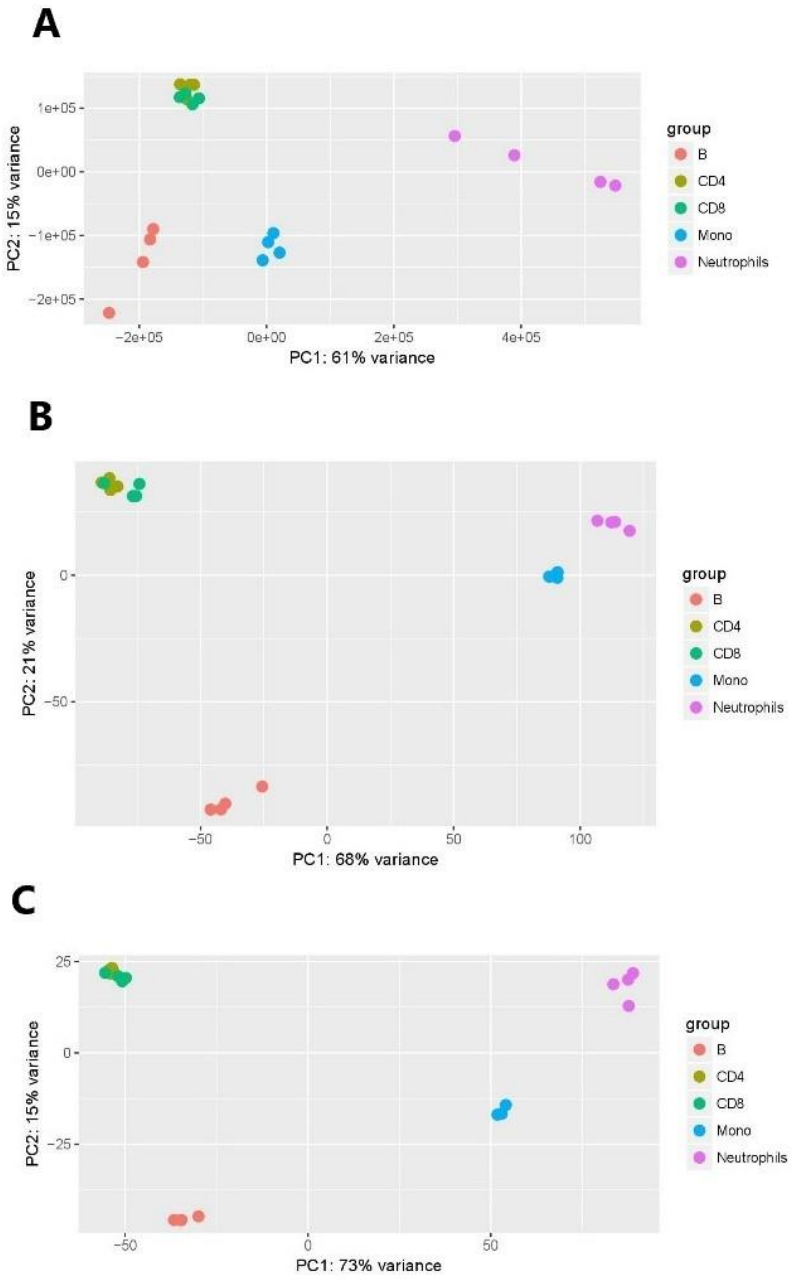


Figure 18. Principal Component Analysis (PCA) of all healthy control samples showing the difference in clustering between the different methods. The methods are the non-normalized raw count (A), the regularized logarithm transformation (rlog) (B), and variance stabilizing transformations (VST) (C). Both transformations produce transformed data on the log2 scale normalized to library size. The results of rlog and VST are showing better segregation of different cell lines and control for the variability between different samples. Also, it is showing a similar expression pattern of CD4, CD8 cells.

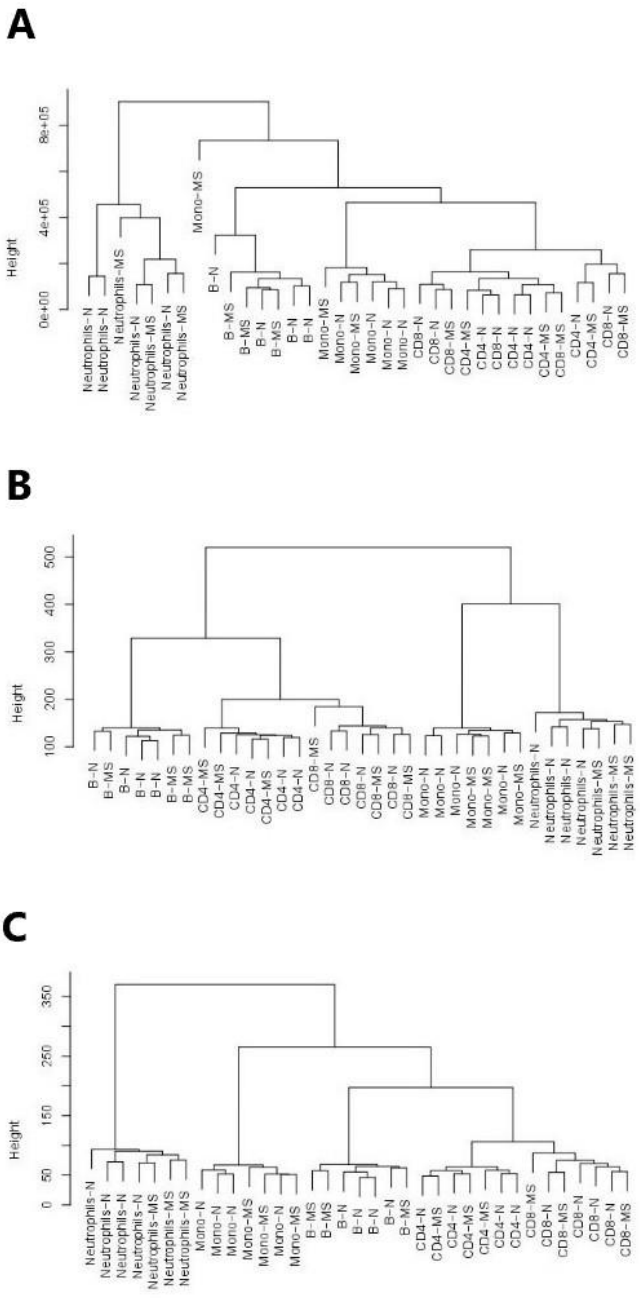


Figure 19. The hierarchal clustering of all MS patients and healthy control samples are showing the difference in clustering between the different methods. The methods are the non-normalized raw count(A), the regularized logarithm transformation (rlog) (B), and variance stabilizing transformations (VST) (C). Both transformations produce transformed data on the log2 scale normalized to library size. The results of rlog and VST showing clustering of the same cell line for both MS patients and healthy controls samples. Cell lines from the two different conditions and the same cell line tend to cluster together.

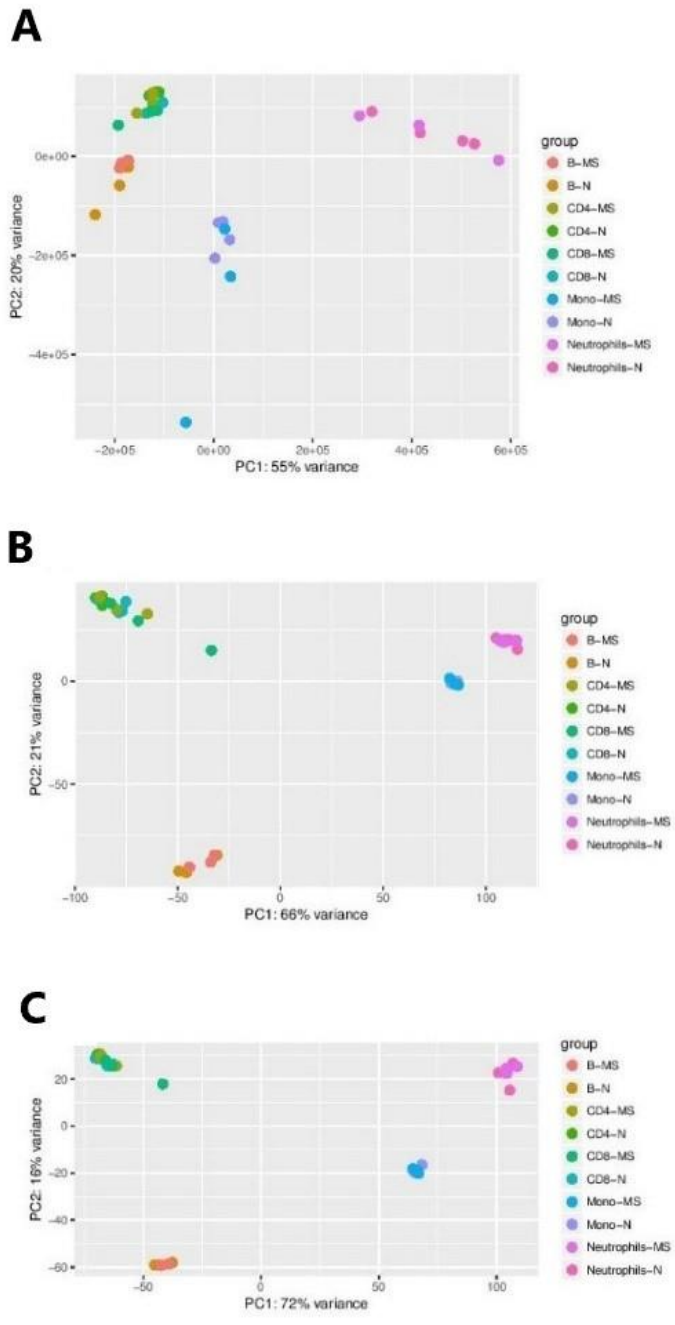


Figure 20. Principal Component Analysis (PCA) of all MS patients and healthy control samples are showing the difference in clustering between the different methods. The methods are the non-normalized raw count (A), the regularized logarithm transformation (rlog) (B), and variance stabilizing transformations (VST) (C). Both transformations produce transformed data on the log2 scale normalized to library size. The results of rlog and VST are showing better segregation of different cell lines and control for the variability between different samples. Also, it is

showing a similar expression pattern of the Cell lines from the two different conditions, and the same cell line tends to cluster together.

Figure 21. Venn diagram of the intersection of the DEGs (MS vs. healthy controls) for each cell line. All cell lines share the SLC2A14 gene in their DEGs list.

Cell line and condition-specific gene regulatory networks

We used PANDA to estimate gene regulatory networks in innate immunity (represented by Monocytes cells) and adaptive immunity (represented by CD4, CD8, and B cells) in each condition (MS patients and HC). For each network, we started with the same TF/target gene prior regulatory network but used the cell-specific gene expression data, which resulted in four gene regulatory networks where each edge connects a TF to a target gene, and the associated edge weight indicates the strength of the inferred regulatory relationship in that cell. These networks can inform us about the genome-wide

regulation of the cell lines and condition analyzed as we compare 795TFs, 8,427 regulatory links, and more than 20,000 target genes.

In both innate and adaptive immunity, we computed the difference between the "out-degree" (sum of edge weights from that TF) in the MS patients' samples and HC samples. Innate and adaptive immunity networks shared nine out of the top ten regulated TF. We found the "SP1" is the top regulated TF in both networks. Innate and adaptive immunity networks shared CDKN1A, MYC, and NFKB1 genes in the top ten most targeted genes. We found that the CDKN1A gene as the most targeted gene in both networks.

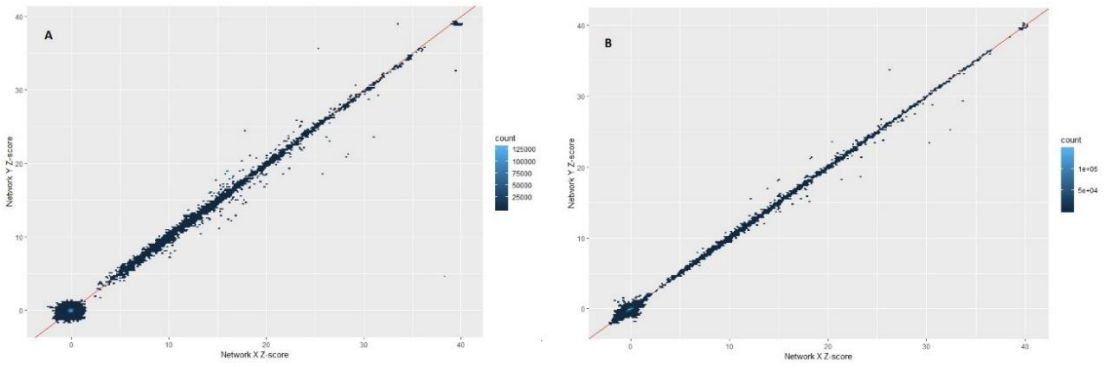
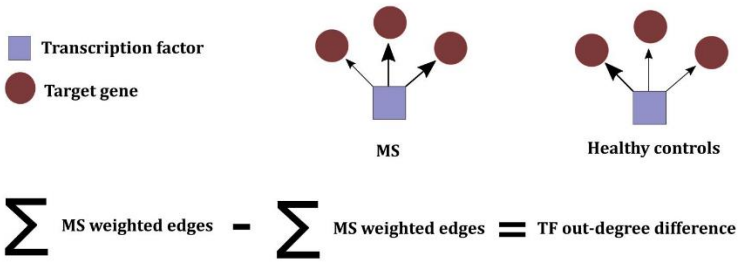


Figure 22. Z-score comparison between MS patients' samples and healthy controls samples in using monocytes cells network (A) and CD4, Cd8, and B cells network (B).

A



B

TF	Official name	Out-degree
SP1	Sp1 transcription factor	1955.49
NFKB1	nuclear factor kappa B subunit 1	1179.83
RELA	RELA proto-oncogene, NF-kB subunit	1114.88
TP53	Tumor protein p53	764.90
E2F1	E2F transcription factor 1	642.59
STAT3	signal transducer and activator of transcription 3	606.50
MYC	MYC proto-oncogene, bHLH transcription factor	577.16
JUN	Jun proto-oncogene, AP-1 transcription factor	430.42
YY1	YY1 transcription factor	414.42
SPI1	spleen focus forming virus (SFFV) proviral integration oncogene	366.34

More targeting in monocytes

SRCAP	Snf2 related CREBBP activator protein	-370.7
IRF8	Interferon regulatory factor 8	-405.936
WWP1	WW domain containing E3 ubiquitin protein ligase 1	-406.124
SREBF2	Sterol regulatory element binding transcription factor 2	-537.021
MTA1	Metastasis associated 1	-561.138
DNMT3L	DNA methyltransferase 3 like	-659.378
TFDP1	Transcription factor Dp-1	-743.191
RB1	RB transcriptional corepressor 1	-1008.43
NFE2L1	Nuclear factor, erythroid 2 like 1	-1186.21
SOX10	SRY-box 10	-2235.27

Less targeting in monocytes

C

TF	Official name	Out-degree
SP1	Sp1 transcription factor	2292.12
NFKB1	nuclear factor kappa B subunit 1	1118.04
RELA	RELA proto-oncogene, NF-kB subunit	1062.90
TP53	Tumor protein p53	817.97
E2F1	E2F transcription factor 1	751.56
STAT3	signal transducer and activator of transcription 3	585.48
MYC	MYC proto-oncogene, bHLH transcription factor	575.99
YY1	YY1 transcription factor	439.75
JUN	Jun proto-oncogene, AP-1 transcription factor	436.58
HIF1A	hypoxia inducible factor 1 subunit alpha	395.18

More targeting in CD4, CD8, B

HIC1	HIC ZBTB transcriptional repressor 1	-393.455
XRCC5	X-ray repair cross complementing 5	-432.734
ISL1	ISL LIM homeobox 1	-469.87
MYB	MYB proto-oncogene, transcription factor	-496.599
SRY	sex determining region Y	-576.246
DRAP1	DR1 associated protein 1	-817.04
TLE3	TLE family member 3, transcriptional corepressor	-881.151
RBL2	RB transcriptional corepressor like 2	-1090.67
NFIB	nuclear factor I B	-1123.83
SOX2	SRY-box 2	-2215.93

More targeting in CD4, CD8, B

Figure 23. Transcription factors differentially targeting genes in MS patients and healthy controls samples. (A) Illustration of the TF out-degree difference between MS samples and healthy controls. Positive values indicate higher targeting in cell lines, and negative values indicate higher targeting in tissues. (B) TFs with the most considerable difference in out-degree comparing MS-vs-healthy controls in monocytes cells. (C) TFs with the most considerable difference in out-degree comparing MS-vs-healthy controls in CD4, CD8, and B cells.

Table 11. Top transcription factors out-degree (Monocytes)

TF	Official name	Out-degree
SP1	Sp1 transcription factor	1955.49
NFKB1	nuclear factor kappa B subunit 1	1179.83
RELA	RELA proto-oncogene, NF-kB subunit	1114.88
TP53	Tumor protein p53	764.90
E2F1	E2F transcription factor 1	642.59
STAT3	signal transducer and activator of transcription 3	606.50
MYC	MYC proto-oncogene, bHLH transcription factor	577.16
JUN	Jun proto-oncogene, AP-1 transcription factor	430.42
YY1	YY1 transcription factor	414.42
SPI1	spleen focus forming virus (SFFV) proviral integration oncogene	366.34

Table 12. Bottom transcription factors out-degree (Monocytes)

TF	Official name	Out-degree
SOX10	SRY-box 10	-2235.26
NFE2L1	nuclear factor, erythroid 2 like 1	-1186.21
RB1	RB transcriptional corepressor 1	-1008.43
TFDP1	transcription factor Dp-1	-743.19
DNMT3L	DNA methyltransferase 3 like	-659.38
MTA1	metastasis associated 1	-561.14
SREBF2	sterol regulatory element binding transcription factor 2	-537.02
WWP1	WW domain-containing E3 ubiquitin protein ligase 1	-406.12
IRF8	interferon regulatory factor 8	-405.93
SRCAP	Snf2 related CREBBP activator protein	-370.70

Table 13. Top transcription factors out-degree (CD4, CD8, and B cells).

TF	Official name	Out-degree
SP1	Sp1 transcription factor	2292.12
NFKB1	nuclear factor kappa B subunit 1	1118.04
RELA	RELA proto-oncogene, NF-kB subunit	1062.90
TP53	Tumor protein p53	817.97
E2F1	E2F transcription factor 1	751.56
STAT3	signal transducer and activator of transcription 3	585.48
MYC	MYC proto-oncogene, bHLH transcription factor	575.99
YY1	YY1 transcription factor	439.75
JUN	Jun proto-oncogene, AP-1 transcription factor	436.58
HIF1A	hypoxia inducible factor 1 subunit alpha	395.18

Table 14. Bottom transcription factors out-degree (CD4, CD8, and B cells).

TF	Official name	Out-degree
SOX2	SRY-box 2	-2215.93
NFIB	nuclear factor I B	-1123.83
RBL2	RB transcriptional corepressor like 2	-1090.67
TLE3	TLE family member 3, transcriptional corepressor	-881.151
DRAP1	DR1 associated protein 1	-817.04
SRY	sex-determining region Y	-576.246
MYB	MYB proto-oncogene, transcription factor	-496.599
ISL1	ISL LIM homeobox 1	-469.87
XRCC5	X-ray repair cross complementing 5	-432.734
HIC1	HIC ZBTB transcriptional repressor 1	-393.455

Table 15. Top gene in-degree (Monocytes)

Gene	In-degree
CDKN1A	1246.00580
MYC	714.38921
VEGFA	466.69242
TP53	312.35393
NFKB1	296.50460
PTGS2	270.49655
FOS	239.13365
JUN	219.95652
CXCL8	206.26191
BAX	188.60459

Table 16. Bottom gene in-degree (Monocytes).

Gene	In-degree
CDK11B	-1217.215663
MSL1	-742.510158
UBE2S	-481.080514
TMEM71	-410.146107
BCAS3	-387.696784
NCOR1	-306.443161
PSMB5	-278.203498
FDPS	-241.057809
ITGAL	-221.848187
CTSB	-207.097773

Table 17. Top gene in-degree (CD4, Cd8, and B cells)

Gene	In-degree
CDKN1A	1264.518
MYC	773.7342
BCL2	376.0229
NFKB1	277.7002
FOS	256.5789
JUN	244.3791
IL6	229.9209
CDKN1B	202.8026
HIF1A	178.9729
BAX	163.3572

Table 18. Bottom gene in-degree (CD4, Cd8, and B cells)

Gene	In-degree
CDK5R1	-1270.14
MX1	-777.258
BCL2A1	-377.505
NEAT1	-276.821
FCGRT	-264.801
ITGAV	-247.321
IKBKE	-230.162
MCL1	-169.049
BBC3	-164.432
RDX	-162.263

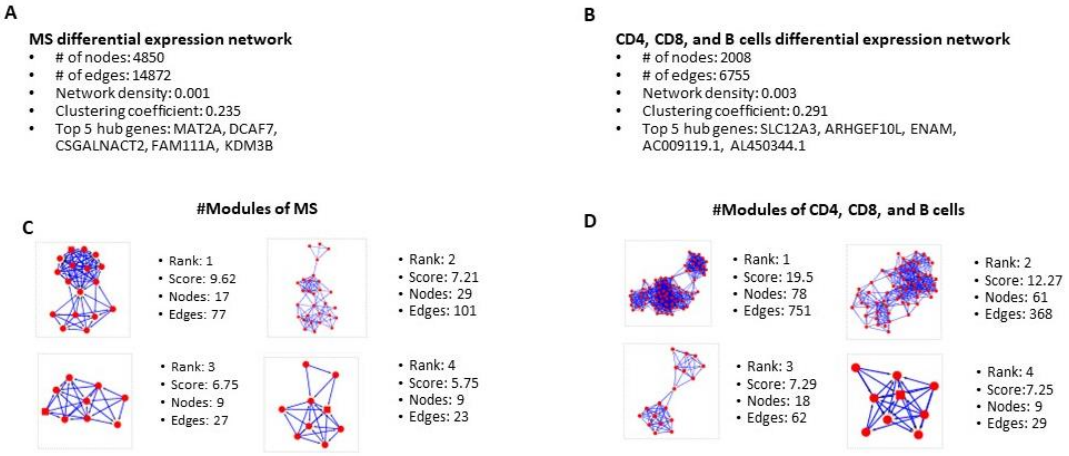


Figure 24.Characteristics of differential networks (MS vs. healthy controls) and belonging modules in innate and adaptive immunity. (A) Topological properties of the differential -expression network in Monocytes samples. (B) Topological properties of the differential -expression network in CD4, CD8, B cell samples. (C) Top four Differentially expressed modules of Monocytes. (d) Top four Differentially expressed modules of CD4, CD8, B cell samples.

Innate and adaptive immunity differential networks

Possible correlations between RNA-Seq expression profiles of innate immunity (represented by Monocytes expression) and adaptive immunity (represented by CD4, Cd8, and B cells expression) in both MS patients and healthy controls were identified employing Pearson correlation coefficients (PCCs), which resulted in four co-expression networks: innate immunity in MS patients, innate immunity in HC, adaptive immunity in MS patients, and adaptive immunity in HC.

Two differential networks, for innate and adaptive immunity, were constructed using the co-expression network (MS patients vs. HC) using HC co-expression networks as a reference. Innate immunity differential network consisted of 14872 links between 4850 genes (Fig. 24A), whereas the adaptive immunity differential expression network consisted of 6755 links between 2008 genes (Fig. 24B). Though the number

of genes exhibiting differential expression pattern was almost two-fold higher in Innate immunity, the density (0.003) and the clustering coefficient (0.291) were higher in adaptive immunity when compared to those in innate immunity (0.001and 0.231, respectively). Innate and adaptive immunity did not share mutual gene hubs.

We computed the maximum common edge subgraph between the two differential networks (innate and adaptive immunity) to point out conserved nodes between the two networks. The analysis yielded 1230 conserved nodes between innate and adaptive immunity networks. The list of the conserved nodes has been used to conduct pathway enrichment analyses of gene sets using KEGG[96] and Reactome[97] as the data sources. The result is shown in (Fig. 25).

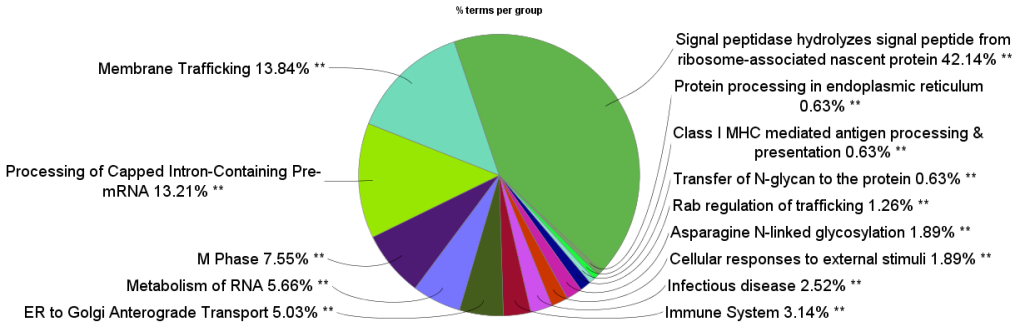


Figure25 . The distribution of the conserved genes into KEGG and Reactome pathways. P-values were determined through a 2-sided hypergeometric test and adjusted via Bonferroni's method. A threshold of adjusted p-value < 0.05 was used to determine the statistical significance of enrichment results.

Discussion

The precise roles of innate and adaptive immunity in MS are still unclear. One crucial question is whether cell types (Monocytes and CD4, CD8, and B cells) reflect the regulatory processes of their primary system (innate and adaptive immunity). By studying gene expression and gene regulatory networks, we were able to uncover patterns of transcriptional regulation that differentiate healthy and diseased states. To the best of our knowledge, this is the first study that compares the differences in regulatory networks between innate and adaptive immunity in MS.

In comparing innate immunity (represented by Monocytes cells) and adaptive immunity (represented by CD4, CD8, and B cells) gene expression, we found that cells of innate immunity (Neutrophils and Monocytes) have a different expression distribution pattern compared to cells of adaptive immunity (CD4, CD8, and B cells), which have almost identical distribution pattern (Fig. 14). PCA and Hierarchical Clustering showed a clear grouping of adaptive immune cells in both HC and MS states, which indicated a close pattern of expression of three cell types (CD4, CD8, and B cells) in MS states (Figures 15-18). PCA and Hierarchal clustering of all samples, from MS patients and HC, showed that cell lines from the two different conditions tend to cluster together, which indicates a slight change in the expression pattern in MS to the HC (Figures 19-20).

We found that SLC2A14 is the only gene shared by all cell lines (Fig. 21). SLC2A14 is a member of the glucose transport family (GLUT), which is a highly conserved integral membrane protein. [98] Shaghaghi et al. found that three alleles,

rs2889504-T, rs10846086-G, and rs10846086-G, in the SLC2A14 gene are associated with increased odds of inflammatory bowel disease. [99] Shulman et al. showed that the rs10845990 variant of SLC2A14 is associated with neurofibrillary tangles formation in the Drosophila model relevant to Alzheimer's disease. [100] The highly expressed SLC2A14 gene may indicate an increase in prefoliation activity in the blood cells and directly causative to MS since SLC2A14 is significantly expressed in blood cells compared to other tissues. [101]

We found that both systems share nine out of the top ten upregulated TFs (Fig. 23), which indicates similar TFs control over both systems. We found that the top five TFs involved in cell differentiation, cell growth, immune responses, response to DNA damage, cell cycle, and chromatin remodeling. SP1 helps with chromatin remodeling and plays a role in a variety of other processes such as cell growth, apoptosis, differentiation, and immune and DNA damage responses. [102] SP1 activation is associated with cytomegalovirus (CMV) infection, which supports the CMV infection role in MS and the association between past CMV infection with MS risk. [103–105] NF-κB regulates multiple aspects of innate and adaptive immune functions and serves as a pivotal mediator of inflammatory responses besides playing a critical role in regulating the survival, activation, and differentiation of innate immune cells and inflammatory T cells. [106] NF-κB1 or NF-κB2 is bound to REL, RELA, RELB to form the NF-κB complex, which explains the upregulated RELA. [107] Inappropriate activation of NF-κB has been linked to inflammatory events associated with autoimmune arthritis, asthma, lung fibrosis, glomerulonephritis, and atherosclerosis. [108] Bonneti et al. found that NF-κB and c-jun

transcription factors are activated in MS lesions. [109] NF-κB1 activation has been linked to the CMV virus as well as the hepatitis B virus (HBV), the hepatitis C virus (HCV), the EBV, and the influenza virus. [110] DeMeritt et al. found that virus-mediated NF-κB activation, through the dysregulation of the IκB kinases complex, plays a primary role in the initiation of the CMV gene cascade. [111] P53 responds to diverse cellular stresses to regulate target genes that induce cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism. [112] The increased expression of P53 could be secondary to oligodendrocyte injury in MS and increased apoptotic activity in the central nervous system. [113] The E2F1 transcription factor can promote proliferation or apoptosis when activated. [114] Iglesias et al. showed that E2f1-deficient mice manifested only mild disability upon induction of Experimental Autoimmune Encephalomyelitis (EAE), which is the MS model in mics. [115] Also, they showed that Peripheral Blood Mononuclear Cells (PBMCs) from Avonex-treated patients had lower expression of E2F targets. [115]

Conclusions

The present findings revealed specific expression patterns of protein-coding and upregulated TFs in MS in both innate and adaptive immunity. The knowledge of an expression network signature may offer valuable insights into the complex pathogenesis of MS; it may also provide potential targets for therapeutic intervention. All the novel changes in gene networks and TFs could be responsible for altered immune response in MS. However, many of these factors, play critical roles in cell differentiation, cell growth, immune responses, response to DNA damage, cell cycle, and chromatin

remodeling as well as in other human diseases, an original list of novel TFs and gene networks with potential implications in MS innate and adaptive immunity is offered.

For future studies, incorporating more metadata (age, sex, ethnicity…etc.) to control for the variation of gene expression is needed. More biological replicates are required to find DEGs with the minuscule difference between the groups. The top TFs list of the study must be validated using CHIP-Sequencing. Also, the association between CMV infection and MS must be explored using serological studies.

Acknowledgments

This work was supported by George Mason University - 2018 SSB Bioinformatics Summer Fellowship, George Mason University - 2019 Dissertation Completion Grant, George Mason University - 2019 Provest Office Summer Research Fellowship, and King Fahad Specialist Hospital - Dammam scholarship.

CHAPTER FOUR: MAPPING EQTLS WITH RNA-SEQ REVEALS NOVEL SUSCEPTIBILITY GENES IN MULTIPLE SCLEROSIS

Abstract

Even though Multiple Sclerosis (MS) is identified to be a partially heritable autoimmune disease, the molecular mechanisms of MS continue to remain poorly understood. While there are recognized genetics risk factors between MS patients and healthy controls, an overview of the differences between various tissues has not been conducted. RNA-sequencing (RNA-Seq) is a powerful technique for the spotting of genetic variants that affect gene expression levels. This study aimed to investigate functionally effective single nucleotide polymorphisms (SNPs) that are unique to MS using RNA-Seq based expression quantitative trait loci (eQTL) analysis in both whole blood and brain tissue. 116 gene-SNP pairs have an FDR < 2.0e-20, which are in chromosomes 1, 2, 5, 7, 17, and X have been found in the brain dataset. We offered candidate SNPs with potential implications in MS. Results from this study will provide the groundwork for the new therapy development of MS.

Background

Multiple Sclerosis

Multiple Sclerosis (MS) is an immune-mediated inflammatory disease in which the fatty myelin sheaths around the axons of the Central Nervous System (CNS) are damaged, leading to demyelination and scarring as well as a broad spectrum of signs and

symptoms.[1] MS is the second most common acquired neurological disorder of young adults, with physical trauma being the most common. The disease shows a spectrum of severity, ranging from an asymptomatic pathological process to mild symptoms to severe disabling illness. The clinical presentation involves two forms, relapsing disorder in which distinct attacks with clinical stability in between, or progressive condition in which gradual worsening of neurological deficits.

Numerous factors are thought to contribute to the cause of MS, including genetic susceptibility and environmental factors. MS affects mainly young people between the ages of 15 and 50 years, with a peak onset at about age 30. There is a substantial gender preference; most MS patients (70-75%) are women.[2]

The incidence and prevalence of MS vary throughout the world. MS affects nearly two million people worldwide with evident variability in geographic distribution.[3] Recognized low, medium, and high-risk zones have been identified. Near the equator, typically in tropical regions, there is low risk, while MS risk north and south of the equator increases with higher latitudes, in both northern and southern hemispheres.[4]

Although the pathogenesis of MS is poorly understood, evidence suggests that both genetic and environmental components play essential roles in disease development, both independently and interactively.[4] The rule of genetics in MS and its interaction with environmental triggers are currently extensively studied. MS is a disease with evident geographic variability in both prevalence and incidence. The role of environmental factors has historically been thought to be necessary. The geographical distribution and familial aggregation of MS have often been ascribed to the rule of infectious agents, but

there is no consensus regarding this theory,[5] A Canadian study examined a population-based sample of 15,000 individuals with MS using standardized, personally administered questionnaires to identify adoptees or those who had adopted relatives. The frequency of MS among first-degree, non-biological relatives living with the index case was no higher than the expected rate from the Canadian population prevalence data and was significantly less than the rate for biological relatives. These findings support the hypothesis that the familial aggregation of MS is genetically determined rather than environmentally determined.[5,6] A significant contributor to the genetic risk is the major histocompatibility complex (MHC) antigen.[7]

The contribution of genetics to MS is supported by many reports showing familial aggregation of the disease, high concordance rates among twins, and increased risk among relatives of patients with MS. People with MS have a 5–26% chance of having one or more affected relatives, which is a much higher chance than one would expect for a disease with no genetic component. [5,6] Furthermore, the relative risk of MS for identical twins, if one is affected, is approximately 200 to 300 times greater than that of the general population.[6,8] Finally, the first-degree relatives of MS patients have a 2–5% risk of also developing the disease.[8]

Several alleles have been identified as heritable risk factors for MS. Genetic complexity, primarily related to human leukocyte antigens (HLA) of the MHC and, to a lesser extent, non-MHC-related genes, plays a significant role in influencing disease susceptibility, phenotypic expression CD4 T cells, which have central role in MS pathogenesis, [22] experience profound changes in gene expression during the initial hours

after activation. Co-stimulation via the CD28 receptor is required for the effective activation of naive T cells. [23] From pre-clinical studies, it is known that the transcription factor is highly induced in a CD28-dependent manner upon T cell activation and is involved in essential CD4 effector T cell functions, participating in the transcriptional regulation of several T cell activation pathways, including a large group of CD28-regulated genes. [24] Furthermore, Levels of blood monocytes secreting IL-6 and IL-12 were higher in patients with untreated MS and other neurological diseases compared to healthy controls, and MS patients blood monocytes also displayed elevated mean fluorescence intensity for the co-stimulatory molecule CD86, and MS patients with longer disease duration (>10 years) and higher disease severity had higher percentages of CD80 expressing monocytes compared to patients with short duration or lower severity.[25]

Mapping QTLs with RNA-seq

Genome Wide Association Studies (GWAS) have effectively identified many genetic loci that play a part in complex-disease susceptibility in humans.[116] Abundant expression quantitative trait loci (eQTL) mapping studies have since been conducted to investigate diseases[117–120], cell-types[121–123], and response to several environmental stimuli.[124] A great restraint on most of such investigations is the use of 3´-targeted microarrays to profile gene expression. Splicing events effect is not likely to be detected,[125], which might explain the limited susceptibility loci localized to causal eQTL signals.[126]

RNA-Seq has been compared to microarray technology as a means of generating transcriptome information since its launch.[26] Both follow a parallel path to answering a

biological question. Nevertheless, there are a few key advantages of RNA-Seq technology, including exploring novel genes, lower noise, and a broad dynamic range of expression levels.[26] RNA-Seq based eQTL mapping studies are started to arise[127], which will significantly increase the likelihood of catching disease-associated eQTLs as per quantification of independent exon expression, as well as relative transcript abundance (novel isoforms).[128–130]

Rationale

Understanding the complex molecular mechanism of MS, including the role of functional SNPs, is crucial to develop effective therapies. This work aims to uncover SNPs that potentially dysregulate many genes and altered the expression pattern in MS using both whole blood and brain tissues. Ultimately, our goal is to investigate if SNPs have a different effect on the cell-specific model in MS. The results could be a ground for further investigating functional SNPs. Integration of biological data (DNA variations, and transcript information) provides valuable insights on the clarification of the disease mechanisms and identification of molecular signatures of human diseases.

Methods

We focused on pointing genetics variants and functional SNPs that play an important role in MS complex molecular pathogenesis from Monocytes cells and Whole Blood cells and brain cells samples.

RNA-Seq data

We obtained the raw expression files (FASTQ format) of whole blood, for both treatment naïve MS patients and HC, from ArrayExpress[31] database (Accession code E-GEOD-66573). A total of 14 samples (8 for HC and 6 for MS patients) was used in the analysis (Table 20). We obtained the raw expression files (FASTQ format) of brain samples, for both treatment naïve MS patients and HC, from Gene Expression Omnibus[131] (Series GSE123496). A total of 50 samples of different brain regions obtained from 10 subjects (5 for HC and 5 for MS patients) was used in the analysis (Table 21).

Table 19. Whole Blood RNA-seq samples information for MS patients and healthy controls , which includes the condition, RNA-seq sample run ID (SRR#), and cell type.

ID	Run	Condition
1	SRR1839791	N
2	SRR1839794	N
3	SRR1839799	N
4	SRR1839800	N
5	SRR1839801	N
6	SRR1839802	N
7	SRR1839803	N
8	SRR1839804	N
9	SRR1839792	MS
10	SRR1839793	MS
11	SRR1839795	MS
12	SRR1839796	MS
13	SRR1839797	MS
14	SRR1839798	MS

Table 20. Brain RNA-seq samples information for MS patients and healthy controls , which includes the condition, RNA-seq sample run ID (SRR#), and cell type.

ID	Run	tissue	condition

1	SRR8307929	corpus callosum	MS
2	SRR8307930	frontal cortex	MS
3	SRR8307931	parietal cortex	MS
4	SRR8307932	hippocampus	MS
5	SRR8307933	internal capsule	MS
6	SRR8307934	corpus callosum	MS
7	SRR8307935	frontal cortex	MS
8	SRR8307936	parietal cortex	MS
9	SRR8307937	hippocampus	MS
10	SRR8307938	internal capsule	MS
11	SRR8307939	corpus callosum	MS
12	SRR8307940	frontal cortex	MS
13	SRR8307941	parietal cortex	MS
14	SRR8307942	hippocampus	MS
15	SRR8307943	internal capsule	MS
16	SRR8307944	corpus callosum	MS
17	SRR8307945	frontal cortex	MS
18	SRR8307946	parietal cortex	MS
19	SRR8307947	hippocampus	MS
20	SRR8307948	internal capsule	MS
21	SRR8307949	corpus callosum	MS
22	SRR8307950	frontal cortex	MS
23	SRR8307951	parietal cortex	MS
24	SRR8307952	hippocampus	MS
25	SRR8307953	internal capsule	MS
26	SRR8307954	corpus callosum	N
27	SRR8307955	frontal cortex	N
28	SRR8307956	parietal cortex	N
29	SRR8307957	hippocampus	N
30	SRR8307958	internal capsule	N
31	SRR8307959	corpus callosum	N
32	SRR8307960	frontal cortex	N
33	SRR8307961	parietal cortex	N
34	SRR8307962	hippocampus	N
35	SRR8307963	internal capsule	N
36	SRR8307964	corpus callosum	N
37	SRR8307965	frontal cortex	N
38	SRR8307966	parietal cortex	N
39	SRR8307967	hippocampus	N
40	SRR8307968	internal capsule	N
41	SRR8307969	corpus callosum	N

42	SRR8307970	frontal cortex	N
43	SRR8307971	parietal cortex	N
44	SRR8307972	hippocampus	N
45	SRR8307973	internal capsule	N
46	SRR8307974	corpus callosum	N
47	SRR8307975	frontal cortex	N
48	SRR8307976	parietal cortex	N
49	SRR8307977	hippocampus	N
50	SRR8307978	internal capsule	N

RNA-Seq data analysis

The quality control on raw reads has been performed using the FastQC[34] tool (Version 0.11.7) for each sample. Raw reads refinement, and clipping have been performed using Trimmomatic[35] (Version 0.36). High-quality reads have been mapped to the reference human genome (GRCh38/hg38) and the human reference transcriptome (Ensembl v70) from the Ensembl[36] genome database using STAR[37] Aligner (Version 2.4.0.1). Unique mapped reads have been used to quantify gene expression in each sample. Gene expression, as reads counts, have been estimated after filtering and normalization of raw reads counts using Rsubread[38] and DEseq2[39] R packages as well as differential expression analysis between MS and HC samples. Multiple testing correction has been performed using Benjamini-Hochberg[40]. Genes with adjusted p-values less than alpha of 0.05 were considered differentially expressed. Clustering, Principal Component Analysis (PCA) of the significant gene list were conducted using the same package.

Variant calling and processing pipeline

GATK best practice[132] workflow for single-nucleotide polymorphism (SNP) and insertion or deletion (Indel) calling on RNA-Seq has been followed to generate raw VCF

files. The human genome (GRCh38/hg38) from the Ensembl[36] genome database has been used as a reference genome for all two datasets. We used bedtools[133] intersect function to find the overlapped variants between samples that are having the same condition, to find variant calls unique to MS intersected samples compared to HC samples for both Whole Blood and Brain datasets. SAMtools / BCFtools[134] have been used to omit duplicated calls, and to filter variants. GATK best practice[132] generic recommendations (QD < 2.0, Q < 40.0, FS > 60.0, SOR > 3.0, MQRankSum < 12.5, ReadPosRankSum < -8.0) have been used as filtering parameters. We used Ensembl database[135] and Ensembl Variant Effect Predictor (VEP)[136] server to annotate, generate summary statistics, and filter variants based in MAF. We used the eulerr[137] tool to generate the Venn diagram.

RNA-Seq based eQTL mapping

The quality control on raw reads has been performed using the FastQC[34] tool (Version 0.11.7) for each sample. Raw reads refinement, and clipping have been performed using Trimmomatic[35] (Version 0.36). High-quality reads have been mapped to the reference human genome (Homo_sapiens.GRCh38.dna.primary_assembly) and the human reference transcriptome (Homo_sapiens.GRCh38.96) from Ensembl[36] genome database using STAR[37] Aligner (Version 2.4.0.1). Unique mapped reads have been used to quantify gene expression in each sample. Gene expression, as reads counts, have been estimated after filtering and normalization of raw reads counts using Rsubread[38] and edgeR[41] R packages. We used the gread[138] R package to extract the common gene annotations (Ensembl ID), and positions form the reference transcriptome file. Picard tools[139] (Version 2.21.1) has been used to sort the bam files. The SAMtools[140] mpileup

function has been utilized to joint call variants of all samples. BCFtools[134] has been used to report SNPs only from the joint called file. VariantAnnotation[141] R package has been used to construct the SNP matrix. We used the MatrixEQTL[142] R package for computational eQTL analysis. Each genotype variable has been treated as categorical, and we modeled its effect on gene expression with a linear regression model, assuming that the noise to be independent and identically distributed across samples. A gene-SNP pair is considered local if the distance between them is less than 1000000 base-pair. Dplyr[143] and data.table[144] R packages have been used for data manipulation.

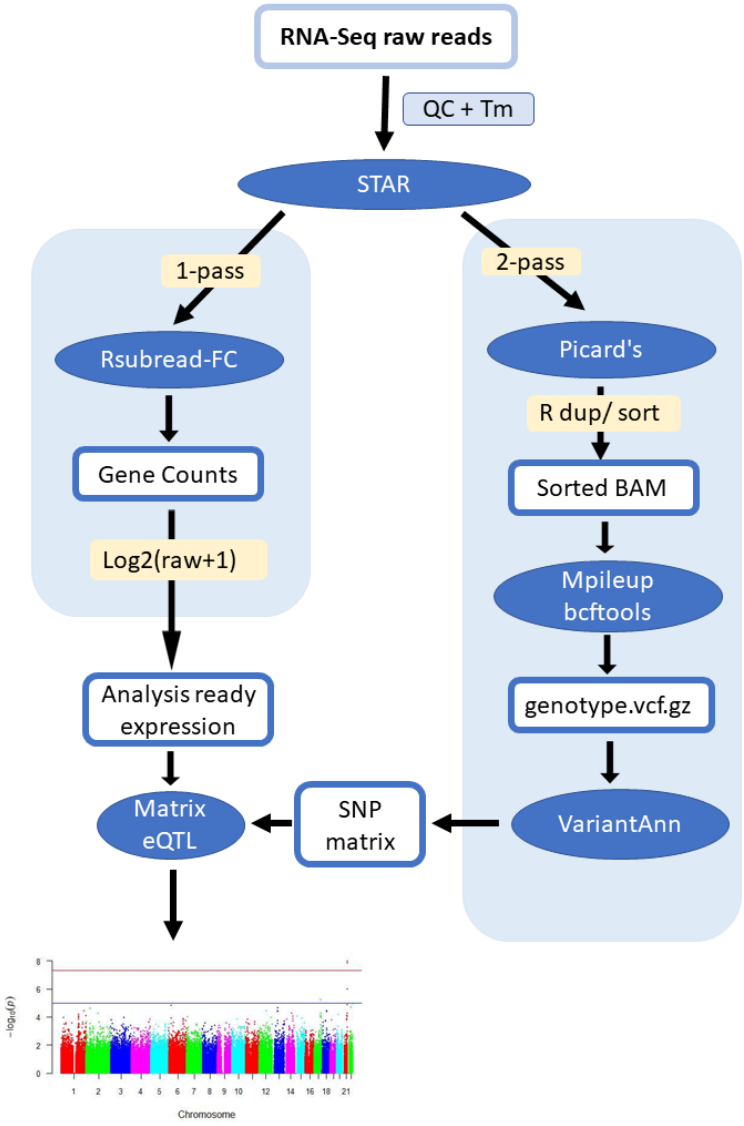


Figure 26. RNA-Seq based eQTL analysis. The analysis started with assessing the quality of the raw FASTAQ files that have been generated by the sequencing machine using FastQC[34] tool; the data are then processed using Trimmomatic[35] tool. To create transcriptome information, High-quality reads have been mapped to the reference human genome and the human reference transcriptome from the Ensembl[36] genome database using STAR[37] Aligner 1-pass mode. Gene expression, as reads counts, have been estimated after filtering and normalization of raw reads counts using Rsubread[38] and edgeR[41] R packages. To generate variants information, we used STAR[37] Aligner 2-pass mode followed by Picard tools[139] to sort the bam files and remove duplicates. The SAMtools[140] mpileup function has been utilized to joint call variants of all samples. BCFtools[134] has been used to report SNPs only from the joint called file. VariantAnnotation[141] R package has been used to construct the SNP matrix. For the eQTL analysis, We used the MatrixEQTL[142] R package for computational eQTL analysis. Each genotype variable has been treated as categorical, and we modeled it effect on gene expression with a linear regression model.

Results

Unique variants by calling pipeline

After the first filtration process (QD < 2.0, Q < 40.0, FS > 60.0, SOR > 3.0, MQRankSum < 12.5, ReadPosRankSum < -8.0), we ended up with 27827 variants(10.7% Novel / 89.3% known) for the brain dataset, and 27094 variants (44.3% Novel 55.7% known) for the whole blood dataset. 914 variants found to be shared between the two datasets (Fig. 26). Variants located in intron and downstream gene regions were more prominent in both datasets, but the coding consequences differ drastically (Fig. 27). In the brain dataset, the coding predicted mainly to have no effect at all (59% synonymous variants) or a slight effect (39% missense variants). On the other hand, a high percentage of the variants in the blood dataset are predicted to have delirious effects (30% frameshift variants).

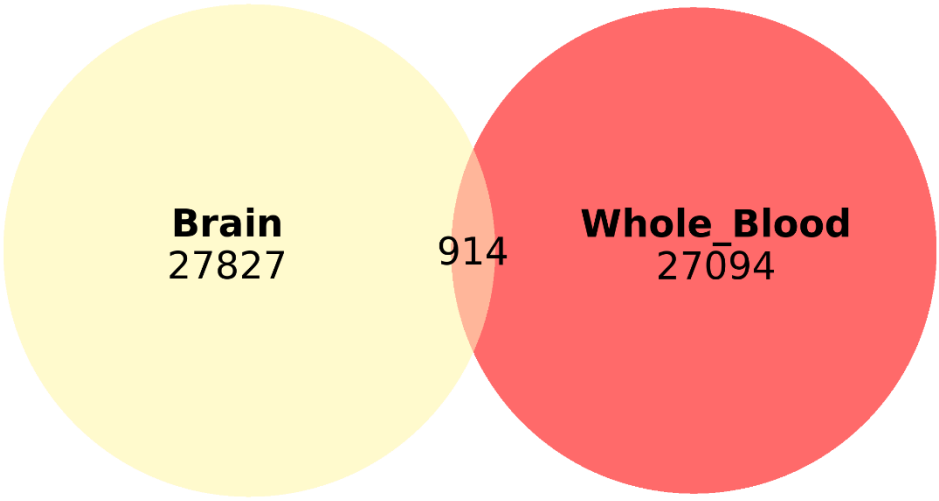


Figure 27. Venn diagram of the number of variants that satisfies the following conditions (QD < 2.0, Q < 40.0, FS > 60.0, SOR > 3.0, MQRankSum < 12.5, ReadPosRankSum < -8.0) in both data sets. 914 variants found to be shared by both.

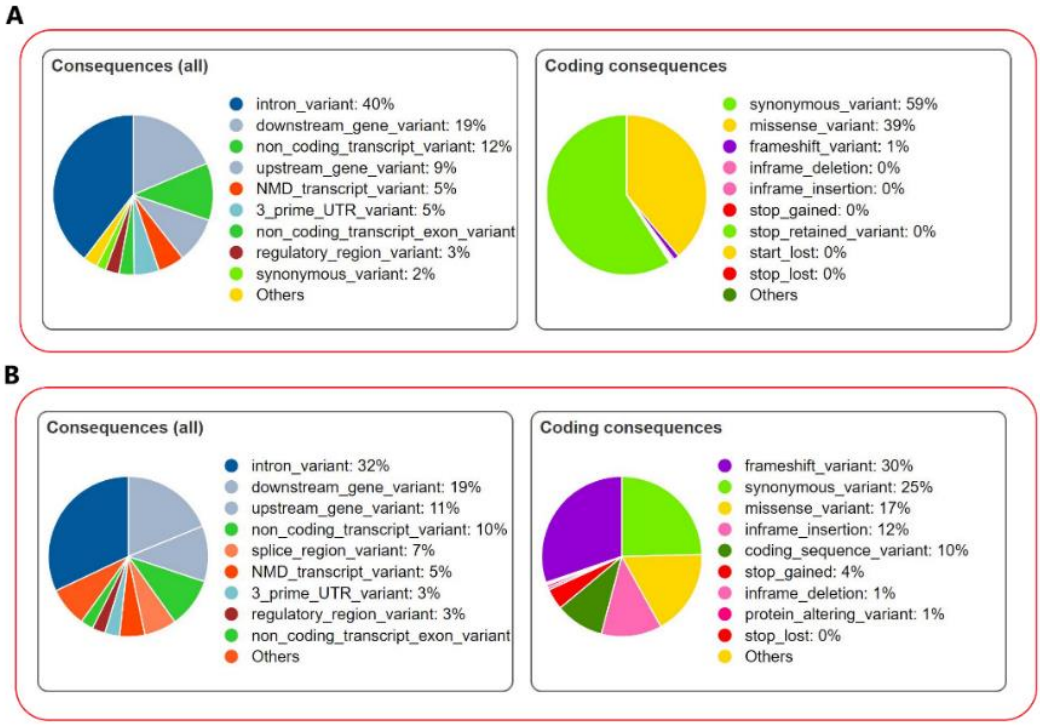


Figure 28. Predicted location(all) and coding consequences of brain dataset variants (A), and whole blood dataset (B). variants located in intron and downstream gene regions were more prominent in both datasets. The coding consequences differ drastically between both datasets.

eQTL analysis hits

In the brain dataset, 94943 gene-SNP pair found to be significant (FDR < 0.05). 116 gene-SNP pair have an FDR < 2.0e-20, which are in chromosomes 1, 2, 5, 7, 17, and X. The top 20 pairs are included in table 21. Both X:153669900_G/T and X:154420998_G/A are associated with ENSG00000126890 (HGNC: CTAG2). 17:35574149_A/G, 17:35724277_G/A, 17:35743280_G/A are associated with ENSG00000261499 (HGNC: NPEPPS). 15 different SNPs are associated with ENSG00000263503 (HGNC: MAPK8IP1P2) (table 21).

In blood samples, only three gene-SNP pair hits have an FDR <0.1. All were in chromosome 1. 1:89105409_T/G found to be associated with ENSG00000284734 (HGNC: AC099063.4). 1:137159_T/C and 1:137159_T/C found to be associated with ENSG00000229344 (HGNC: MTCO2P12).

Table 21. Top 20 gene-SNP pair hits of brain samples (based on FDR corrected p-values). SNPs include the chromosome, the location of the variant in the chromosome. Test statistics were computed using a linear regression model. We used the Benjamini Hochberg method as a P-value adjustment method and to calculate the FDR.

SNPs	Ensembl gene ID	Statistic	p-value	FDR
X:153669900_G/T	ENSG00000126890	1148.325	5.50E-40	1.42E-33
X:154420998_G/A	ENSG00000126890	756.4182	6.44E-36	6.96E-30
17:35574149_A/G	ENSG00000261499	731.8077	1.35E-35	6.96E-30
17:35724277_G/A	ENSG00000261499	731.8077	1.35E-35	6.96E-30
17:35743280_G/A	ENSG00000261499	731.8077	1.35E-35	6.96E-30
17:45985549_G/T	ENSG00000263503	621.2764	5.14E-34	2.21E-28
17:45645823_G/A	ENSG00000263503	595.7799	1.30E-33	3.49E-28
17:46007310_C/T	ENSG00000263503	595.7799	1.30E-33	3.49E-28
17:45436075_C/T	ENSG00000263503	587.6616	1.76E-33	3.49E-28
17:46002673_T/G	ENSG00000263503	573.0078	3.08E-33	3.49E-28
17:45639519_A/G	ENSG00000263503	566.1877	4.02E-33	3.49E-28
17:46003698_A/G	ENSG00000263503	566.1877	4.02E-33	3.49E-28
17:45637652_T/C	ENSG00000263503	561.0187	4.92E-33	3.49E-28
17:45981350_G/T	ENSG00000263503	559.6198	5.20E-33	3.49E-28
17:46000342_G/T	ENSG00000263503	557.7623	5.59E-33	3.49E-28
17:45436185_C/G	ENSG00000263503	557.6049	5.63E-33	3.49E-28
17:45632049_C/T	ENSG00000263503	555.6837	6.08E-33	3.49E-28
17:45636559_A/G	ENSG00000263503	555.6837	6.08E-33	3.49E-28
17:45641777_A/G	ENSG00000263503	555.6837	6.08E-33	3.49E-28
17:45988535_C/T	ENSG00000263503	555.6837	6.08E-33	3.49E-28

Table 22. Top 3 gene-SNP pair hits of whole blood samples (based on FDR corrected p-values). SNPs include the chromosome, the location of the variant in the chromosome. Test statistics were computed using a linear regression model. We used the Benjamini Hochberg method as a P-value adjustment method and to calculate the FDR.

SNPs	Ensembl gene ID	Statistic	p-value	FDR

1:89105409_T/G	ENSG00000284734	-13.8902	2.55E-08	0.073374
1:137159_T/C	ENSG00000229344	13.42196	3.65E-08	0.073374
1:137159_T/C	ENSG00000198744	12.59958	7.04E-08	0.094289

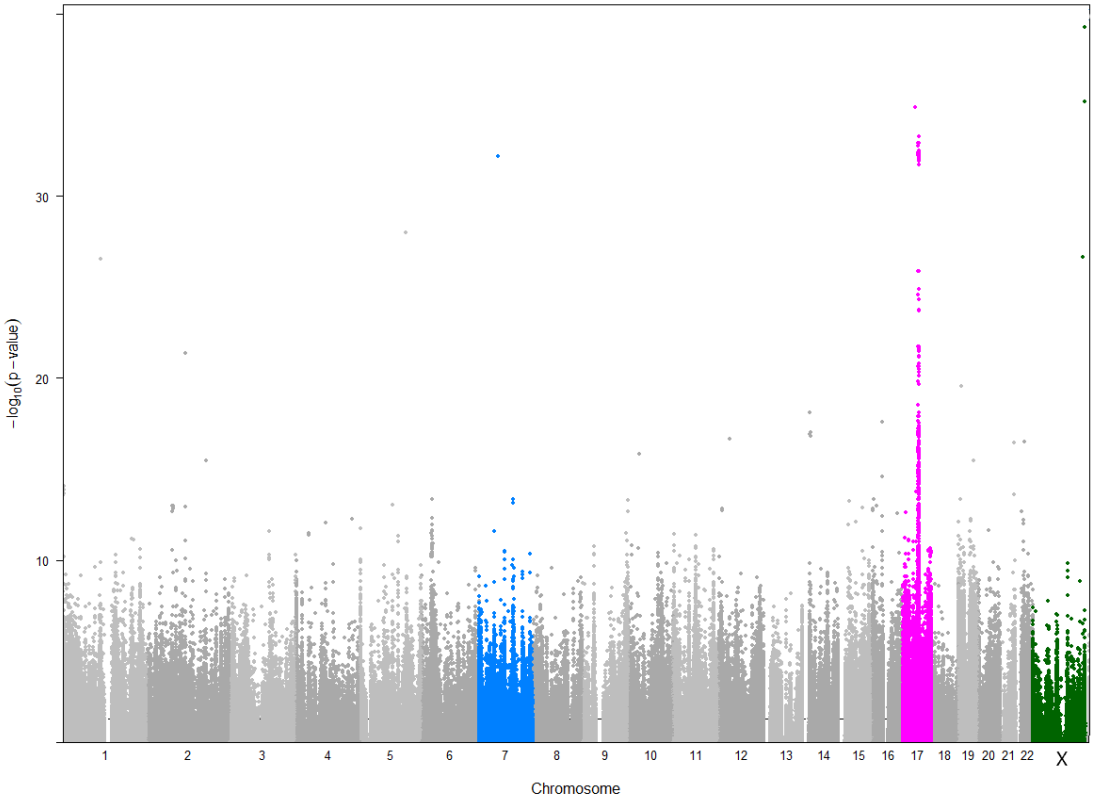


Figure 29. Manhattan plot SNP pair hits of brain samples.

Discussion

The precise effect of SNPs on cell-specific gene expression in MS is still unclear. This work aims to uncover SNPs that potentially dysregulate many genes and altered the expression pattern in MS using both whole blood and brain tissues. By studying the link

between gene expression and SNPs in different tissues, we were able to uncover patterns SNP-Gene links in MS. To the best of our knowledge, this is the first study that compares the differences in regulatory networks between innate and adaptive immunity in MS.

In the brain dataset, 116 gene-SNP pair have an FDR < 2.0e-20, which are in chromosomes 1, 2, 5, 7, 17, and X. two specific variants on the X chromosome (X:153669900_G/T and X:154420998_G/A) found to be linked to changed expression of CTAG2 gene expression. This protein is expressed by many human cancers, but not by normal tissues, with the exception of testis and placenta.[145] Zarour et al., found that the CTAG2 gene has an immunogenic role, and its products have the capability to stimulate T-helper 1 type CD4+ T cells. [146] This finding may support the role of CD4+ T cells in the pathogenesis of MS. [147] The CTA New York Esophageal Squamous Cell Carcinoma-1 (NY-ESO-1) antigen, which is encoded by the gene CTGAG1B, is widely believed to be a good candidate target for immunotherapy and some promising results have been obtained in early phase I/II studies. [148] Immunotherapy targeting NY-ESO-1 could be tested in patients with MS. NPEPPS is a protein-coding gene, which codes for Puromycin-sensitive aminopeptidase. It could be identified in cortical and cerebellar neurons and its part of Class I MHC mediated antigen processing and presentation and Innate Immune System pathways. [149]

In the blood dataset, only three gene-SNP pair hits have an FDR <0.1. This is mainly due to the limited number of samples used in the analysis. AC099063.4 is the Antisense RNA that controls the expression of the Guanylate Binding Protein 4 (GBP4) gene. [150] GBP4 is part of NOD-like receptor signaling pathway pathways, which are

involved in the pathogenesis of MS. [151] Berben et al. found that ubiquitin ligase *Peli1* knock-out experimental autoimmune encephalomyelitis mice had less inflammation in the central nervous system. [152] Several proteins related to the interferon signaling pathway were among the most upregulated in the *Peli1* knock-out mice compared to the wild type, such as IFIT3, IRGM1, and the GTPases IIGP1, GBP2, and GBP4. [152]

Conclusions

The present findings revealed a candidate SNPs that have functional implications in MS. The knowledge about the functional role of SNPs may offer valuable insights into the complex pathogenesis of MS; it may also provide potential targets for therapeutic intervention. Current immunotherapy, like the one targeting NY-ESO-1, could be used in a clinical-based study to see its effect on the course of MS.

For future studies, incorporating more metadata (age, sex, ethnicity…etc.) to control for the variation of gene expression is needed. More biological replicates are required to have enough power to link the SNPs with expression patterns in the eQTL analysis.

Acknowledgments

This work was supported by George Mason University - 2018 SSB Bioinformatics Summer Fellowship, George Mason University - 2019 Dissertation Completion Grant, George Mason University - 2019 Provest Office Summer Research Fellowship, and King Fahad Specialist Hospital - Dammam scholarship.

CHAPTER FIVE: CONCLUSION AND FUTURE DIRECTION

Conclusions

This dissertation explored different computational methods, aimed to explore different genomic systems, for analyzing RNA-Seq expression data obtained from MS patients. The main goal is a better understanding of the complex molecular mechanism of MS and, hopefully, a groundwork for the new therapy development of MS. We ended up with candidates' genes, TFs, and SNPs with potential implications in MS.

One main drive of this dissertation is using RNA-Seq expression data to represent transcriptome change in all studies. We choose RNA-Seq expression data and not microarray, which another transcriptome information generator because RNA-Seq has few important advantages. [26] First, using microarray technology limits the researcher in spotting transcripts that linked to existing genomic sequencing information. RNA-Seq experiments, instead, work well for examining both known transcripts and explore new ones.[26] Second, RNA-seq delivers a low background signal because DNA sequences can be unambiguously mapped to unique regions of the genome. As a result, noise in the experiment is effortlessly eliminated during analysis. Hybridization issues seen with microarrays, such as cross-hybridization or non-ideal hybridization kinetics, are also removed in RNA-Seq experiments, which offers another signal-to-noise advantage.[26] Finally, RNA-seq can quantify a broad dynamic range of expression levels, with absolute rather than relative values.[26]

In the first study, we investigated specific transcriptome changes occurring in monocytes of patients with MS compared to Healthy Controls (HC) patients. Monocytes have been chosen because most of the previous transcriptome studies covered adaptive immunity cells (B, CD4, and CD8). Also, it has been experimentally proven that monocytes have a central role in MS pathogenesis. Yamasaki et al. found that Monocyte-derived macrophages initiate demyelination at disease onset in the experimental autoimmune encephalomyelitis (EAE) model. [153] Data analysis revealed that 6120 genes were significantly altered between the two groups (16% up-regulated and 17% down-regulated in MS group compared to healthy controls). The main findings of the study are: 1) The identification of specific changes occurring in the transcriptome of MS patients compared to healthy subjects. In particular, the bioinformatic analysis revealed significant alteration of the Chemokine signaling pathway, Jak-STAT signaling pathway, Toll-like receptor signaling pathway, NOD-like receptor signaling pathway. The KEGG hsa04062 Chemokine signaling pathway was the most significant up-regulated pathway in the functional scoring analysis. ; 2) The identification of specific changes occurring in the transcriptome of MS patients compared to healthy subjects not previously linked to MS. In particular, Osteoclast differentiation and Endocytosis; 3) The identification of some genes, which previous studies have not reported them, had an association with MS (RPS4Y1, XIST, DDX3Y, KDM5D, EIF1AY, and TXLNGY). 4) The number of the top ten DEGs based on p.adjusted value is first to be linked directly to MS (RPS4Y1, XIST, KDM5D, KDM6A, TXLNGY, UTY, PRKY). 5) the number of the top ten DEGs based

on p.adjusted value that are involved in gene regulation (RPS4Y1, XIST, DDX3Y, KDM5D, KDM6A, EIF1AY, ZFY).

In the second study, we focused on the Transcription Factors (TFs) that are important in MS. experimental Evidence suggests that TFs are playing a role in the pathogenesis of MS and other autoimmune diseases. [9] the goal of this study is to infer upregulated TFs of both innate (Monocytes) and adaptive immunity (B, CD4, and CD8) in MS. We did that to Infer TFs networks from expression data and knowledge-based bipartite networks using a message-passing algorithm. [78] We found that both adaptive immunity and innate immunity share nine out of the top ten upregulated TFs (SP1, NFKB1, RELA, TP53, E2F1, STAT3, MYC, JUN, YY1) Also, We found that the top five TFs involved in cell differentiation, cell growth, immune responses, response to DNA damage, cell cycle, and chromatin remodeling, all have been proven experimentally to be linked to MS.[9,10,13] all these findings indicates similar TFs control over both systems.

The third study was focused on investigating functionally effective single nucleotide polymorphisms (SNPs) that are unique to MS using RNA-Seq based expression quantitative trait loci (eQTL) analysis. The goal was to compare the WB and brain tissues, and to examine if SNPs have a different effect on the cell-specific model in MS. we found that brain dataset has 116 gene-SNP pair have, which have an FDR < 2.0e-20, and are in chromosomes 1, 2, 5, 7, 17, and X. Genes included CTAG2, which has been found by Zarour et al. that it has an immunogenic role, and its products have the capability to stimulate T-helper 1 type CD4+ T cells and support the role of CD4+ T cells

in the pathogenesis of MS. [146,147] Also, CTA New York Esophageal Squamous Cell Carcinoma-1 (NY-ESO-1), which is encoded by the gene CTGAG1B, and widely believed to be a good candidate target for immunotherapy, and some promising results have been obtained in early phase I/II studies. [148] In the blood dataset, only three gene-SNP pair hits have an FDR <0.1. This is mainly due to the limited number of samples used in the analysis. The genes included AC099063.4, which is the Antisense RNA that controls the expression of the Guanylate Binding Protein 4 (GBP4) gene. [150] GBP4 is part of NOD-like receptor signaling pathway pathways, which are involved in the pathogenesis of MS. [151] Berben et al. found that ubiquitin ligase *Peli1* knock-out experimental autoimmune encephalomyelitis mice had less inflammation in the central nervous system. [152] Several proteins related to the interferon signaling pathway were among the most upregulated in the *Peli1* knock-out mice compared to the wild type, such as IFIT3, IRGM1, and the GTPases IIGP1, GBP2, and GBP4. [152]

Future Direction

For future studies, we need to incorporate more detailed metadata to control for the variation of gene expression between samples. Variables like age, sex, ethnicity have been experimentally proven to influence expression in different conditions. Viñuela et al. found evidence that up to 60% of age effects on transcription levels shared across tissues, and 47% of those on splicing. [154] Dillman et al. demonstrated that there are robust age-related alterations in gene expression in the human brain and that genes encoding for neuronal synaptic function may be particularly sensitive to the aging process. [155] Gal-Oz

et al. detected a clear differential expression pattern of genes coding for competent of macrophages from three different tissues, which may explain the strong activation of innate immune pathways prior to pathogen invasion in females. [156]

More biological replicates in each arm (MS and healthy participants) are needed for future studies. For example, only four biological replicates have been used in each arm in the first study, which will identify 40%–60% of the significantly differentially expressed (SDE) genes. [157]We need at least 20 biological replicates To achieve >85% for all SDE genes regardless of fold change. [157]

Experimental studies can be used to validate the results of all studies. reverse transcription-polymerase chain reaction/real-time polymerase chain reaction combined technique (qRT-PCR) is the method of choice to validate the top DEGs list of the first study.). [158] G protein-coupled receptor kinase 2 (GRK2), a crucial part of the top upregulated KEGG pathway, found to be downregulated in Relapsing-remitting MS (RRMS) patients compared to stroke patients and healthy controls. [159] GRK levels can be assessed experimentally using cell cultures obtained from MS patients and healthy controls. [159] For the second study, Chromatin Immunoprecipitation Sequencing (CHIP-Sequencing) can be used to validate the upregulated TFs in both innate and adaptive immunity. [160] Also, the association between CMV infection and MS may be explored using serological studies. [105] In the third study, Current immunotherapy, like the one targeting NY-ESO-1, could be used in a clinical-based study to see its effect on the course of MS.

REFERENCES

1. Compston A, Coles A. Multiple sclerosis. *Lancet*. 2008;372(9648):1502-1517. doi:10.1016/S0140-6736(08)61620-7
2. Duquette P, Pleines J, Girard M, Charest L, Senecal-Quevillon M, Masse C. The increased susceptibility of women to multiple sclerosis. *Can J Neurol Sci*. 1992;19(4):466-471. http://www.ncbi.nlm.nih.gov/pubmed/1423044.
3. Zwibel HL, Smrtka J. Improving quality of life in multiple sclerosis: an unmet need. *Am J Manag Care*. 2011;17 Suppl 5 Improving:S139-45. http://www.ncbi.nlm.nih.gov/pubmed/21761952. Accessed April 9, 2019.
4. Ebers GC. Environmental factors and multiple sclerosis. *Lancet Neurol*. 2008;7(3):268-277. doi:10.1016/S1474-4422(08)70042-5
5. Ebers GC, Sadovnick AD, Risch NJ. A genetic basis for familial aggregation in multiple sclerosis. *Nature*. 1995;377(6545):150-151. doi:10.1038/377150a0
6. Sadovnick AD, Ebers GC, Dyment DA, Risch NJ. Evidence for genetic basis of multiple sclerosis. The Canadian Collaborative Study Group. *Lancet (London, England)*. 1996;347(9017):1728-1730. http://www.ncbi.nlm.nih.gov/pubmed/8656905.
7. Haines JL, Terwedow HA, Burgess K, et al. Linkage of the MHC to familial multiple sclerosis suggests genetic heterogeneity. The Multiple Sclerosis Genetics Group. *Hum Mol Genet*. 1998;7(8):1229-1234. http://www.ncbi.nlm.nih.gov/pubmed/9668163.

8. Willer CJ, Dyment DA, Risch NJ, Sadovnick AD, Ebers GC, Canadian Collaborative Study Group. Twin concordance and sibling recurrence rates in multiple sclerosis. *Proc Natl Acad Sci*. 2003;100(22):12877-12882. doi:10.1073/pnas.1932604100

9. Eggert M, Klüter A, Zettl UK, Neeck G. Transcription factors in autoimmune diseases. *Curr Pharm Des*. 2004;10(23):2787-2796. http://www.ncbi.nlm.nih.gov/pubmed/15379667.

10. Peng SL. Transcription factors in autoimmune diseases. *Front Biosci*. 2008;13:4218-4240. http://www.ncbi.nlm.nih.gov/pubmed/18508507.

11. Fortunato G, Calcagno G, Bresciamorra V, et al. Multiple Sclerosis and Hepatitis C Virus Infection Are Associated with Single Nucleotide Polymorphisms in Interferon Pathway Genes. *J Interf Cytokine Res*. 2008;28(3):141-152. doi:10.1089/jir.2007.0049

12. Taki S. Type I interferons and autoimmunity: lessons from the clinic and from IRF-2-deficient mice. *Cytokine Growth Factor Rev*. 13(4-5):379-391. http://www.ncbi.nlm.nih.gov/pubmed/12220551.

13. Kristjansdottir G, Sandling JK, Bonetti A, et al. Interferon regulatory factor 5 (IRF5) gene variants are associated with multiple sclerosis in three distinct populations. *J Med Genet*. 2008;45(6):362-369. doi:10.1136/jmg.2007.055012

14. Kuipers HF, Biesta PJ, Montagne LJ, van Haastert ES, van der Valk P, van den Elsen PJ. CC chemokine receptor 5 gene promoter activation by the cyclic AMP response element binding transcription factor. *Blood*. 2008;112(5):1610-1619.

doi:10.1182/blood-2008-01-135111

15. Nguyen VT, Benveniste EN. Critical Role of Tumor Necrosis Factor-α and NF-κB in Interferon-γ-induced CD40 Expression in Microglia/Macrophages. *J Biol Chem*. 2002;277(16):13796-13803. doi:10.1074/jbc.M111906200

16. Klotz L, Schmidt S, Heun R, Klockgether T, Kölsch H. Association of the PPAR$γ$ gene polymorphism Pro12Ala with delayed onset of multiple sclerosis. *Neurosci Lett*. 2009;449(1):81-83. doi:10.1016/j.neulet.2008.10.066

17. Grekova MC, Salerno K, Mikkilineni R, Richert JR. Sp3 expression in immune cells: a quantitative study. *Lab Invest*. 2002;82(9):1131-1138. http://www.ncbi.nlm.nih.gov/pubmed/12218073.

18. Montes M, Zhang X, Berthelot L, et al. Oligoclonal myelin-reactive T-cell infiltrates derived from multiple sclerosis lesions are enriched in Th17 cells. *Clin Immunol*. 2009;130(2):133-144. doi:10.1016/j.clim.2008.08.030

19. Avasarala JR, Chittur S V, George AD, Tine JA. Microarray analysis in B cells among siblings with/without MS - role for transcription factor TCF2. *BMC Med Genomics*. 2008;1(1):2. doi:10.1186/1755-8794-1-2

20. Gerhauser I, Alldinger S, Baumgärtner W. Ets-1 represents a pivotal transcription factor for viral clearance, inflammation, and demyelination in a mouse model of multiple sclerosis. *J Neuroimmunol*. 2007;188(1-2):86-94. doi:10.1016/j.jneuroim.2007.05.019

21. Venken K, Hellings N, Hensen K, et al. Secondary progressive in contrast to relapsing-remitting multiple sclerosis patients show a normal $CD4^+$ $CD25^+$

regulatory T-cell function and FOXP3 expression. *J Neurosci Res*. 2006;83(8):1432-1446. doi:10.1002/jnr.20852

22. Hollenbach JA, Oksenberg JR. The immunogenetics of multiple sclerosis: A comprehensive review. *J Autoimmun*. 2015;64:13-25. doi:10.1016/j.jaut.2015.06.010

23. Lenschow DJ, Walunas TL, Bluestone JA. CD28/B7 SYSTEM OF T CELL COSTIMULATION. *Annu Rev Immunol*. 1996;14(1):233-258. doi:10.1146/annurev.immunol.14.1.233

24. Martínez-Llordella M, Esensten JH, Bailey-Bucktrout SL, et al. CD28-inducible transcription factor DEC1 is required for efficient autoreactive CD4 $^{+}$T cell response. *J Exp Med*. 2013;210(8):1603-1619. doi:10.1084/jem.20122387

25. Kouwenhoven M, Teleshova N, Ozenci V, Press R, Link H. Monocytes in multiple sclerosis: phenotype and cytokine profile. *J Neuroimmunol*. 2001;112(1-2):197-205. http://www.ncbi.nlm.nih.gov/pubmed/11108949.

26. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10(1):57-63. doi:10.1038/nrg2484

27. Khatri P, Sirota M, Butte AJ. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. Ouzounis CA, ed. *PLoS Comput Biol*. 2012;8(2):e1002375. doi:10.1371/journal.pcbi.1002375

28. Glazko G V, Emmert-Streib F. Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets. *Bioinformatics*. 2009;25(18):2348-2354. doi:10.1093/bioinformatics/btp406

29. Bayerlová M, Jung K, Kramer F, Klemm F, Bleckmann A, Beißbarth T. Comparative study on gene set and pathway topology-based enrichment methods. *BMC Bioinformatics*. 2015;16(1):334. doi:10.1186/s12859-015-0751-5

30. Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*. 2009;10(1):161. doi:10.1186/1471-2105-10-161

31. ArrayExpress < EMBL-EBI.

32. Binder MD, Fox AD, Merlo D, et al. Common and Low Frequency Variants in MERTK Are Independently Associated with Multiple Sclerosis Susceptibility with Discordant Association Dependent upon HLA-DRB1*15:01 Status. Gibson G, ed. *PLoS Genet*. 2016;12(3):e1005853. doi:10.1371/journal.pgen.1005853

33. Home - GEO - NCBI. https://www.ncbi.nlm.nih.gov/geo/. Accessed April 9, 2019.

34. Andrews S. (2010). Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed March 6, 2018.

35. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170

36. Ensembl genome browser 91. https://useast.ensembl.org/index.html. Accessed March 6, 2018.

37. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq

aligner. *Bioinformatics*. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635

38. Liao Y, Smyth GK, Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res*. 2013;41(10):e108--e108. doi:10.1093/nar/gkt214

39. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8

40. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B*. 1995;57(1):289-300. doi:10.1111/j.2517-6161.1995.tb02031.x

41. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139-140. doi:10.1093/bioinformatics/btp616

42. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res*. 2012;40(10):4288-4297. doi:10.1093/nar/gks042

43. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol*. 2010;11(2):R14. doi:10.1186/gb-2010-11-2-r14

44. Spies D, Ciaudo C. Dynamics in Transcriptomics: Advancements in RNA-seq Time Course and Downstream Analysis. *Comput Struct Biotechnol J*. 2015;13:469-477. doi:10.1016/j.csbj.2015.08.004

45. Costa V, Angelini C, De Feis I, Ciccodicola A. Uncovering the Complexity of Transcriptomes with RNA-Seq. *J Biomed Biotechnol*. 2010;2010:1-19. doi:10.1155/2010/853916

46. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11(10):R106. doi:10.1186/gb-2010-11-10-r106

47. Wunsch M, Hohmann C, Milles B, et al. The Correlation between the Virus- and Brain Antigen-Specific B Cell Response in the Blood of Patients with Multiple Sclerosis. *Viruses*. 2016;8(4):105. doi:10.3390/v8040105

48. FURTH R, SLUITER W, DISSEL JT. Genetic Control of Macrophage Responses. *Ann N Y Acad Sci*. 1986;465(1 Tenth Interna):15-25. doi:10.1111/j.1749-6632.1986.tb18476.x

49. Karlmark KR, Tacke F, Dunay IR. Monocytes in health and disease - Minireview. *Eur J Microbiol Immunol (Bp)*. 2012;2(2):97-102. doi:10.1556/EuJMI.2.2012.2.1

50. Fisher EM, Beer-Romero P, Brown LG, et al. Homologous ribosomal protein genes on the human X and Y chromosomes: escape from X inactivation and possible implications for Turner syndrome. *Cell*. 1990;63(6):1205-1218. http://www.ncbi.nlm.nih.gov/pubmed/2124517. Accessed April 19, 2019.

51. Hamvas RMJ, Zinn A, Keer JT, et al. Rps4 maps near the inactivation center on the mouse X chromosome. *Genomics*. 1992;12(2):363-367. doi:10.1016/0888-7543(92)90386-7

52. Stenberg AE, Sylvén L, Magnusson CGM, Hultcrantz M. Immunological parameters in girls with Turner syndrome. *J Negat Results Biomed*. 2004;3:6.

doi:10.1186/1477-5751-3-6

53. Gu S, Xie R, Liu X, Shou J, Gu W, Che X. Long Coding RNA XIST Contributes to Neuronal Apoptosis through the Downregulation of AKT Phosphorylation and Is Negatively Regulated by miR-494 in Rat Spinal Cord Injury. *Int J Mol Sci*. 2017;18(4):732. doi:10.3390/ijms18040732

54. Cruz-Orengo L, Daniels BP, Dorsey D, et al. Enhanced sphingosine-1-phosphate receptor 2 expression underlies female CNS autoimmunity susceptibility. *J Clin Invest*. 2014;124(6):2571-2584. doi:10.1172/JCI73408

55. Lahn BT, Page DC. Functional coherence of the human Y chromosome. *Science*. 1997;278(5338):675-680. doi:10.1126/SCIENCE.278.5338.675

56. Rosinski K V, Fujii N, Mito JK, et al. DDX3Y encodes a class I MHC-restricted H-Y antigen that is expressed in leukemic stem cells. *Blood*. 2008;111(9):4817-4826. doi:10.1182/blood-2007-06-096313

57. Scler Sputtek M, Salinas-Riester G, Kroeger N, et al. No proinflammatory signature in CD34+ hematopoietic progenitor cells in multiple sclerosis patients On behalf of: European Committee for Treatment and Research in Multiple Sclerosis Americas Committee for Treatment and Research in Multiple Sclerosis Pan-Asian Committee for Treatment and Research in Multiple Sclerosis Latin American Committee on Treatment and Research of Multiple Sclerosis can be found at: Multiple Sclerosis Journal Additional services and information for. 2012. doi:10.1177/1352458511434067

58. Kooistra SM, Helin K. Molecular mechanisms and potential functions of histone

demethylases. *Nat Rev Mol Cell Biol*. 2012;13(5):297-311. doi:10.1038/nrm3327

59. Huynh JL, Casaccia P. Epigenetic mechanisms in multiple sclerosis: implications for pathogenesis and treatment. *Lancet Neurol*. 2013;12(2):195. doi:10.1016/S1474-4422(12)70309-5

60. Greenfield A, Carrel L, Pennisi D, et al. The UTX gene escapes X inactivation in mice and humans. *Hum Mol Genet*. 1998;7(4):737-742. doi:10.1093/hmg/7.4.737

61. M�ller G, Schempp W. Mapping the human ZFX locus to Xp21.3 by in situ hybridization. *Hum Genet*. 1989;82(1):82-84. doi:10.1007/BF00288279

62. Consortium IMSG, Briggs FBS, Shao X, et al. Genome-wide association study of severity in multiple sclerosis. *Genes Immun*. 2011;12(8):615-625. doi:10.1038/gene.2011.34

63. Brynedal B, Khademi M, Wallström E, Hillert J, Olsson T, Duvefelt K. Gene expression profiling in multiple sclerosis: A disease of the central nervous system, but with relapses triggered in the periphery? *Neurobiol Dis*. 2010;37(3):613-621. doi:10.1016/J.NBD.2009.11.014

64. Sporici R, Issekutz TB. CXCR3 blockade inhibits T-cell migration into the CNS during EAE and prevents development of adoptively transferred, but not actively induced, disease. *Eur J Immunol*. 2010;40(10):2751-2761. doi:10.1002/eji.200939975

65. Høglund RA, Hestvik AL, Holmøy T, Maghazachi AA. Expression and functional activity of chemokine receptors in glatiramer acetate–specific T cells isolated from multiple sclerosis patient receiving the drug glatiramer acetate. *Hum Immunol*.

2011;72(2):124-134. doi:10.1016/j.humimm.2010.10.016

66. Rot A, von Andrian UH. C hemokines in I nnate and A daptive H ost D efense : Basic Chemokinese Grammar for Immune Cells. *Annu Rev Immunol*. 2004;22(1):891-928. doi:10.1146/annurev.immunol.22.012703.104543

67. Acosta-Rodriguez E V, Rivino L, Geginat J, et al. Surface phenotype and antigenic specificity of human interleukin 17–producing T helper memory cells. *Nat Immunol*. 2007;8(6):639-646. doi:10.1038/ni1467

68. Broux B, Pannemans K, Zhang X, et al. CX3CR1 drives cytotoxic CD4+CD28− T cells into the brain of multiple sclerosis patients. *J Autoimmun*. 2012;38(1):10-19. doi:10.1016/j.jaut.2011.11.006

69. Franchi L, Warner N, Viani K, Nuñez G. Function of Nod-like receptors in microbial recognition and host defense. *Immunol Rev*. 2009;227(1):106-128. doi:10.1111/j.1600-065X.2008.00734.x

70. Mahla RS, Reddy MC, Prasad DVR, Kumar H. Sweeten PAMPs: Role of Sugar Complexed PAMPs in Innate Immunity and Vaccine Biology. *Front Immunol*. 2013;4:248. doi:10.3389/fimmu.2013.00248

71. O'Shea JJ, Plenge R. JAK and STAT Signaling Molecules in Immunoregulation and Immune-Mediated Disease. *Immunity*. 2012;36(4):542-550. doi:10.1016/j.immuni.2012.03.014

72. Egwuagu CE, Larkin, III J. Therapeutic targeting of STAT pathways in CNS autoimmune diseases. *JAK-STAT*. 2013;2(1):e24134. doi:10.4161/jkst.24134

73. Zaheer S, Wu Y, Bassett J, Yang B, Zaheer A. Glia Maturation Factor Regulation

of STAT Expression: A Novel Mechanism in Experimental Autoimmune Encephalomyelitis. *Neurochem Res*. 2007;32(12):2123-2131. doi:10.1007/s11064-007-9383-0

74. Chen C, Liu X, Wan B, Zhang JZ. Regulatory Properties of Copolymer I in Th17 Differentiation by Altering STAT3 Phosphorylation. *J Immunol*. 2009;183(1):246-253. doi:10.4049/jimmunol.0900193

75. O'Shea JJ, Kontzias A, Yamaoka K, Tanaka Y, Laurence A. Janus kinase inhibitors in autoimmune diseases. *Ann Rheum Dis*. 2013;72 Suppl 2(0 2):ii111-5. doi:10.1136/annrheumdis-2012-202576

76. Racke MK, Drew PD. Toll-like receptors in multiple sclerosis. *Curr Top Microbiol Immunol*. 2009;336:155-168. doi:10.1007/978-3-642-00549-7_9

77. Levine M, Tjian R. Transcription regulation and animal diversity. *Nature*. 2003;424(6945):147-151. doi:10.1038/nature01763

78. Glass K, Huttenhower C, Quackenbush J, Yuan G-C. Passing Messages between Biological Networks to Refine Predicted Interactions. Semsey S, ed. *PLoS One*. 2013;8(5):e64832. doi:10.1371/journal.pone.0064832

79. Stuart JM. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science (80-)*. 2003;302(5643):249-255. doi:10.1126/science.1087447

80. de la Fuente A. From 'differential expression' to 'differential networking' – identification of dysfunctional regulatory networks in diseases. *Trends Genet*. 2010;26(7):326-333. doi:10.1016/j.tig.2010.05.001

81. Ideker T, Krogan NJ. Differential network biology. *Mol Syst Biol*. 2012;8:565. doi:10.1038/msb.2011.99

82. Hsu C-L, Juan H-F, Huang H-C. Functional Analysis and Characterization of Differential Coexpression Networks. *Sci Rep*. 2015;5(1):13295. doi:10.1038/srep13295

83. Walley AJ, Jacobson P, Falchi M, et al. Differential coexpression analysis of obesity-associated networks in human subcutaneous adipose tissue. *Int J Obes*. 2012;36(1):137-147. doi:10.1038/ijo.2011.22

84. Doig TN, Hume DA, Theocharidis T, Goodlad JR, Gregory CD, Freeman TC. Coexpression analysis of large cancer datasets provides insight into the cellular phenotypes of the tumour microenvironment. *BMC Genomics*. 2013;14:469. doi:10.1186/1471-2164-14-469

85. Wolf DM, Lenburg ME, Yau C, Boudreau A, van 't Veer LJ. Gene co-expression modules as clinically relevant hallmarks of breast cancer diversity. *PLoS One*. 2014;9(2):e88309. doi:10.1371/journal.pone.0088309

86. Linsley PS, Speake C, Whalen E, Chaussabel D. Copy Number Loss of the Interferon Gene Cluster in Melanomas Is Linked to Reduced T Cell Infiltrate and Poor Patient Prognosis. Castro MG, ed. *PLoS One*. 2014;9(10):e109760. doi:10.1371/journal.pone.0109760

87. Han H, Cho J-W, Lee S, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res*. 2018;46(D1):D380--D386. doi:10.1093/nar/gkx1013

88. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47-e47. doi:10.1093/nar/gkv007

89. Shannon P, Markiel A, Ozier O, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res*. 2003;13(11):2498-2504. doi:10.1101/gr.1239303

90. Landeghem S Van, Parys T Van, Dubois M, Inzé D, de Peer Y Van. Diffany: an ontology-driven framework to infer, visualise and analyse differential molecular networks. *BMC Bioinformatics*. 2016;17(1):18. doi:10.1186/s12859-015-0863-y

91. Assenov Y, Ramírez F, Schelhorn S-E, Lengauer T, Albrecht M. Computing topological parameters of biological networks. *Bioinformatics*. 2008;24(2):282-284. doi:10.1093/bioinformatics/btm554

92. Chin C-H, Chen S-H, Wu H-H, Ho C-W, Ko M-T, Lin C-Y. cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst Biol*. 2014;8(Suppl 4):S11. doi:10.1186/1752-0509-8-S4-S11

93. Bader GD, Hogue CW V. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*. 2003;4:2. http://www.ncbi.nlm.nih.gov/pubmed/12525261. Accessed May 9, 2019.

94. Larsen SJ, Baumbach J. CytoMCS: A Multiple Maximum Common Subgraph Detection Tool for Cytoscape. *J Integr Bioinform*. 2017;14(2). doi:10.1515/jib-2017-0014

95. Bindea G, Mlecnik B, Hackl H, et al. ClueGO: a Cytoscape plug-in to decipher

functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*. 2009;25(8):1091-1093. doi:10.1093/bioinformatics/btp101

96. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 2000;28(1):27-30. doi:10.1093/nar/28.1.27

97. Fabregat A, Jupe S, Matthews L, et al. The Reactome Pathway Knowledgebase. *Nucleic Acids Res*. 2018;46(D1):D649-D655. doi:10.1093/nar/gkx1132

98. Wu X, Freeze HH. GLUT14, a Duplicon of GLUT3, Is Specifically Expressed in Testis as Alternative Splice Forms. *Genomics*. 2002;80(6):553-557. doi:10.1006/GENO.2002.7010

99. Amir Shaghaghi M, Zhouyao H, Tu H, et al. The *SLC2A14* gene, encoding the novel glucose/dehydroascorbate transporter GLUT14, is associated with inflammatory bowel disease. *Am J Clin Nutr*. 2017;106(6):1508-1513. doi:10.3945/ajcn.116.147603

100. Shulman JM, Chipendo P, Chibnik LB, et al. Functional Screening of Alzheimer Pathology Genome-wide Association Signals in Drosophila. *Am J Hum Genet*. 2011;88(2):232-238. doi:10.1016/J.AJHG.2011.01.006

101. Amir Shaghaghi M, Murphy B, Eck P. The *SLC2A14* gene: genomic locus, tissue expression, splice variants, and subcellular localization of the protein. *Biochem Cell Biol*. 2016;94(4):331-335. doi:10.1139/bcb-2015-0089

102. Tan NY, Khachigian LM. Sp1 phosphorylation and its regulation of gene transcription. *Mol Cell Biol*. 2009;29(10):2483-2488. doi:10.1128/MCB.01828-08

103. Yurochko AD, Mayo MW, Poma EE, Baldwin AS, Jr, Huang ES. Induction of the

transcription factor Sp1 during human cytomegalovirus infection mediates upregulation of the p65 and p105/p50 NF-kappaB promoters. *J Virol*. 1997;71(6):4638. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC191685/. Accessed May 18, 2019.

104. Vanheusden M, Broux B, Welten SPM, et al. Cytomegalovirus infection exacerbates autoimmune mediated neuroinflammation. *Sci Rep*. 2017;7(1):663. doi:10.1038/s41598-017-00645-3

105. Sundqvist E, Bergström T, Daialhosein H, et al. Cytomegalovirus seropositivity is negatively associated with multiple sclerosis. *Mult Scler J*. 2014;20(2):165-173. doi:10.1177/1352458513494489

106. Liu T, Zhang L, Joo D, Sun S-C. NF-κB signaling in inflammation. *Signal Transduct Target Ther*. 2017;2:17023. doi:10.1038/SIGTRANS.2017.23

107. Baldwin AS. THE NF-κB AND IκB PROTEINS: New Discoveries and Insights. *Annu Rev Immunol*. 1996;14(1):649-681. doi:10.1146/annurev.immunol.14.1.649

108. Chen F, Castranova V, Shi X, Demers LM. New insights into the role of nuclear factor-kappaB, a ubiquitous transcription factor in the initiation of diseases. *Clin Chem*. 1999;45(1):7-17. http://www.ncbi.nlm.nih.gov/pubmed/9895331. Accessed May 18, 2019.

109. Bonetti B, Stegagno C, Cannella B, Rizzuto N, Moretto G, Raine CS. Activation of NF-kappaB and c-jun transcription factors in multiple sclerosis lesions. Implications for oligodendrocyte pathology. *Am J Pathol*. 1999;155(5):1433-1438. http://www.ncbi.nlm.nih.gov/pubmed/10550297. Accessed May 18, 2019.

110. Hiscott J, Kwon H, Génin P. Hostile takeovers: viral appropriation of the NF-kappaB pathway. *J Clin Invest*. 2001;107(2):143-151. doi:10.1172/JCI11918

111. DeMeritt IB, Milford LE, Yurochko AD. Activation of the NF-kappaB pathway in human cytomegalovirus-infected cells is necessary for efficient transactivation of the major immediate-early promoter. *J Virol*. 2004;78(9):4498-4507. doi:10.1128/jvi.78.9.4498-4507.2004

112. Surget S, Khoury MP, Bourdon J-C. Uncovering the role of p53 splice variants in human malignancy: a clinical perspective. *Onco Targets Ther*. 2013;7:57-68. doi:10.2147/OTT.S53876

113. Wosik K, Antel J, Kuhlmann T, Brück W, Massie B, Nalbantoglu J. Oligodendrocyte injury in multiple sclerosis: a role for p53. *J Neurochem*. 2003;85(3):635-644. http://www.ncbi.nlm.nih.gov/pubmed/12694389. Accessed May 20, 2019.

114. Chen D, Pacal M, Wenzel P, Knoepfler PS, Leone G, Bremner R. Division and apoptosis of E2f-deficient retinal progenitors. *Nature*. 2009;462(7275):925-929. doi:10.1038/nature08544

115. Iglesias A, Camelo S, Hwang D, Villanueva R, Stephanopoulos G, Dangond F. Microarray detection of E2F pathway activation and other targets in multiple sclerosis peripheral blood mononuclear cells. *J Neuroimmunol*. 2004;150(1-2):163-177. doi:10.1016/j.jneuroim.2004.01.017

116. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc*

Natl Acad Sci. 2009;106(23):9362-9367. doi:10.1073/pnas.0903103106

117. Hao K, Bossé Y, Nickle DC, et al. Lung eQTLs to Help Reveal the Molecular Underpinnings of Asthma. Williams SM, ed. *PLoS Genet*. 2012;8(11):e1003029. doi:10.1371/journal.pgen.1003029

118. Li Q, Seo J-H, Stranger B, et al. Integrative eQTL-Based Analyses Reveal the Biology of Breast Cancer Risk Loci. *Cell*. 2013;152(3):633-641. doi:10.1016/j.cell.2012.12.034

119. Zou F, Chai HS, Younkin CS, et al. Brain Expression Genome-Wide Association Study (eGWAS) Identifies Human Disease-Associated Variants. Gibson G, ed. *PLoS Genet*. 2012;8(6):e1002707. doi:10.1371/journal.pgen.1002707

120. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. *Nat Rev Genet*. 2015;16(4):197-212. doi:10.1038/nrg3891

121. Grundberg E, Small KS, Hedman ÅK, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet*. 2012;44(10):1084-1089. doi:10.1038/ng.2394

122. Schadt EE, Molony C, Chudin E, et al. Mapping the Genetic Architecture of Gene Expression in Human Liver. Abecassis G, ed. *PLoS Biol*. 2008;6(5):e107. doi:10.1371/journal.pbio.0060107

123. Nica AC, Parts L, Glass D, et al. The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study. Barsh G, ed. *PLoS Genet*. 2011;7(2):e1002003. doi:10.1371/journal.pgen.1002003

124. Fairfax BP, Humburg P, Makino S, et al. Innate Immune Activity Conditions the

Effect of Regulatory Variants upon Monocyte Gene Expression. *Science (80-)*. 2014;343(6175):1246949-1246949. doi:10.1126/science.1246949

125. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. 2008;18(9):1509-1517. doi:10.1101/gr.079558.108

126. Zhang X, Johnson AD, Hendricks AE, et al. Genetic associations with expression for genes implicated in GWAS studies for atherosclerotic cardiovascular disease and blood phenotypes. *Hum Mol Genet*. 2014;23(3):782-795. doi:10.1093/hmg/ddt461

127. Sun W, Hu Y. eQTL Mapping Using RNA-seq Data. *Stat Biosci*. 2013;5(1):198-219. doi:10.1007/s12561-012-9068-3

128. Pickrell JK, Marioni JC, Pai AA, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*. 2010;464(7289):768-772. doi:10.1038/nature08872

129. Battle A, Mostafavi S, Zhu X, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res*. 2014;24(1):14-24. doi:10.1101/gr.155192.113

130. Kumar V, Westra H-J, Karjalainen J, et al. Human Disease-Associated Genetic Variation Impacts Large Intergenic Non-Coding RNA Expression. Cheung VG, ed. *PLoS Genet*. 2013;9(1):e1003201. doi:10.1371/journal.pgen.1003201

131. Edgar R. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30(1):207-210.

doi:10.1093/nar/30.1.207

132. *Calling Variants in RNAseq — GATK-Forum*. https://gatkforums.broadinstitute.org/gatk/discussion/3891/calling-variants-in-rnaseq. Accessed March 8, 2018.

133. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-842. doi:10.1093/bioinformatics/btq033

134. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352

135. Zerbino DR, Achuthan P, Akanni W, et al. Ensembl 2018. *Nucleic Acids Res*. 2018;46(D1):D754-D761. doi:10.1093/nar/gkx1098

136. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122. doi:10.1186/s13059-016-0974-4

137. Larsson J. Area-Proportional Euler and Venn Diagrams with Ellipses [R package eulerr version 6.0.0]. https://cran.r-project.org/web/packages/eulerr/index.html. Accessed September 28, 2019.

138. Arunkumar Srinivasan. gread: Fast Reading and Processing of Common Gene Annotation and Next Generation Sequencing Format Files. 2019.

139. Picard Tools - By Broad Institute. http://broadinstitute.github.io/picard/. Accessed October 3, 2019.

140. Li H. A statistical framework for SNP calling, mutation discovery, association

mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27(21):2987-2993. doi:10.1093/bioinformatics/btr509

141. Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M. VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics*. 2014;30(14):2076-2078. doi:10.1093/bioinformatics/btu168

142. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*. 2012;28(10):1353-1358. doi:10.1093/bioinformatics/bts163

143. A Grammar of Data Manipulation [R package dplyr version 0.8.3]. https://cran.r-project.org/web/packages/dplyr/index.html. Accessed October 3, 2019.

144. Matt Dowle and Arun Srinivasan. data.table: Extension of `data.frame`. 2019. https://cran.r-project.org/package=data.table.

145. Sweeney SM, Cerami E, Baras A, et al. AACR project genie: Powering precision medicine through an international consortium. *Cancer Discov*. 2017;7(8):818-831. doi:10.1158/2159-8290.CD-17-0151

146. Mandic M, Almunia C, Vicel S, et al. The alternative open reading frame of LAGE-1 gives rise to multiple promiscuous HLA-DR-restricted epitopes recognized by T-helper 1-type tumor-reactive CD4+ T cells. *Cancer Res*. 2003;63(19):6506-6515. http://www.ncbi.nlm.nih.gov/pubmed/14559844. Accessed December 15, 2019.

147. Chitnis T. The Role of CD4 T Cells in the Pathogenesis of Multiple Sclerosis. *Int Rev Neurobiol*. 2007;79:43-72. doi:10.1016/S0074-7742(07)79003-7

148. Thomas R, Al-Khadairi G, Roelands J, et al. NY-ESO-1 based immunotherapy of cancer: Current perspectives. *Front Immunol*. 2018;9(MAY). doi:10.3389/fimmu.2018.00947

149. Tobler AR, Constam DB, Schmitt-Gräff A, Malipiero U, Schlapbach R, Fontana A. Cloning of the Human Puromycin-Sensitive Aminopeptidase and Evidence for Expression in Neurons. *J Neurochem*. 2002;68(3):889-897. doi:10.1046/j.1471-4159.1997.68030889.x

150. Vestal DJ. The guanylate-binding proteins (GBPs): Proinflammatory cytokine-induced members of the dynamin superfamily with unique GTPase activity. *J Interf Cytokine Res*. 2005;25(8):435-443. doi:10.1089/jir.2005.25.435

151. Miranda-Hernandez S, Baxter AG. Role of toll-like receptors in multiple sclerosis. *Am J Clin Exp Immunol*. 2013;2(1):75-93. http://www.ncbi.nlm.nih.gov/pubmed/23885326. Accessed December 15, 2019.

152. Reehorst Lereim R, Oveland E. The Brain Proteome of the Ubiquitin Ligase Peli1 Knock-Out Mouse during Experimental Autoimmune Encephalomyelitis. *J Proteomics Bioinform*. 2016;9(9). doi:10.4172/jpb.1000408

153. Yamasaki R, Lu H, Butovsky O, et al. Differential roles of microglia and monocytes in the inflamed central nervous system. *J Exp Med*. 2014;211(8):1533-1549. doi:10.1084/jem.20132477

154. Viñuela A, Brown AA, Buil A, et al. Age-dependent changes in mean and variance of gene expression across tissues in a twin cohort. *Hum Mol Genet*. 2018;27(4):732-741. doi:10.1093/hmg/ddx424

155. Dillman AA, Majounie E, Ding J, et al. Transcriptomic profiling of the human brain reveals that altered synaptic gene expression is associated with chronological aging. *Sci Rep*. 2017;7(1). doi:10.1038/s41598-017-17322-0

156. Gal-Oz ST, Maier B, Yoshida H, et al. ImmGen report: sexual dimorphism in the immune system transcriptome. *Nat Commun*. 2019;10(1). doi:10.1038/s41467-019-12348-6

157. Schurch NJ, Schofield P, Gierliński M, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*. 2016;22(6):839-851. doi:10.1261/rna.053959.115

158. Freeman WM, Walker SJ, Vrana KE. Quantitative RT-PCR: Pitfalls and potential. *Biotechniques*. 1999;26(1):112-125. doi:10.2144/99261rv01

159. Vroon A, Kavelaars A, Limmroth V, et al. G Protein-Coupled Receptor Kinase 2 in Multiple Sclerosis and Experimental Autoimmune Encephalomyelitis. *J Immunol*. 2005;174(7):4400-4406. doi:10.4049/jimmunol.174.7.4400

160. Jothi R, Cuddapah S, Barski A, Cui K, Zhao K. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res*. 2008;36(16):5221-5231. doi:10.1093/nar/gkn488