University of Eastern Piedmont

Department of Health Science

PhD Programme in Medical Sciences and Biotechnology

XXX Cycle

FOLLOW-UP AND FINE-SCALE MAPPING OF MULTIPLE SCLEROSIS LOCI IDENTIFIED IN GENOME WIDE ASSOCIATION STUDIES

PhD Candidate: Miriam Zuccalà

Tutor: Professor Sandra D'Alfonso

Coordinator: Professor Marisa Gariglio

Index

Abstract

Background: Multiple sclerosis (MS) is a chronic inflammatory demyelinating disease of the central nervous system (CNS). It is a multifactorial disease, in fact both environmental and genetic factors contribute to the etiology of the disease. During my PhD program, I focused my research on genetic factors which could contribute to increase the MS susceptibility risk in the Italian population. In the last few years, international studies analyzing large datasets have identified 200 loci involved in the susceptibility of the disease in addition to the known HLA region but very few fine mapping analyses have been conducted to identify the primary causal variant or gene. This discovery was mainly due to the contribution of three international studies in 2011 (IMSGC, Nature, 2011), 2013 (IMSGC, Nature Genetics 2013), and 2017 (IMSGC, 2017) in which our laboratory has took part with 2633 MS patients and 3164 healthy controls (HC). Thanks to this collaboration, we used data derived from the first two studies (IMSGC, Nature, 2011) (IMSGC, Nature Genetics 2013), to perform a genome-wide association analysis, in order to identify genetic markers of susceptibility to MS specific for our population.

Aims: The general aim of our studies was to identify and functionally characterize sequence variations associated to the risk to develop MS in the Italian continental population. To this end we performed two different parallel analyses:

1) a fine mapping analysis of already known associated MS loci in order to find the primary associated variant or gene;

2) a burden test analysis on rare and low frequency variants on already known MS associated loci and genes.

Results: From these analysis, we identified the strongest non-HLA signal in the Italian population maps in the Tumor Necrosis Factor (ligand) superfamily member 14 (*TNFSF14*) gene encoding for LIGHT, a transmembrane glycoprotein expressed on various immune cells and involved in dendritic cells (DC) maturation. We demonstrated through a fine-mapping approach that an intronic variant, rs1077667, is the primarily associated one. Cis-eQTL data from different databases showed that carriers of MS risk allele have a lower *TNFSF14* RNA expression in EBV-transformed lymphoblastoid cell lines (Geuvadis, Bioportal, Gtex) and in peripheral blood mononuclear cells (PBMCs) (Gtex). These data are consistent with the imbalance against the risk allele observed in heterozygous individuals (p<0.0001, RNAseq on 97 lymphoblastoid cells, Geuvadis). Consistently, in PBMC of 84 Italian MS and 80 healthy controls (HC), individuals with MS risk genotype produced lower levels of *TNFSF14* transcript (p=1.1e-4) and MS patients were the minor producers (p=0.031). Analysis on peripheral blood of HC (N=37) with flow cytometry showed that in myeloid DC

(CD11c+) the homozygous individuals for the risk allele had a higher percentage of LIGHT positive cells (p-value=0.04). Subsequently, we focused our attention on *TNFRSF14* gene which encodes for HVEM (LIGHT receptor) and we tried to define the primary associated variant in this region as previously done for *TNFSF14* gene. We found a cluster of variants in high linkage disequilibrium (LD) with the most associated variant (rs3748817) in the international cohort but we were not able to distinguish among them, which is the primary associated variant because of the high LD in this region. So, we tried to purse various approaches in order to prioritize these variants and the most consistent data were observed for eQTL analysis. In fact, we found in different databases an eQTL effect for the associated variant rs3748817 and its proxy rs2258734 (r2=1), a variant that maps in the promoter of *TNFRSF14* gene. In details the risk alleles of both these variants significantly correlated with a decrease expression of *TNFRSF14* in different data sets, especially in brain-cerebellum and in EBV-transformed lymphocytes from Gtex and Geuvadis data sets and in whole blood from Blood eQTL browser. No association was found for the expression of surface HVEM protein and the risk genotype of rs3748817 by flow cytometry analysis in whole blood in different cell types. Parallel to this analysis, we explored the interactome of TNFSF14 and TNFRSF14 pathway, at the purpose to conduct an analysis for rare variants. We selected 31 genes for NGS analysis resulting from those genes which showed an interaction with *TNFSF14* or *TNFRSF14* genes experimentally valuated or predicted from online tools and a SNP with an already reported association in the international MSchip project (IMSGC, 2017). We found a significant burden for three genes: *EIF3E* for rare regulative variants, *RUVBL2* and *CDC37* for rare missense variants. We thus conducted gene-gene interaction analysis for the two genes *TNFSF14* and *TNFRSF14,* modelling multiple loci jointly, searching for non-additive effects beyond the single SNP effects. We adopted a "candidate-interactions" strategy, leveraging information from freely available protein-protein interaction (PPI) and pathways resources. Potentially interacting SNPs were hence prioritized in order to narrow down the search space, extracting candidate interacting pairs from three sources: PPI resources (STRING, Reactome, GPS-PROT, PINA), 3 KEGG pathways, HLA class I and class II genes (22 genes) for a total of 561 interactions among *TNFSF14* and *TNFRSF14* and 370 genes. Epistatic interactions were tested in four cohorts but no significant interaction after Bonferroni correction (p<10-4) was found for pairs of SNPs in common among the 4 data sets. Only for 71 pairs of SNPs has been found a significant interaction after Bonferroni correction (p<10-4) but in only one data-set. These results belong to the interaction of *TNFSF14* with 5 genes: *C3, PLCG2*, *PTPN11* and 2 *HLA* genes. Finally, we calculated the weight genetic risk score (wGRS) on 13 genes belonging to TNFSF14 pathways in two different data sets and we found a significant wGRS for both, confirming that variants in this pathway have a role in MS susceptibility. The area under the ROC curve was however very small

(<0.6), indicating that these variants account only for a small fraction of total MS genetic susceptibility.

For the second aim we tested the effect of rare (MAF < 1%) and at low frequency (MAF 1-5%) variants as burden test comparing the total numbers of alternative and reference alleles in patients and controls in loci already known to be associated to MS. During our Discovery phase, we sequenced by Next Generation Sequencing (NGS) approach, 100 genes in 600 MS patients and 408 healthy controls grouped in pools of 12 individuals. We performed Burden Test analysis using specific statistical algorithms obtaining a list of 17 genes showing a significant difference of the number of variants among MS patients and controls. In order to replicate these results, we sequenced the entire coding region of these genes by NGS in an independent sample consisting of 504 MS patients and 504 healthy controls, analysed in pools of 12 individuals. In the Replication phase we found 3 genes (*MYC, TUBD1* and *EFCAB13*) that showed a significant burden in at least one of the 6 filters for coding variants with at least one of the statistic programs. The meta-analysis between the two studies confirmed a significant burden with one of the statistical test for *MYC* and *TUBD1* genes for filters involving coding variants, *NPEPPS* for the filter involving regulatory variants and for *EFCAB13* for the "disruptive" filter (stop-gain, stop-loss, splicing variants). In particular *EFCAB13* seemed to show the most promising result. It was the only gene resulted to have a statistically significant p-value with 2 statistic test (C-ALPHA and WSS) both in discovery and replication study and in meta-analysis among the two data sets. Among the two studies, it is the gene most enriched in disruptive variants (totaling 8 between discovery and replication). Six of these variants were observed both in discovery and in replication subset, and among them, 4 variants (3 stop-gained and 1 splice acceptor) showed a concordant trend for minor allele count between patients and controls.

Conclusions:

In conclusion, thanks to our analysis, we identified an intronic variant (rs1077667) in the *TNFSF14* gene (encoding for LIGHT protein) as the primary associated one in the Italian population and we were able to define its functional role in the regulation of gene transcription and protein production. In *TNFRSF14* gene region but we were not able to identify the primary associated variant due to high linkage disequilibrium (LD). Despite this, we observed a cis-eQTL effect for different variants in this region on *TNFRSF14* gene expression. So, based on these evidences, we proposed for these variants a possible role in gene regulation (especially for a SNP in the gene promoter, in high LD with the associated variant in the international studies). Although we did not confirm this effect on protein production in a specific cell population, further analysis will be required to confirm our hypothesis and to try to better investigate the regulative role for the most interesting variants inside the region of

TNFRSF14 gene region. Gene-gene interaction analysis, burden test and weight genetic risk score on TNFSF14-TNFRSF14 pathway seemed to confirm our hypothesis that also genes which interact with *TNFSF14*, can also play a role in MS pathogenesis. Further analysis will be required to better investigate the causative variant in these genes and to study in deeper the role of this pathway in MS pathogenesis.

Parallel to these analysis, we conducted a research of rare functional variants in MS associated loci in order to assess if the genes in these regions showed an imbalance of rare variant frequencies (burden) between MS patients and healthy controls. *EFCAB13* was the gene that seemed to show the most promising result especially for disruptive variants (stop-gain, stop-loss and slicing). *EFCAB13* encodes for EF-hand calcium-binding domain-containing protein 13 which is a poorly characterized calcium binding adaptor protein. Further analysis will be required in order to define its functional role and to better investigate its function in MS pathogenesis.

Riassunto

Introduzione: La sclerosi multipla (MS) è una malattia infiammatoria cronica demielinizzante del sistema nervoso centrale. È una malattia multifattoriale, infatti sia fattori genetici che ambientali contribuiscono alla sua eziologia. Durante il mio programma di dottorato in scienze e biotecnologie mediche ho condotto una ricerca incentrata su possibili fattori genetici che potessero essere coinvolti nell'eziologia della malattia. Negli ultimi anni grazie a studi internazionali è stato possibile identificare 200 loci coinvolti nella suscettibilità alla malattia oltre alla nota regione del complesso maggiore di istocompatibilità (MHC). Nonostante questi progressi, pochi studi di mappatura fine ("fine-mapping") sono stati condotti al fine di identificare il gene o la variante causale. Queste scoperte sono state rese possibili grazie al contributo di tre studi internazionali nei quali il nostro laboratorio ha preso parte con un totale di 2633 pazienti e 3164 controlli sani rispettivamente nel 2011 (IMSGC, Nature, 2011), 2013 (IMSGC, Nature Genetics 2013), and 2017 (IMSGC, 2017). Grazie a questa collaborazione, abbiamo utilizzato i dati derivati dai primi due studi (IMSGC, Nature, 2011) (IMSGC, Nature Genetics 2013), per eseguire un'analisi di associazione su tutto il genoma al fine di identificare i marcatori genetici di suscettibilità alla MS specifici per la nostra popolazione.

Scopo del lavoro: L'obiettivo generale del nostro studio è stato quello di identificare e caratterizzare funzionalmente le variazioni di sequenza (polimorfismi a singolo nucleotide, SNP) associate al rischio di sviluppare la sclerosi multipla (SM) nella popolazione continentale italiana. A tal fine abbiamo eseguito due diverse analisi parallele:

1) un'analisi di mappatura fine di loci MS già noti, al fine di identificare la variante o il gene primariamente associati;

2) un'analisi del carico ("burden") delle varianti rare e a bassa frequenza su loci e geni già noti per la loro associazione con la MS.

Risultati: Le nostre analisi ci hanno portato ad identificare il segnale di associazione più forte nella popolazione italiana nel gene "Tumor Necrosis Factor (ligand) superfamily member 14 (*TNFSF14*)" che codifica per LIGHT, una glicoproteina transmembrana espressa su varie cellule del sistema immunitario e coinvolta nel processo di maturazione delle cellule dendritiche (CD). Abbiamo dimostrato tramite un approccio "fine-mapping" che la variante primariamente associate nel gene è la variante intronica rs1077667. Dati di cis-eQTL derivanti da differenti database hanno mostrato come i portatori dell'allele di rischio (C) abbiano un'espressione minore del trascritto del gene *TNFSF14* nelle linee cellulari linfoblastoidi trasformate con EBV (dati da Geuvadis, Bioportal, Gtex)

e nelle cellule mononucleate del sangue periferico (PBMCs) (dati da Gtex). Questi dati sono inoltre in linea con dati di letteratura derivanti dal sequenziamento dell'RNA con metodiche di sequenziamento di nuova generazione (NGS) che riportano negli individui eterozigoti un'espressione differenziale del gene a sfavore dell'allele di rischio (p<0.0001, dati di RNAseq su 97 linee cellulari linfoblastoidi da Geuvadis). In maniera coerente abbiamo dimostrato che nei PBMC di 84 pazienti SM italiani e 80 controlli sani, gli individui con il genotipo di rischio per MS producevano livelli minori del trascritto del gene *TNFSF14* e che i pazienti erano in assoluto i minori produttori. Analisi di citofluorimetria condotte su sangue periferico di 37 controlli sani hanno evidenziato che nella popolazione delle cellule dendritiche mieloidi (CD11c+), gli individui omozigoti per l'allele di rischio presentavano una percentuale maggiore di cellule LIGHT positive.

Successivamente a questi risultati, abbiamo focalizzato la nostra attenzione sul gene *TNFRSF14* che codifica per HVEM (recettore di LIGHT) con l'intento di identificare la variante primariamente associata in questa regione come fatto in precedenza per *TNFSF14*. Le nostre analisi hanno identificato un cluster di varianti in alto linkage disequilibrium (LD) con la variante più associata (rs3748817) nella coorte internazionale ma non siamo stati in grado di discriminare quale tra queste sia la variante primariamente associata a causa dell'elevato LD di questa regione. Abbiamo tentato diversi approcci al fine di prioritizzare questo gruppo di varianti e i risultati più promettenti si sono rivelati essere quelli provenienti dalle analisi di cis-eQTL. Infatti, abbiamo identificato in differenti database un effetto cis-eQTL per la variante più associata rs3748817 e per il suo proxy rs2258734 (r2=1), una variante che mappa nel promotore del gene *TNFRSF14*. In particolare, abbiamo riscontrato che gli alleli di rischio per entrambe le varianti correlavano significativamente con una ridotta espressione del gene *TNFRSF14* in differenti data set, specialmente nel cervelletto e in linee cellulari linfoblastoidi trasformate con EBV (come riportato nei database Gtex e Geuvadis) e nel sangue periferico (da Blood eQTL browser). Successivamente abbiamo eseguito delle analisi in citofluorimetria per la proteina di superficie HVEM in differenti tipi cellulari presenti nel sangue periferico di controlli sani ma non è stata riscontrata un'associazione tra i livelli di espressione della proteina e il genotipo di rischio della variante associata rs3748817. Parallelamente a quanto riportato finora, ci siamo inoltre focalizzati sul pathway di TNFSF14 e TNFRSF14, al fine di condurre un'analisi per le varianti rare. Abbiamo selezionato a tal fine 31 geni che sono stati sequenziati con metodica NGS. La selezione di questi geni è stato il risultato di una ricerca condotta in letteratura e tramite tool bioinformatici di interazioni geniche validate sperimentalmente o predette e aventi una variante associata ad SM come riportato nella casistica internazionale derivante dal progetto MSchip (IMSGC, 2017). Abbiamo trovato un carico significativo (burden) per 3 geni: *EIF3E* per varianti rare regolatorie, *RUVBL2* e *CDC37* per varianti rare missenso.

Successivamente abbiamo condotto un'analisi di interazione genica per i due geni *TNFSF14* e *TNFRSF14* e i loro possibili interattori sfruttando le informazioni disponibili dalla letteratura derivanti da analisi di interazione proteina-proteina (PPI) e da analisi di pathway. Gli SNP in geni che potenzialmente potevano interagire sono stati quindi prioritizzati estrapolando coppie geniche derivanti da tre fonti: PPI (STRING, Reactome, GPS-PROT, PINA), 3 percorsi KEGG, geni MHC di classe I e classe II (22 geni) per un totale di 561 interazioni tra *TNFSF14* e *TNFRSF14* e 370 geni. Le interazioni epistatiche sono state testate in 4 coorti ma non è stata trovata alcuna interazione dopo correzione di Bonferroni (p<10-4) per coppie di SNP in comune tra i 4 data set. Soltanto per 71 coppie di SNP è stata trovata un'interazione significativa dopo correzione di Bonferroni ma solo in un data-set. Questi risultati appartengono all'interazione del gene *TNFSF14* con 5 geni: *C3, PLCG2, PTPN11* e 2 geni MHC. Infine, abbiamo calcolato uno score genetico di rischio (wGRS) su 13 geni appartenenti al pathway di TNFSF14 in due diversi data set e abbiamo trovato un wGRS significativo per entrambi, confermando come anche queste varianti in questo pathway abiano un ruolo nella suscettibilità alla MS. Tuttavia, l'aria sottesa alla curva ROC mostrava un valore piuttosto basso (<0.6), e ciò è indice del fatto che queste varianti contribuiscono solo ad una piccola frazione della suscettibilità genetica totale alla SM.

Per il secondo obiettivo del nostro studio, abbiamo testato l'effetto delle varianti rare e a basa frequenza (frequenza allele minore, MAF < 1%) come "burden test" comparando il numero totale di alleli alternativi e di riferimento in pazienti e controlli in loci già noti per essere associati alla SM. Durante una prima fase preliminare abbiamo sequenziato tramite approccio NGS, 100 geni in 600 pazienti e 408 controlli sani raggruppati in pool di 12 individui. Abbiamo eseguito un'analisi del carico totale delle varianti ("burden test") usando specifici algoritmi statistici, ottenendo in tal modo una lisa di 17 geni che mostravano una differenza significativa nel numero delle varianti tra pazienti e controlli. Al fine di replicare questi risultati, abbiamo sequenziato l'intera regione codificante di questi geni in NGS in una casistica indipendente di 504 pazienti e 504 controlli sani, analizzati in pool di 12 individui. Nella fase di replica abbiamo trovato 3 geni (*MYC, TUBD1* and *EFCAB13*) che mostravano un burden significativo in almeno uno dei sei filtri adottati per le varianti nella porzione codificante del gene, con almeno uno dei programmi statistici. La meta-analisi tra i due studi ha confermato un burden significativo con almeno uno dei test statistici per i geni *MYC* e *TUBD1* per i filtri che selezionano le varianti codificanti, *NPEPPS* per il filtro che seleziona le varianti con funzione regolatoria e per *EFCAB13* per il filtro che seleziona quelle varianti altamente dannose quali varianti che portano alla formazione di un codone di stop e varianti che alterano lo splicing. In particolare, *EFCAB13* sembra mostrare il risultato più promettente. Tra i due studi è il gene più arricchito in queste varianti in totale 8 tra fase preliminare e di replica. Sei di queste varianti erano

condivise nei due data set e tra loro quattro (3 varianti di stop ed una di splicing) mostravano un trend concordante per la conta dell'allele minore tra pazienti e controlli.

Conclusioni: In conclusione, grazie alla nostra analisi, abbiamo identificato una variante intronica (rs1077667) nel gene *TNFSF14* (codificante per la proteina LIGHT) come primariamente associata nella popolazione italiana e siamo stati in grado di definire il suo ruolo funzionale nella regolazione della trascrizione genica e traduzione proteica. Nella regione del gene *TNFRSF14* non siamo stati in grado di identificare la variante primariamente associata a causa dell'alto linkage disequilibrium della regione. Nonostante ciò abbiamo osservato un effetto cis-eQTL per differenti varianti in questa regione sull'espressione del gene. Così, basandoci su queste evidenze, proponiamo per queste varianti un possibile ruolo nella regolazione del gene *TNFRSF14* (specialmente per uno SNP nella regione del promotore, in alto LD con la variante associata negli studi internazionali). Benché non confermiamo questo effetto sulla produzione proteica in una specifica popolazione cellulare, ulteriori analisi saranno eseguite per confermare la nostra ipotesi e provare a investigare nel dettaglio il ruolo regolatorio per le varianti più interessanti nella regione del gene *TNFRSF14*. Analisi di interazione genica, burden test e score di rischio genetico sul pathway di TNFSF14-TNFRSF14 sembrano confermare la nostra ipotesi che anche geni che interagiscono con *TNFSF14* possono anche rivestire un ruolo nella patogenesi della sclerosi multipla. Ulteriori analisi saranno necessarie per identificare la variante causativa in questi geni e studiare nel dettaglio il ruolo di questo pathway nella patogenesi della malattia.

Parallelamente a queste analisi, abbiamo condotto una ricerca di varianti rare funzionali in loci già associati ad SM al fine di determinare se questi geni in queste regioni possano mostrare un carico di varianti rare (burden) differenziale tra pazienti e controlli sani. *EFCAB13* è risultato essere il gene più promettente, in particolare per le varianti altamente dannose che portano alla formazione di codoni di stop e/o di alterazione dello splicing. *EFCAB13* codifica per la proteina 13 contenente un dominio helix-loop-helix e un dominio legante il calcio la cui funzione è stata poco studiata.

Ulteriori analisi saranno necessarie al fine di identificare il ruolo funzionale della proteina sopra citata e in particolare il suo ruolo nella patogenesi della sclerosi multipla.

1. Introduction

1.1 Multiple sclerosis

Multiple sclerosis (MS) is a chronic inflammatory demyelinating disease of the central nervous system (CNS), that leads to changes of nerve conduction due to damage resident cells, primarily oligodendrocytes and neurons (McFarlin and McFarland, 1982a, b). The progressive loss of myelin in different areas of the central nervous system slows the transmission of the nervous signals with the consequent reduction or loss of essential functions. These areas have different size and are called plaques, which can evolve from an initial inflammatory phase to a chronic phase in which they become like scars ("sclerosis") and this leads to a significant physical and cognitive disability.

The plaques are typical inflammatory lesions caused by the attack of the autoimmune system, by the activated autoreactive T lymphocytes (especially CD8+) towards the myelin sheath. T lymphocytes can drive the inflammatory event with production of proinflammatory cytokines (Interferon gamma (INFγ), Tumor necrosis factor alpha (TNFα) and Interleukin 2 (IL-2)) and recall of further mononuclear cells that cross the blood-brain barrier such as B lymphocytes, macrophages, which phagocytize myelin fragments, and polymorphonucleates, which release cytotoxic and cytolytic substances.

The pathogenetic model of MS currently proposed, schematically represented in figure 1, sees the activation of pro-inflammatory T lymphocytes in the periphery. The activation is caused by the recognition of T cells receptor (TCR), antigens presented on the major histocompatibility complex of class II (Major histocompatibility complex, MHC-II), from the antigen presenting cells (Antigen Presentig Cell, APC). These T lymphocytes migrate, adhere and penetrate the blood-brain barrier through molecular adhesion mechanisms and with the intervention of proteases and cytokines.

Within the central nervous system, the T lymphocytes are reactivated by MHC-II on the APC and begin to produce pro-inflammatory cytokines, which promote the inflammatory state in the CNS with consequent activation of effector molecules such as macrophages, B lymphocytes and other T lymphocytes. Macrophages and T lymphocytes attack the myelin sheath through cytotoxic mediators, especially TNF-α, radical species of oxygen (O2) and nitric oxide (NO). B lymphocytes differentiate into plasma cells that secrete demyelinating antibodies. The latter activate other macrophages and the cascade of the complement that cause myelin damage (McFarland and Martin, 2007).
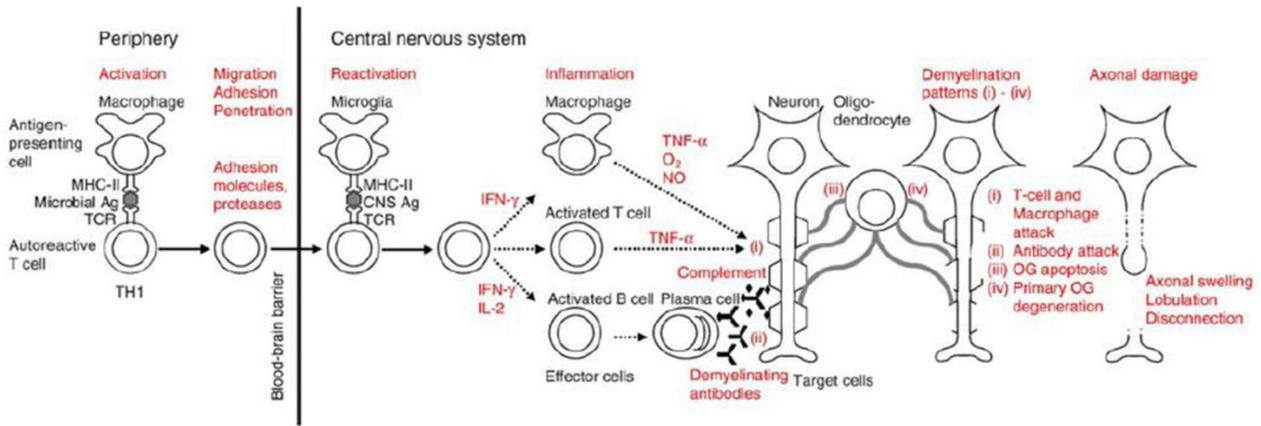
Figure 1: Pathogenetic mechanism of MS

The plaques, generated by this autoimmune process, are called multifocal both in the spatial sense, as they can appear in different areas of the central nervous system, that in a temporal sense because some plaques regress completely, but in general their number increases over time. These regions of inflammation can be detected by neuroimaging techniques, such as magnetic resonance imaging (MRI) (as shown in figure 2). The Revised McDonald Criteria, published in 2017 by the International Panel on the Diagnosis of Multiple Sclerosis, include specific guidelines for using MRI and cerebrospinal fluid analysis to speed the diagnostic process. The MRI can be used to look for a second area of damage in a person who has experienced only one attack (also called a relapse or an exacerbation) of MS-like symptoms — referred to as clinically-isolated syndrome (CIS). The MRI can also be used to confirm that damage has occurred at two different points in time. In some circumstances, the presence of oligoclonal bands in a person's cerebrospinal fluid analysis can be used instead of dissemination in time to confirm the MS diagnosis.
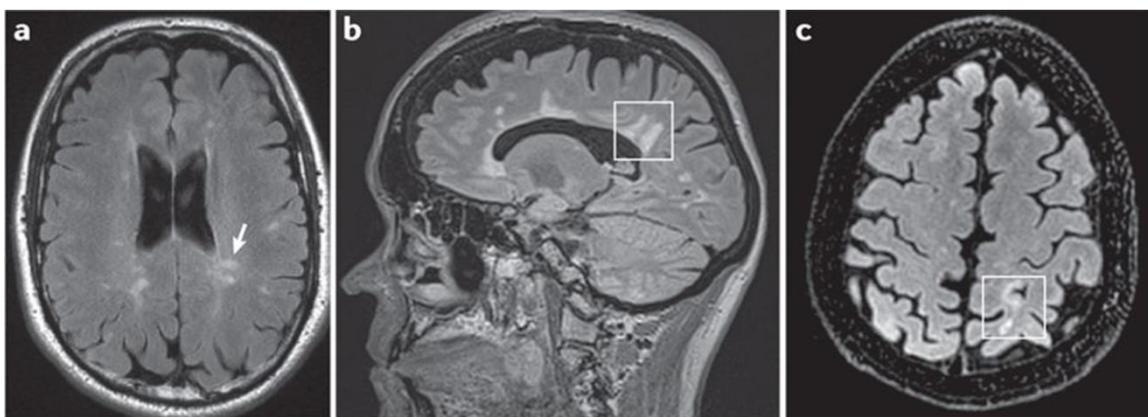
Figure 2: MRI imagine of an individual with MS. The typical plaques are shown with the arrow (a) and inside the square (b-c).

The clinical events of the disease can vary depending on the affected areas (brain and spinal cord) and the symptoms can affect different functions of the organism, regulated by the central nervous system; such as movement and coordination with a general sense of fatigue, sensitivity, sight, balance, speech, sphincter functions and sometimes even cognitive functions.

According to the World Health Organization (WHO) classification, the burden of the disease on the quality of life of the MS patient can be described in terms of;

- "impairment" (set of neurological deficits);

- "handicap" (limitations in social and work activities);

- "disability" (limitations in daily life activities).

The degree of severity of the disease is evaluated through a score from 0 to 10 defined by the clinical scale Expanded Disability Status Scale (EDSS), proposed by the American neurologist Kurtzke in 1983 (Kurtzke, 1983), and shown in figure 3.
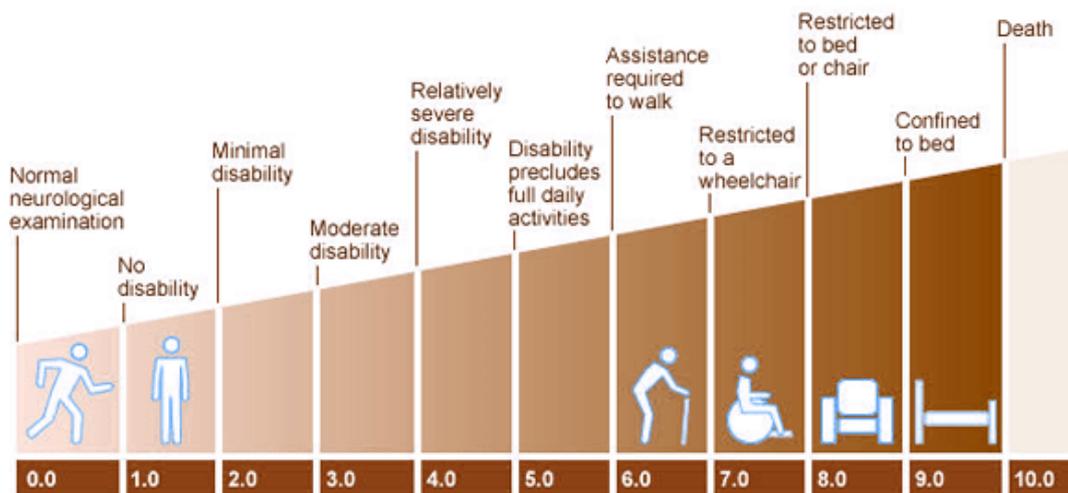
Figure 3: EDSS scale

In 1996, the International Advisory Committee on Clinical Trials of MS originally identified four disease courses. The original four disease types were: Relapsing-Remitting MS, Primary-Progressive MS, Secondary-Progressive MS, Progressive-Relapsing MS. In 2013, the committee changed this classification, including the clinically isolated syndrome (CIS), and eliminating the progressive-relapsing MS (PRMS) (Lublin et al., 2014). In addition, modifiers have been added to promote more effective conversations about disease activity and progression and shared decision-making about treatment options. The modifiers, such as "active" and "not active," incorporate information from MRIs, relapses and degree of disability. CIS is now recognized as the first clinical presentation of a disease that shows characteristics of inflammatory demyelination that could be MS but, has yet to fulfill criteria of dissemination in time. Sometimes, people who have a CIS will not go on to develop

MS. An MRI will better determine the likelihood that someone who has a CIS will develop MS. If lesions on a brain MRI are seen with a CIS, there is a higher chance the patient will develop MS. About the 85% of patients have a relapsing-remitting form (RRMS) which is characterized by clearly defined attacks of new or increasing neurologic symptoms. These attacks – also called relapses or exacerbations – are followed by periods of partial or complete recovery (remissions). During remissions, all symptoms may disappear, or some symptoms may continue and become permanent within the first two decades of the onset. However, about the half of RRMS cases develop a secondary progressive form (SPMS) in which disability gradually increases over time, with or without evidence of disease activity (relapses or changes on MRI). Indeed, the remaining 15% is affected by primary progressive multiple sclerosis (PPMS), associated to a fast progression since the beginning (Kingwell et al., 2015) (Figure 4).

Figure 4: Types of Multiple Sclerosis
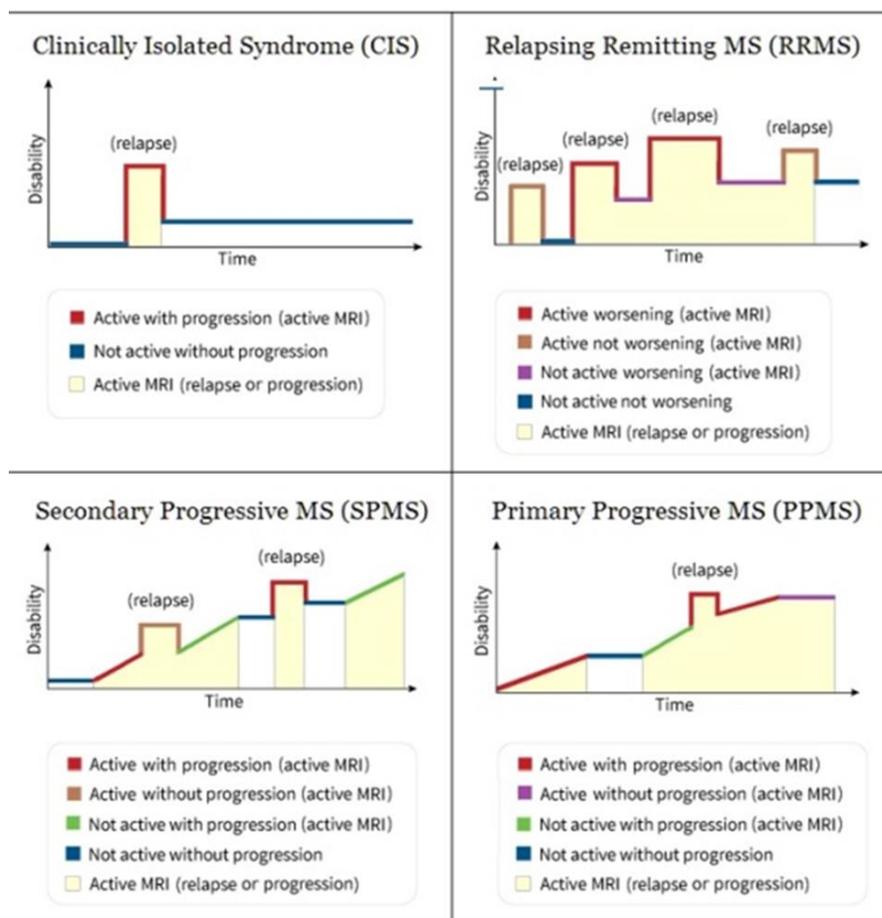
1.2 Epidemiology

MS has a variable onset, between 15 and 50 years, even if it occurs mainly among young adults, between 30 and 40 years, and mainly in the female sex, in a ratio of 1 to 2, compared to men. MS is fairly uncommon in children and teenagers, with prevalence rates approximately in the range of 2.2-5% for those under the age of 16, and MS prevalence rates of pre-pubescent children are as low as 0.1-0.7% (Huppke et al., 2014).

The WHO estimates that there are over 2.5 million affected people in the world. In the United States, the disease affects about 400,000 people and in Italy it is estimated that there are about 50,000 affected individuals. The incidence and prevalence of MS differ depending on the region of the world with most affected patients distant from the equator. The disease is very frequent among the Caucasian populations (especially those living in the north-west of Europe), in North America, in the south-east of Australia and in New Zealand, South-Africa and South America, while there is a low incidence in Asia and in the Caribbean regions. Figure 5 shows the distribution of different MS prevalence in countries around the world.
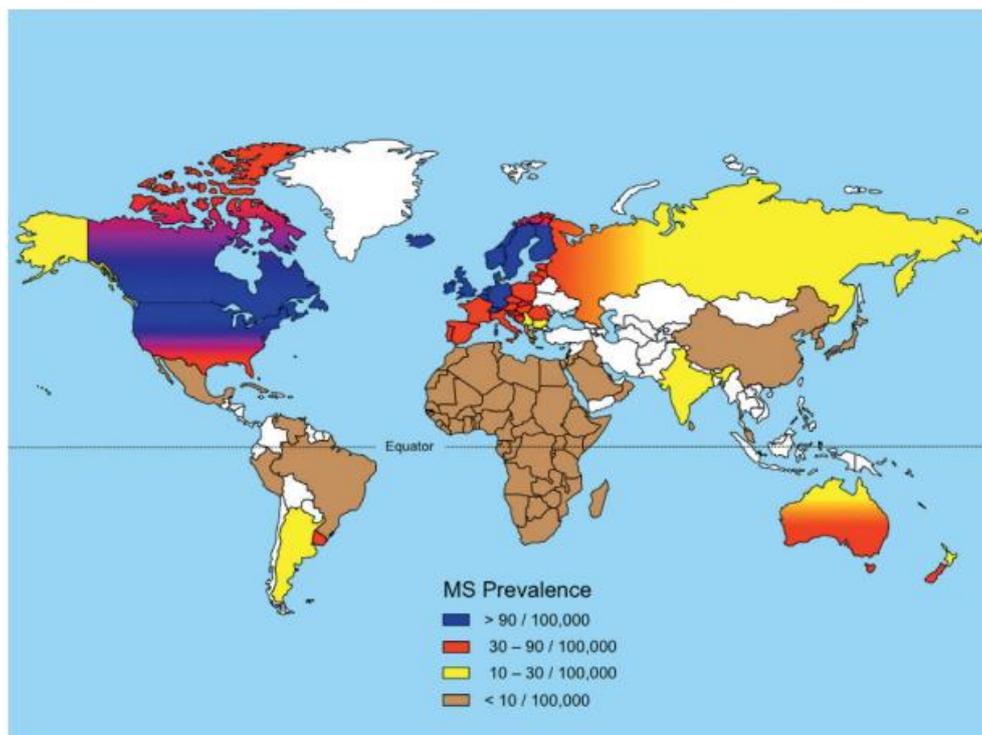

Figure 5: Prevalence of Multiple Sclerosis in the world (per 100,000 individuals).

In Europe, the incidence seems to follow a north-south gradient, with a higher prevalence in the countries of the north, especially in Scandinavia, and low in the countries of the south, except for

Sardinia which shows a prevalence twice higher than the rest of the Italian population, comparable to the prevalence of North European countries. The prevalence of MS in the Italian population shows different rates depending on the regions in particular, in the Central and the South of Italy there are 53 cases per 100,000 inhabitants, while in the North there are 81 cases per 100,000 inhabitants (Totaro et al., 2000). The prevalence in Sardinia is higher than that observed in mainland, to about 150 cases per 100,000 inhabitants (Pugliatti et al., 2001).

1.3 Therapy

More than a dozen disease-modifying medications have been approved by the U.S. Food and Drug Administration (FDA) to treat relapsing forms of MS. These medications reduce the frequency and severity of relapses (also called attacks or exacerbations), reduce the accumulation of lesions in the brain and spinal cord as seen on magnetic resonance imaging (MRI) and may slow the accumulation of disability. Severe exacerbations are most commonly treated with high-dose corticosteroids to reduce the inflammation. Among the injectable drugs, three preparations of IFNβ were for first approved in 1993. Their exact mechanism remains unknown, but the overall effect of this cytokine is an anti-inflammatory, regulatory response. Glatiramer acetate, another injectable that is composed of four amino acids, also produces an anti-inflammatory effect. In clinical trials, these injectable drugs consistently reduced the annual relapse rate (aRR) in patients with RRMS and modestly reduced the time to an increase in disability assessed with the EDSS. Monoclonal antibodies are currently used to treat many autoimmune neurological disorders, including MS and neuromyelitis Optica spectrum disorders (NMOSDs). The first to be used in MS was natalizumab, a humanized monoclonal antibody that binds to α-4 integrin, a component of very late antigen 4 (VLA4), which is present on lymphocytes. Natalizumab prevents the interaction between VLA4 and its endothelial ligand vascular cell adhesion molecule, thereby preventing lymphocytes from crossing the blood–brain barrier. Fingolimod, an analogue of sphingosine 1-phosphate (S1P) that acts as an S1P antagonist, was the first oral drug to be approved for the treatment of RRMS. Fingolimod prevents T cells from leaving the secondary lymph organs because this move depends on the S1P receptor 1 (S1P1); this effect results in a decrease in the number of circulating lymphocytes. Teriflunomide, the second oral drug to be approved by the FDA, was approved in September 2012. Teriflunomide is an active metabolite of leflunomide that inhibits the proliferation of blasting B and T cells. Fumaric acids have been used for decades to treat psoriasis, and in March 2013, the third oral treatment to be approved for the treatment of RRMS was the second-generation fumaric acid, dimethyl fumarate. Preclinical studies demonstrated that dimethyl fumarate has immunomodulatory and antioxidant properties. The

immunomodulatory properties probably relate to the fact that the drug induces a shift in the cytokine profile of T helper (TH) cells from pro-inflammatory (T helper 1 (TH1) cells) to anti-inflammatory (T helper 2 (TH2) cells). In November 2014, alemtuzumab was launched into the market after its FDA approval. Alemtuzumab is a humanized monoclonal antibody against CD52, a receptor that is present on lymphocytes, monocytes and other immune and non-immune cells. Daclizumab is a humanized monoclonal antibody that is administered subcutaneously once per month. This drug was released in May 2016, although safety concerns led to its withdrawal in March 2018. Daclizumab modulates IL-2 signalling by binding to the IL-2 receptor subunit-α (also known as CD25). This binding seems to induce immune tolerance through the expansion of immunoregulatory CD56 bright natural killer cells and the reduction of early T cell activation. Evidence that B cells are involved in the activation of pro-inflammatory T cells, secretion of pro-inflammatory cytokines and the production of autoantibodies led to the development of another generation of monoclonal antibodies that are targeted to these cells. Such antibodies are the CD20-binding antibodies rituximab and ocrelizumab, which deplete mature B cell pools. Based on its immunosuppressant properties, the cancer drug mitoxantrone was also tested in patients with secondary progressive MS and primary progressive MS in a study published in 2002 and was shown to effectively decrease relapse counts and disability progression. The EMA and FDA have recently approved ocrelizumab, which will become the first drug that is licensed for the treatment of primary progressive MS and the oral drug cladribine as therapeutic option for RRMS patients (Tintore et al., 2018). Until now, these drugs do not represent a definite resolution in the treatment of MS. In fact, they play a role in modifying and slowing the course of the disease, reducing the number of attacks of the most common relapsing-remitting form of the disease. MS is a highly disabling pathology that mainly affects young adults and for this reason, it represents a considerable effort in terms of spending on public health. Finding the genetic factors of susceptibility to the disease is crucial to reveal the mechanisms and pathways involved in the onset of CNS damage to understand also the fundamentals for the protection of itself. These results are necessary to highlight new therapeutic targets and develop new and effective therapies.
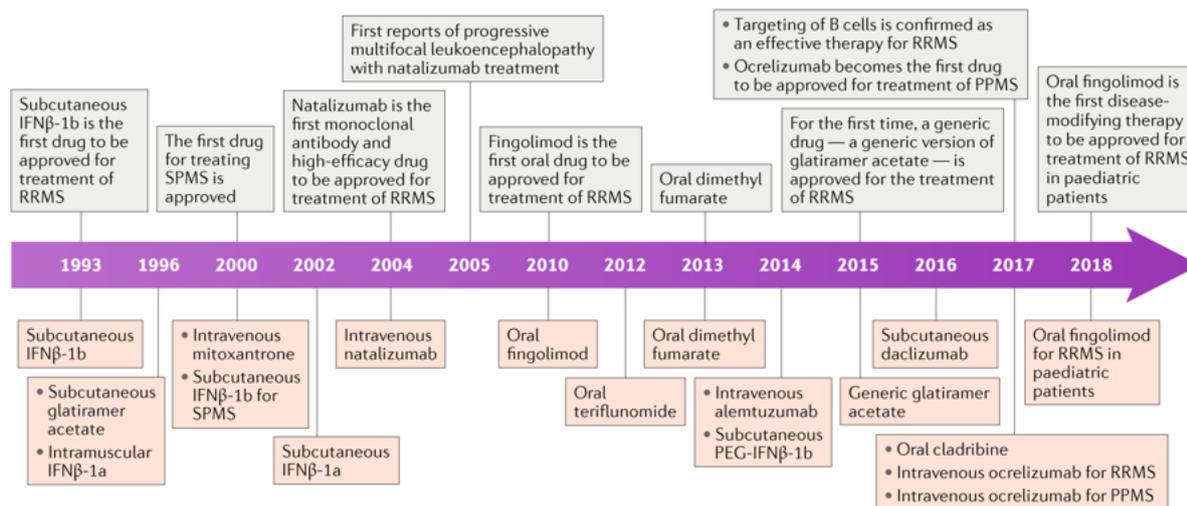
Figure 6: graphical representation of approved drugs for MS across 25 years (image from Treatment of multiple sclerosis — success from bench to bedside, Mar Tintore et al., Nat Rev Neurol, 2018).

1.4 Animal models of MS

Since MS is a complex disease, there is no a single animal model that can capture the entire spectrum of heterogeneity of human MS. Several researchers have recently raised the question whether these animal models could really represent a good model for MS since they do not perfectly reflect all the aspects of the human disease. Disease initiation is usually highly artificial in the animal models (induced by active immunization with an autoantigen). Also, the time-frame of the clinical symptoms onset is different between humans and mice. In humans, physiological processes underlying the disease are undetected for years before the onset of clinical manifestations, while symptoms in the animal models can be detected within weeks or even days after induction of the disease. Moreover, the treatment in these therapy studies started very early during the induced autoimmune disease, whereas any therapy for humans is administered in a late phase of the disease. However, over the last few years, animal models have been used to study the pathogenic mechanisms of MS. The most commonly studied animal models of MS are the experimental autoimmune/allergic encephalomyelitis (EAE); viral induced models, mainly Theiler's murine encephalomyelitis virus (TMEV) infection and consequential chronic demyelination and toxin-induced models of demyelination, such as the cuprizone and the lyso-phosphatidylcholine (lyso-lecithin) models (Procaccini et al., 2015). EAE model, which is the most commonly studied, is induced in susceptible mice through immunization with self-antigens derived from basic myelin protein with Freund's adjuvant and pertussis toxin to potentiate the humoral immune response. The relevant immunogen is derived from self-CNS proteins such as myelin basic protein (MBP), proteolipid protein (PLP) or myelin oligodendrocyte

glycoprotein (MOG). Immunization of SJL/J mice with the immunodominant epitope of PLP ($PLP_{139-151}$) induces a relapsing–remitting (RR) disease course (Tuohy et al., 1989), while disease induced by the immunodominant MOG_{35-55} peptide in C57BL6/J mice is of chronic nature. Viral infections of the CNS can induce demyelination in mice and the best studied are the picornavirus, such as Theiler's murine encephalomyelitis virus (TMEV) and certain strains of the coronavirus, such mouse hepatitis virus (MHV). Unlike EAE, the disease is always chronic-progressive in susceptible mice and TMEV can induce inflammatory demyelinating disease only in mice (Owens, 2006) and not in other different species, such as rodents and primates. On the contrary of EAE model, in TMEV infection, axonal damage precedes demyelination (Tsunoda et al., 2003) and the distribution of damaged axons observed during the early phase corresponds to regions, where subsequent inflammatory demyelination occurs during the chronic phase. This evidence suggests that axonal degeneration triggers recruitment of T cells and macrophages into the CNS, leading to subsequent loss of myelin. While EAE is the most commonly used model to reflect the autoimmune origin of MS, toxic demyelination is more suitable to study the de-and re-myelination processes (Blakemore and Franklin, 2008). Two are the most common agents utilized to induce demyelination: cuprizone and lysolecithin. In conclusion despite all the limitations, animal model of EAE will continue to play a key role as a first-line model system in the development of novel therapeutic approaches for MS, especially for shedding light on specific mechanistic questions.

1.5 Etiopathogenesis of the disease: environmental factors

Although multiple sclerosis (MS) is recognized as a disorder involving the immune system, the interplay of environmental factors and individual genetic susceptibility seems to influence MS onset and clinical expression, as well as therapeutic responsiveness. Multiple human epidemiological and animal model studies have evaluated the effect of different environmental factors, such as viral infections, vitamin intake, sun exposure, or still dietary and life habits on MS prevalence. Previous Epstein-Barr virus infection, especially if this infection occurs in late childhood, and lack of vitamin D (VitD) currently appear to be the most robust environmental factors for the risk of MS, at least from an epidemiological standpoint. Ultraviolet radiation (UVR) activates VitD production but there are also some elements supporting the fact that insufficient UVR exposure during childhood may represent a VitD-independent risk factor of MS development, as well as negative effect on the clinical and radiological course of MS. Recently, there has been a growing interest in the gut-brain axis, a bidirectional neuro-hormonal communication system between the intestinal microbiota and the central nervous system (CNS). Indeed, components of the intestinal microbiota may be pro-

inflammatory, promote the migration of immune cells into the CNS, and thus be a key parameter for the development of autoimmune disorders such as MS. Interestingly most environmental factors seem to play a role during childhood. Thus, if childhood is the most fragile period to develop MS later in life, preventive measures should be applied early in life. For example, adopting a diet enriched in VitD, playing outdoor and avoiding passive smoking would be extremely simple measures of primary prevention for public health strategies. However, these hypotheses need to be confirmed by prospective evaluations, which are obviously difficult to conduct. In addition, it remains to be determined whether and how VitD supplementation in adult life would be useful in alleviating the course of MS once this disease has already started. A better knowledge of the influence of various environmental stimuli on MS risk and course would certainly allow the development of add-on therapies or measures in parallel to the immunotherapies currently used in MS (Pantazou et al., 2015).

1.6 Etiopathogenesis of the disease: genetic factors

1.6.1 The role of common variants in the susceptibility to MS

Multiple sclerosis is a multifactorial disease, in fact both environmental and genetic factors contribute to the etiology of the disease. It is known that the rate of recurrence in families is of 20%, the concordance in monozygotic twins is of 24-30%, while in dizygotic is only 3-5%, which is comparable to that of normal brothers (Hansen et al., 2005; Mumford et al., 1994). The genetic susceptibility is mainly due to human HLA II region (Human Leukocyte Antigen), and especially with *HLA-DRB1*15.01* allele, with an increasing of the risk of three times (Lincoln et al., 2005).

In the early 2000s, the introduction of chip-based technologies with the capacity to genotype simultaneously hundreds of thousands of SNPs allowed the development of a new analytical methodology known as genome-wide association studies or GWAS: a hypothesis-free method in which SNPs spaced across the entire genome are screened for association with a particular trait in case–control datasets composed of genetically unrelated individuals (Manolio, 2010). Compared to classic linkage studies that rely on extended families, the possibility to test unrelated individuals allows collecting much larger datasets, substantially increasing the statistical power of gene-discovery studies. In the last few years, international studies analyzing large datasets at the genome-wide level, have identified 200 loci involved in the susceptibility of the disease in addition to the HLA region. This discovery was mainly due to the contribution of three international studies in 2011 (IMSGC, Nature, 2011), 2013 (IMSGC, Nature Genetics 2013), and 2017 (International et al., 2017). The first MS GWAS was reported in 2007 by the IMSGC employing 931 family trios (one affected

child and both parents). The screening confirmed with genome-wide significance the association of the previously identified locus containing the interleukin-7 receptor α (*IL7Rα*) gene and detected a novel non-*HLA* disease-risk locus, defined by the presence of the interleukin-2 receptor α (*IL2Rα*) gene. In the following years, between 2007 and 2011, seven additional GWA studies of comparable size and one meta-analysis were performed, adding 21 new loci to the roster of MS risk variants. However, theoretical power estimations showed that all the studies conducted at that time were substantially underpowered to capture risk variants with odd ratios less than 1.2, which were the values expected for most of the MS risk variants (Sawcer et al., 2010). For that reason, the IMSGC decided in 2011 to embark on the largest MS GWAS with the collaborative effort of the Welcome Trust Case Control Consortium 2 (WTCCC2). This new study employed nearly 10,000 MS cases and 20,000 healthy controls of European ancestry (collected by 23 research groups working in 15 different countries) and analyzed approximately 450,000 SNPs. This study has confirmed 23 of the 26 known MS associated loci and has identified 29 novel susceptibility loci (p-value $<5 \times 10^{-8}$) and further 5 new regions with strong evidence for association (p-value $<5 \times 10^{-7}$) (IMSGC, Nature, 2011). Gene Ontology analyses have shown that in the 30% of association regions, the nearest gene to the lead SNP is an immune system gene (figure 7). These are genes involved in the lymphocyte function, especially in T-cell activation and proliferation. In details there are genes coding for cytokine pathway (*CXCR5, IL2RA, IL7R, IL7, IL12RB1, IL22RA2, IL12A, IL12B, IRF8, TNFRSF1A, TNFRSF14, TNFSF14*), co-stimulatory (*CD37, CD40, CD58, CD80, CD86, CLECL1*) and signal transduction (*CBLB, GPR65, MALT1, RGS1, STAT3, TAGAP, TYK2*). There are also molecules relates to previously reported environmental risk factors such as vitamin D (*CYP27B1, CYP24A1*), genes involved in therapies for multiple sclerosis including natalizumab (*VCAM1*) and daclizumab (*IL2RA*) and only two genes with a role in axonal neurodegeneration (*GALC, KIF21B*). Each of these genes contributes only minimally to the total risk of development of the disease (odds ratio, OR~1.2) and the most part of the heritability of MS (about 80%) remains unexplained.
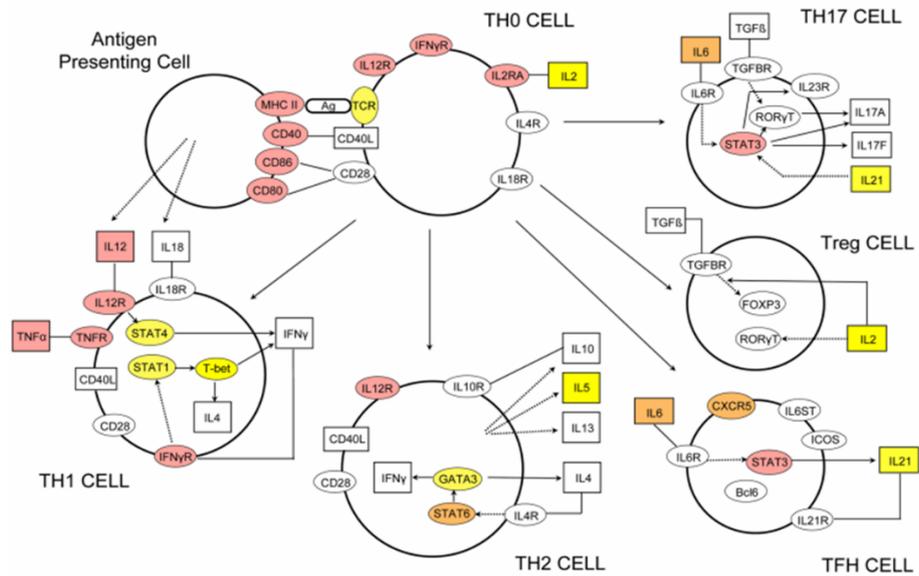
Figure 7: Graphical representation of the T helper cell differentiation pathway (reproduced from IMSCG, Nature, 2011).

Assuming that there are genetic susceptibility factors shared by autoimmune diseases, in 2013, IMSGC have undertaken the Immunochip project (Illumina iSelect custom beadchip platform), drawing a platform array containing 196,524 SNPs in 186 loci emerged in genome-wide association studies previously conducted, associated with at least one of 12 autoimmune diseases (autoimmune thyroid disease, ankylosing spondylitis, Crohn's disease, celiac disease, IgA deficiency, multiple sclerosis, primary biliary cirrhosis, psoriasis, rheumatoid arthritis, systemic lupus erythematosus, type I diabetes and ulcerative colitis), including 55 genes associated with MS (IMSGC, 2013). The analysis was performed on 14,498 MS patients and 24,091 HC belonging to 12 countries. The results were replicated in 14,802 patients and 26,703 controls. They have identified 48 new regions associated with MS and have confirmed 49 already known regions. Overall, in 2013 the results of the Immunochip project have doubled the number of genetic risk factors involved in the disease (leading to 103, in addition to the HLA region, figure 8) and they have confirmed the role of these in the immune response control.

Figure 8: Genetic atlas of multiple sclerosis. The circus plot summarizes all the known MS-associated risk loci. The outer most track indicates the numbered autosomal chromosomes, while the second track shows the closest gene to the top hit within each locus (previously identified associations are in grey). The third track indicates the physical position of the 184 fine-mapped intervals (in green). The inner most track indicates −log(p) for each SNP (scaled from 0 to 12 which truncates the signal in several regions). Also, contour lines are given at the a priori discovery (−log(p) = 4) and genome-wide significance (−log(p) = 7.3) thresholds. Orange indicates −log(p) ≥ 4 and <7.3, while red indicates −log(p) ≥ 7.3. (Reproduced from IMSCG, Nature Genetics, 2013).

In this work also, they found 11 statistically independent effects in the *MHC* region: six *HLA-DRB1* and one *HLA-DPB1* alleles in the centromeric class II region of the locus; one *HLA-A* and two *HLA-B* alleles in the telomeric class I region; and one in the class III region between MHC class I polypeptide-related sequence B (*MICB*) and leukocyte-specific transcript 1 (*LST1*) (Beecham et al., 2013). More recently, the analysis of independent high-density *MHC* region SNP data from multiple cohorts of European ancestry has provided, in addition to novel and previously identified *HLA* class II risk alleles (*DRB1*15:01, DRB1*13:03, DRB1*03:01, DRB1*08:01,* and *DQB1*03:02*) and independent *HLA* class I protective alleles (*A*02:01, B*44:02, B*38:01,* and *B*55:01*), evidence for

two interactions involving pairs of class II alleles: *DQA1*01:01– DRB1*15:01* and *DQB1*03:01– DQB1*03:02* (Moutsianas et al., 2015).

Finally, with the latest GWAS study in 2017 (IMSGC, 2017), the number of statistically independent associations with MS susceptibility has been increased to 233. Of these, 32 were within the major histocompatibility complex (MHC) region, one in chromosome X, and 200 in the autosomal non-MHC genome suggestive effects, including 416 effects that had evidence of statistical replication but did not reach the level of genome-wide statistical significance. The genome-wide and suggestive effects jointly could explain about 48% of the estimated heritability. This study has confirmed the enrichment for MS susceptibility loci in many different immune cell types and tissues on the other hand, they did not find an enrichment in tissue-level CNS profiles. Analyzing data generated from human iPSC-derived neurons as well as from purified primary human astrocytes and microglia, they observed that enrichment for MS genes is seen in human microglia but not in astrocytes or neurons. This study has been the follow up of GWAS and Immunochip IMSGC project. It has been tested in 39,000 individuals from 9 different populations and analyzing 331,536 SNPs. Seen the involvement of our genetics laboratory in this work during my PhD program, this work will be discussed more in detail in the next chapter of this thesis.

Although these efforts in identifying these associated variants and loci and although several signals are near genes involved in immunologic processes, the effector mechanisms for most associations remain unknown. The translation of GWAS data into biological functions has been challenging. The principal reason for this shortcoming consists in the pervasive linkage disequilibrium (LD) along the human genome, which hinders the identification of true causative variants. LD refers to the tendency of genetic loci in physical proximity to segregate together during meiosis, leading DNA to be inherited in large blocks through generations. This peculiarity of genome architecture substantially impairs GWAS resolution since SNPs in the same LD block are inherited together as well. Thus, statistically significant GWAS risk variants are usually proxy for the real causative variants, which can be located up to several megabases away within the same LD block. In addition, the identification of the causative variants is further complicated by the fact that most of them are not translated but rather map to regulatory elements (promoters, enhancers, silencers, and other transcription factor–binding sites) (Baranzini and Oksenberg, 2017). Until now, very few fine mapping analyses have been conducted to identify the primary causal variant or gene. The first putative causal variant identified in MS was the SNP rs6897932 located within the exon 6 of the *IL7R* gene, coding for the trans-membrane segment of the receptor. This SNP was shown to disrupt an exonic splicing silencer, affecting the relative amounts of soluble and membrane-bound isoforms of the protein (Gregory et al., 2007). Recent evidence has shown that the RNA helicase DEAD box polypeptide 39B (DDX39B)

is also a potent activator of *IL7R* exon 6, and the SNP rs2523506 located in the *DDX39B* 5'UTR increases MS risk by reducing *DDX39B* mRNA translation. A similar effect was described for the intronic SNP rs2104286 in the *IL2RA* gene as well. In fact, this risk variant was also found to alter the soluble/membrane-bound ratio of IL2RA protein by driving the expression of higher levels of its soluble form (Galarza-Muñoz et al., 2017). Another well-characterized example is the case of *TNFRSF1A* gene, encoding tumor necrosis factor receptor 1 (TNFR1) a member of the TNF receptor superfamily. It was identified as the causal variant in the gene, a SNP (rs1800693), already discovered through GWAS to be associated with multiple sclerosis, which directs the expression of a soluble form of *TNFR1*. This isoform can bind the TNF, so it mimics the effect of TNF-blocking drugs. It is known that TNF-blocking drugs can promote onset or exacerbation of MS, but they have proven highly efficacious in the treatment of autoimmune diseases for which there is no association with rs1800693 (Gregory et al., 2012). More recently, a study has reported that the nonsynonymous exonic SNP rs11808092 in the ecotropic viral integration site 5 (*EVI5*) gene induces changes in superficial hydrophobicity patterns of the coiled-coil domain of EVI5 protein, which, in turns, affects the EVI5 interactome. In particular, they demonstrated that EVI5 protein bearing the risk allele selectively interacts with sphingosine 1-phosphate lyase (SGPL1), an enzyme important for the creation of the S1P gradient—which is relevant to adaptive immune response and the therapeutic management of MS (Didonna et al., 2015). For the Sardinian population, a GWAS analysis followed by a fine mapping approach led to the identification of a variant in *TNFSF13B* locus (BAFF-var), primarily associated with the regulation of *BAFF* transcription. This variant creates an alternative polyadenylation signal that generates a shorter 3′UTR transcript lacking a miRNA binding site, which leads to increased levels of soluble BAFF, an higher number of B cells and immunoglobulins, reduced levels of monocytes, and an increased risk of autoimmunity (Steri et al., 2017). Fine-mapping analysis will be also matter of this thesis and, two chapters will be dedicated to describing this analysis on two MS associated locus in order to find the primary associated variant or gene.

It is not inconceivable that the potential for the discovery of additive risk variance extractable from large genomic screens will be quickly exhausted. The remaining fraction of the risk commonly known as "missing heritability" is likely due to still unknown common variants characterized by much smaller effects, below the detection limits of the GWA studies conducted so far. Some authors have proposed that a substantial portion of the missing heritability lies in genetic interactions between known variants, the so-called phantom heritability (Zuk et al., 2012). Also, likewise gene by environment interactions, cis/trans-regulators of allelic expression, unidentified rare and penetrant semi-private variants, population and/or disease heterogeneity, neglecting the analysis of sex chromosomes, and hidden epigenetic effects may all contribute to the missing heritability.

1.6.2 The role of rare variants in the susceptibility to MS

Despite the success of GWAS in finding common SNPs associated with disease, common variants explain only a small percentage of the familial aggregation of common complex diseases. So far, the identified MS loci explain only the half of the genetic component involved in the disease; therefore, we still do not have a complete view of the mechanisms at the basis of MS susceptibility. This means that probably many other low-frequency allelic variants and rare variants (MAF<5%), contribute significantly to the etiology of MS. Accordingly, the research is still involved to identify new MS susceptibility marker, including rare variants (O'Gorman et al., 2013).

Further studied have been performed on families with at least two MS cases with two different aims: on one side, to try to discover new rare variants associated with MS susceptibility and on the other side, to understand if some susceptibility loci were transmitted among individuals within the family. only 0,2% of MS families have 4 or more MS patients (Dyment et al., 2008). So far, several Linkage studies on multiplex MS families have been performed. In particular, the study performed by The International Multiple Sclerosis Genetic Consortium, published in 2005, was able to demonstrate a significant linkage in the MHC region, and in addition to this known region it evidenced a suggestive linkage also for chromosome 17q23 and 5q33 although they did not reach statistical significance (IMSGC, 2005). Regarding the rare variants different studies have been performed to find new susceptibility loci. For example, De Jager et al. in a study of 2009 identified rare variants in *TNFRS1A* gene, *CD6* gene, near *IRF 8* gene (De Jager et al., 2009c), associations confirmed also by The International Multiple Sclerosis Genetics Consortium in the study of 2011 (IMSGC, 2011). Another study was able to identify a variant in *TYK2* gene in which the aminoacidic substitution (1104A) made by the polymorphism seems to be involved in a protective role against MS (Ban et al., 2009). An interesting and peculiar case is the study conducted by Ramagopalan et al. that identified a role of rare missense variants in *CYP27B1* in MS susceptibility (Ramagopalan et al., 2011) but when two independent attempts of replication of the data were performed, no statistical evidence were obtained (Barizzone et al., 2013). In addition, Wang Z. in a study of 2016 has identified a mutation in *NR1H3* gene in a MS family associated to the disease susceptibility. In particular, they have proposed that this mutation was responsible in the family of a Mendelian form of MS (Wang et al., 2016). However, the replication of the data conducted by The International Multiple Sclerosis Genetics Consortium on a large dataset failed to confirm this association (IMSGC, 2016). These examples highlight that the discovery of rare variants associated to MS susceptibility gives still controversial results. Regarding the discovery of rare variants associated with autoimmune diseases a study of 2013, conducted by Hunt et al, has highlighted that the most used method for identifying them (genotyping in a large

cohort the variants discovered in the initial small cohort) fails in its intent. In fact, from their data, they have found that rare coding variants have not an important role in the susceptibility of the common autoimmune diseases, but they have supposed that many common variants with a weak signal can be involved in the susceptibility (Hunt et al., 2013). More recently, a study published from IMSGC have tried to study in deep the role of low-frequency and rare variants in the susceptibility to MS. In details this study analyzed 32,367 MS cases and 36,012 controls though an array platform called Exome chip (Illumina technology), which contains almost 200,000 SNPs (rare synonymous and non-synonymous SNPs and common synonymous SNPs) mapping in coding regions enriched in rare variants (MAF<0.01) (IMSGC, 2018). This work has seen the involvement of 12 countries around the world, including 1,530 Italian MS patients and 1,581 controls. They found a significant association for 7 low-frequency variants in 6 genes outside the HLA region. Two of these variants were in genes identified by MS GWAS and showed linkage disequilibrium with the common-variant associations previously reported (Sawcer et al., 2011), while the remaining signals were novel and did not show linkage disequilibrium to common variant association signals in GWASs. The identified genes showed a clear immunological function, particularly in T cells development. This work concludes that nearly 5% of heritability is explained by coding low-frequency variants and that more low-frequency and rare-variant associations remain to be discovered and it will be necessary larger sample sizes to increase statistical power.

A possible strategy to find an association for new rare variants is to study genes already known to carry established MS associated common variants. In fact, Rivas et al, in a study of 2015, studying rare variants associated with inflammatory bowel disease, suggested the possibility that the same genes carrying common disease variants can be affected also by more penetrant rare variants and that the deep sequencing techniques can be useful for understanding the possible role of rare variants in complex disorders (Rivas et al., 2011). This has been the strategy followed by our lab that will be described in a chapter of this thesis.

1.7 LIGHT/HVEM pathway.

Tumor necrosis factor superfamily (TNFSF) molecules play an important role in the activation, proliferation, differentiation, and migration of immune cells into the central nervous system (CNS) and they also have a role in the pathogenesis of neuroinflammation and CNS autoimmunity. While expressions of TNFSF ligands are induced largely on professional antigen-presenting cells (APCs; dendritic cells, B cells, macrophages), their expression is also reported on T cells, NK cells, mast cells, eosinophils, basophils, endothelial cells, thymic epithelial cells, and smooth muscle cells.

Tumor necrosis factor ligand superfamily member 14 (TNFSF14), also known as LIGHT, is a type II transmembrane glycoprotein expressed by activated T lymphocytes, natural killer and immature dendritic cells. *TNFSF14* gene is on chromosome 19p13.3, covers 5.1 KB and includes 4 exons: the first encodes the first 73 amino acids of the polypeptide which constitute the cytoplasmic tail, the transmembrane domain and the start of the extracellular region; the second and third exon coding for the beginning instead of the trimerization domain, while the fourth coding for the remaining trimerization domain (amino acids 101-240) and includes glycosylation site. This protein binds 2 different receptors: HVEM (herpes virus entry mediator) on T lymphocytes and natural killer cells working as a costimulatory molecule inducing proliferation and secretion of IFN-γ, and LTβR (lymphotoxin β receptor) on stromal cells and monocytes, inducing the pro-inflammatory genes expression through activation of NF-kB. The result of the signaling is context specific depending upon the cell type displaying receptor because the binding of LIGHT with the two receptors can determine the cytoplasmatic engagement of TRAF. If it engages TRAF 2/5 the resulting signaling pathway induces the activation of NF-kB and so this results in cell survival and inflammation; but in different context LIGHT- LTβR signaling can induce cell death because of the engagement of TRAF 3 and the activation of caspases. LIGHT also engages decoy receptor-3 (DcR3), a soluble TNFSF receptor lacking transmembrane and signaling domains that probably acts to limit bioavailability of LIGHT. (Granger and Rickert, 2003). There are three physical forms of LIGHT that vary in cellular location: full-length mRNA encodes a typical TNF family transmembrane glycoprotein of 240 aa, an alternative spliced isoform encodes for a non-glycosylated molecule of 204 aa lacking the transmembrane domain that is retained in the cell cytosol and finally a third soluble form derives from LIGHT cleavage by metalloprotease activity (Granger et al., 2001). In literature it is known that TNFSF14 cooperates with CD154 (CD40 ligand) in dendritic cells (DC) maturation, with particular potentiation of allogeneic T cell proliferation and cytokine secretion of IL-12, IL-6, and TNFa (Morel et al., 2001); furthermore, licensed human natural killer cells aid dendritic cell maturation via TNFSF14 (Holmes et al., 2014). In a published work in 2013 on Journal of Immunology was demonstrated that LIGHT-deficient mice developed severe experimental autoimmune encephalomyelitis (EAE) that resulted in an atypically high mortality rate. In the same work they demonstrated that LIGHT expression was crucially involved in controlling activated macrophages/microglia during autoimmune CNS inflammation (Maña P et al, 2013). The receptor of LIGHT, HVEM, also known as TNFR superfamily 14 (TNFRSF14), plays important roles in the immune system such as T-cell costimulation, regulation of DC homeostasis, autoimmune-mediated inflammatory responses, as well as host defense against pathogens. Northern blotting reveals that it is widely expressed in nearly all internal organs with highest expression in lung, kidney, and liver. It

is also expressed in T cells, B cells, dendritic cells (DCs), NK cells, peripheral blood monocytes, neutrophils and all other types of cells within the lymphoid tissue (Cai and Freeman, 2009). In addition to LIGHT, the identified HVEM ligands include CD160, BTLA (B- and T-lymphocyte attenuator), and LTa (lymphotoxin-a) (as shown in figure 9). The binding of LIGHT or LTa to HVEM delivers a costimulatory signal, whereas the binding of BTLA or CD160 to HVEM delivers a coinhibitory signal. Thus, HVEM is a bidirectional switch regulating T-cell activation in a costimulatory or coinhibitory fashion whose outcome depends on the ligand engaged. The cysteine-rich domain 1 (CRD1) of HVEM is essential for the binding of coinhibitory ligands CD160 and BTLA but not costimulatory ligand LIGHT. Deletion or blockade of HVEM CRD1 abolishes the binding of CD160 and BTLA, but not LIGHT, and converts HVEM to a dominant costimulatory molecule, possibly through the loss of negative signaling by CD160∕BTLA (Cai and Freeman, 2009).

HVEM is a receptor that signals through TRAF2 leading to NFkB activation. It may signal also through STAT3 activation influencing the expression of genes involved in host defense in epithelial, Th17 and innate lymphoid cells (Shui et al., 2012). In particular, the NFkB and the STAT 3 signaling pathway may cooperate in differentiation of Th17 cells since STAT3 and, in the NFkB pathway, the c-Rel and RelA/ p65 transcription factors may promote RORt expression in T cells and thus enhance Th17 differentiation (Shui and Kronenberg, 2013). TNFSF14 competes with HSV gD for binding to TNFRSF14, whereas BTLA uses a different binding site. Moreover, TNFSF14 also binds the lymphotoxin ß receptor (LTßR), and Decoy Receptor 3 (Dc3) that binds also FasL. TNFRSF14 has been reported to influence several models of autoimmune diseases such as autoimmune diabetes, autoimmune encephalopathy, concanavalin A-mediated hepatitis, collagen-induced arthritis and several colitis models, with either pro- or anti-inflammatory roles depending on the context (Šedý et al., 2014). BTLA is an inhibitory coreceptor with similarities to CTLA-4 and PD-1 belonging to the CD28 family (Watanabe et al., 2003). It is expressed on a wide range of hematopoietic cells including CD4+T cells, CD8+ T cells, B cells, NKT cells, NK cells, macrophages, and dendritic cells. Moreover, it is highly expressed on follicular T helper cells (Tfh cells). Ligation of BTLA induces its tyrosine phosphorylation and SHP-1/SHP-2 association and then, attenuates IL-2 production and proliferation of T cells. These findings suggest that BTLA functions as an inhibitory coreceptor through the interaction with TNFRSF14. BTLA-deficient mice exhibit enhanced specific antibody responses and sensitivity to experimental autoimmune encephalomyelitis (EAE), rapid rejection of partially MHC-mismatched cardiac allograft, acceleration of experimental colitis and development of an autoimmune hepatitis- (AIH-) like disease and lymphocytic infiltration in multiple organs. Moreover, BTLA plays a protective role in autoimmune diseases in MRL-lpr mice (Oya et al., 2011). The survival of memory and effector T cells has emerged as an important immune function of the

TNFRSF14 network. Btla-/- CD4+ or CD8+ T cells display increased proliferation when activated in vitro consistent with the inhibitory signaling of BTLA (Derré et al., 2010). Inclusion of BTLAFc as a surrogate ligand for TNFRSF14 to cultures of CD4+ and CD8+ T cells substantially enhances proliferation of Btla-/-T cells, but not the rate of division, suggesting that TNFRSF14 impacts cell survival. Rel A nuclear translocation correlates with the survival of Btla-/- T cells treated with BTLA-Fc providing a mechanism linking BTLA-activated TNFRSF14 signaling to cell survival gene expression (Ware and Sedý, 2011). Altogether these data show as the TNFSF14 system is quite complex since it involves several receptor/ligand interactions which may have bidirectional signaling effects on several types of immune cells.
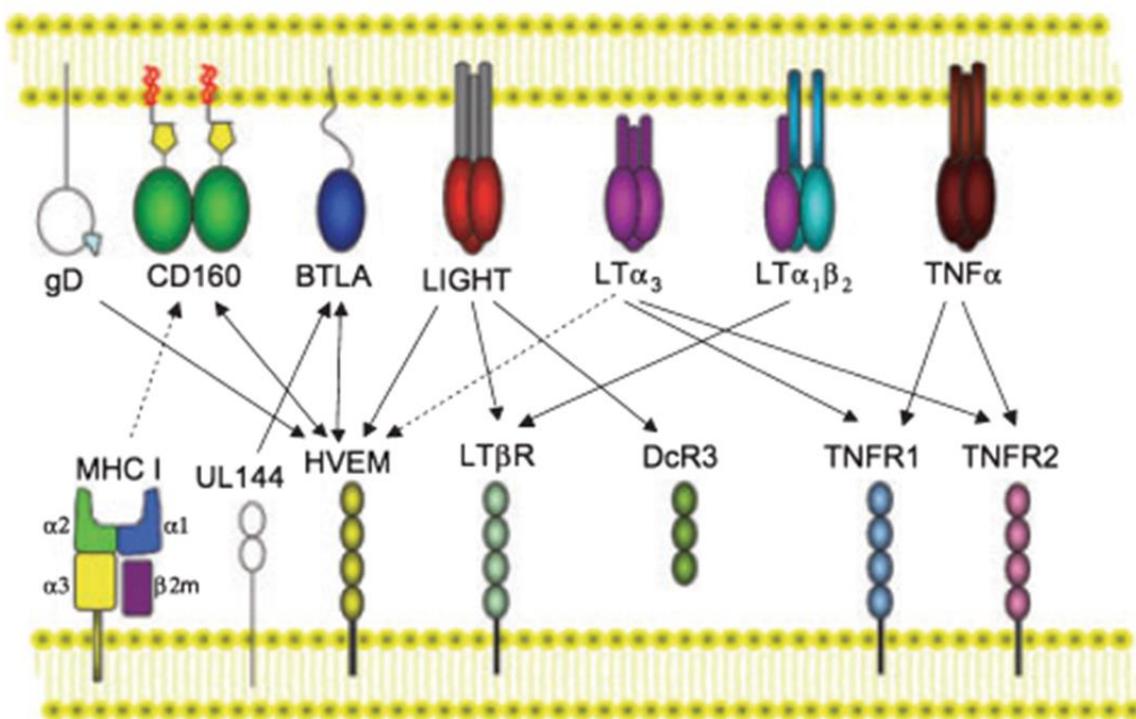
Figure 9: Complex binding pattern of herpes virus entry mediator (HVEM) ligands. CD160, B- and T-lymphocyte attenuator (BTLA), LIGHT (lymphotoxin-like, exhibits inducible expression, and competes with herpes simplex virus glycoprotein D for HVEM, a receptor expressed by T lymphocytes), and lymphotoxin a (LTa) all bind to HVEM. CD160 also binds weakly to classical and non-classical MHC I molecules. In addition, LIGHT binds to LTbR and DcR3. LTa3 and tumour necrosis factor a (TNFa) both bind to TNFR1 and TNFR2, while LTa3 binds to HVEM as well. Furthermore, LTa can couple with LTb and form a LTa1b2 heterotrimer, which binds to LTbR (dashed lines indicate weak binding; arrowheads indicate known signalling directions). (Reproduced from (Cai and Freeman, 2009)

Aim of the study and flowchart of the project

The general aim of our studies was to identify and functionally characterize sequence variations associated to the risk to develop Multiple Sclerosis (MS) in the Italian continental population. To this end we performed two different parallel analyses:

3) a fine mapping analysis of already known associated MS loci identified in Genome Wide Association studies in order to find the primary associated variant or gene;

4) a burden test analysis on rare and low frequency variants on already known associated MS loci and genes.

As shown in the flowchart of the project, we started to perform a Genome Wide Association Study in the Italian population, in order to identify genetic markers of susceptibility to MS specific for our population. The samples (1711 patients and 2234 controls) were genotyped at the genome wide level, thanks to the collaboration to two big International Studies: a Genome-Wide Association-Study (GWAS) and the International Immunochip project, both performed by International Multiple Sclerosis Genetics Consortium and published respectively on 2011 and 2013. From these results, we moved to the selection of MS risk loci for sequencing analysis by NGS approach. This analysis allowed us to perform a fine mapping on two loci resulting to be associated in our sample set: *TNFSF14* and *TNFRSF14,* that will be described in the chapters 2 and 4 of this thesis, and a burden test analysis to study the cumulative effect of rare and low frequency variants on MS associated genes (in chapter 3).

Flowchart of the project
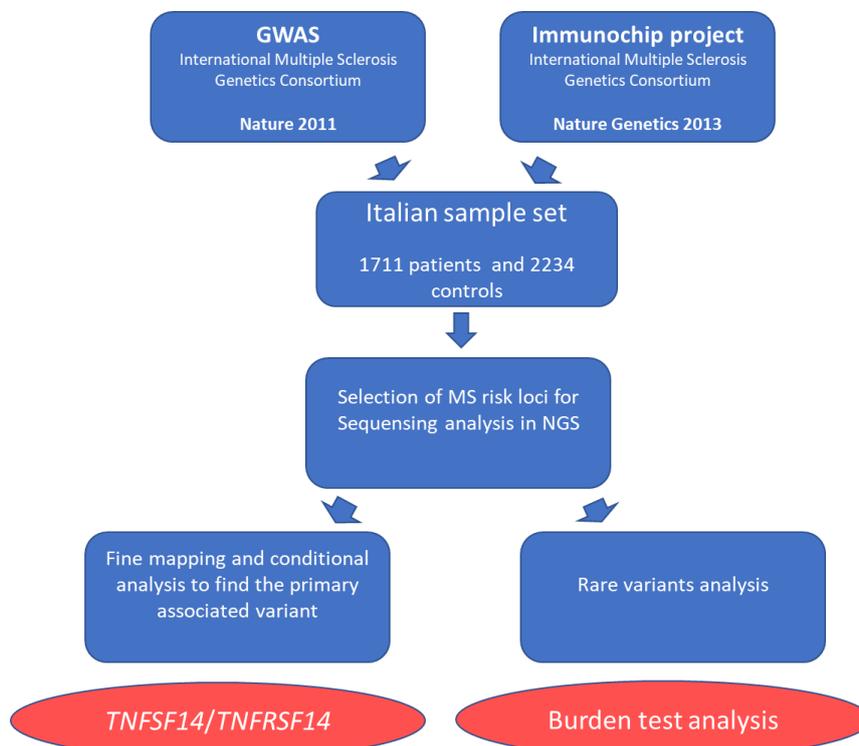
2. Materials and methods

2.1 Samples

A total of 8101 individuals of continental Italian ancestry were recruited across Italian MS centres after approval by the ethics committee of the local hospitals and obtaining written informed consent for genetic analysis for research purposes from all the recruited individuals. MS patients were diagnosed according to McDonald criteria while controls do not have family history of autoimmune diseases and have similar geographic provenience of patients. A peripheral blood sample with EDTA was obtained by each participant to the study. We included in our study 3903 MS patients (2.0:1 female/male ratio) and 4198 healthy controls (HC) (1:1.8 female/male) ratio, 31.67 ($\pm$10) mean age of disease onset, mean EDSS 3.16 ($\pm$2.25), 7% with primary progressive MS (PP).

For 84 patients and 80 controls (for whom fresh biological material was available) we also obtained peripheral mononuclear blood cells (PBMCs). Nucleic acids were extracted from whole blood according to standard protocols (salting out or QIAamp®DNA Blood Mini kit and RNeasy Plus Mini Kit provided by QIAGEN). The PBMCs were isolated by density gradient centrifugation using Lympholyte-H (Cedarline, Burlington, NC, USA) and stored at -80°C with RNAlater (QIAGEN, GmbH, Hilden, Germany) for RNA preservation.

2.2 Sequencing analysis

We sequenced 1104 MS and 912 HC in two different experiments, pooled in groups of 12 individuals each. The libraries were prepared with the "SureSelectXT Target Enrichment System for Illumina Paired-End Multiplexed Sequencing Library" (Agilent Technologies). The DNA quantity has been properly balanced in each pool in order to equally represent each genome.

In the first experiment consisting of 600 MS patients and 408 HC, paired-end multiplexed sequencing was performed on the Illumina GaIIx platform (Illumina, San Diego, CA), combining 6 pools tagged with different index sequences in each lane and producing 2×85 bp read lengths. One of the MS pools (12 patients) did not pass quality controls, so it was discarded from the following analysis. In the second experiment consisting of 504 MS patients and 504 HC, paired end multiplexed sequencing was performed on the Illumina NextSeq 500 (Illumina San Diego) platform, producing 2x150 bp read length.

The two datasets (1092 MS and 912 HC post QC) were analyzed with the same bioinformatic pipeline. The raw-reads were first checked for quality using FastQC (Andrews, 2015). The QC-

checked paired end (PE) reads of each pool were mapped to NCBI human reference genome (build GRCh37) using BWA (v0.7.5) (Li and Durbin, 2009) and the duplicate reads due to PCR amplification during library preparation were removed using *samtools* (Li et al., 2009). A variant caller specifically designed for pooled samples (Bansal, 2010) was used to call the variants. Genomic and functional annotation of the variations was performed with ANNOVAR (Wang et al., 2010). Allelic frequencies (AF) in patients and controls were estimated using an ad-hoc custom pipeline, which was developed to guarantee accurate AF estimation with pooled NGS data (Anand et al., 2016) Specifically, a threshold was applied to single pool alternative AF in order to remove spurious reads. The thresholds (0.26 for the first experiment, 0.24 for the second one) were empirically determined as described in Anand et al. (Anand et al., 2016). In order to remove false positive variants, we chose those variants with sequencing call quality >100.

The 600 MS patients in the first sequencing experiment had been previously individually genotyped either with the Illumina 660Q chip or with the Immunochip platform (Beecham et al., 2013; Sawcer et al., 2011) and AF comparison with these platforms were used to demonstrate a high correlation with AF in the pools (R^2=0.987). Similarly, we also observed a high correlation between pooled AF and frequencies reported in public databases (100 genomes_EUR R^2=0.980, ExAC R^2=0.970).

2.3 Replication and fine-mapping in *TNFSF14*

We genotyped 62 *TNFSF14* variants on an independent, individually typed set of 1745 (867 MS and 878 HC) samples using using a TaqMan® OpenArray™ Genotyping System (Applied Biosystems, Foster City, CA, USA). DNA samples were loaded at a concentration of 50 ng/mL and amplified according to the manufacturer's instructions. The auto-calling method, implemented in the TaqMan Genotyper software version 1.3, was used to assign genotypes. Seven SNPs were removed from the analysis due to failure in the design of the probes, and 13 variants were removed after QC due to bad clustering. All remaining SNPs showed a call rate > 90%. Individuals showing a call rate < 80% were removed during QC, yielding a final dataset of 867 MS and 878 HC. Association effects sizes from this cohort were meta-analyzed with those of 2 other sample sets:

a) 734 MS and 1250 HC (GWAS dataset 1) genotyped with Human610-Quad platform (Sawcer et al., 2011) and imputed with Mach software (Li et al., 2010) on the 1000 genomes dataset (Abecasis et al., 2012) Pre-imputation QC was performed as described elsewhere (Sawcer et al., 2011). After imputation we retained SNPs showing imputation quality index Rsq>0.3 and MAF>0.01.

b) 1236 MS and 370 HC (GWAS dataset 2) genotyped with Illumina HumanOmniExpress-12 BeadChip and HumanOmni-2.5 BeadChip (~ 550k markers in overlap) and imputed on 1000

Genomes phase 3 ALL reference panel (Abecasis et al., 2012) using SHAPEIT (Delaneau et al., 2011) for pre-phasing step and Minimac3 for genotype imputation (Fuchsberger et al., 2015). After imputation we retained SNPs showing Rsq>0.8 and MAF>0.01.

2.4 Rare variants analysis in *TNFSF14*

Rare variants (MAF < 0.01) observed in sequencing were not analyzed for association at single variant level due to lack of power, instead, we investigated the cumulative role of rare variants with predicted functional role in *TNFSF14* performing a burden test analysis. Rare variants were annotated (ANNOVAR) (Wang et al., 2010) and filtered on the basis of *in silico* predicted function. We considered missense and synonymous variants. No splicing or nonsense variation was observed. The burden of rare variants was estimated comparing the total numbers of alternative and reference alleles in patients and controls. The statistical significance was assessed with a Fisher test using R software.

2.5 Burden test analysis

We have calculated the Burden Test in order to evaluate the cumulative effect of rare (Minor Allele Frequency: MAF < 1%), and of low frequency (MAF 1-5%) potential functional variants, derived from the sequencing analysis of the pools. We have found and used 3 different algorithms:

- Weighted-Sum Statistic (WSS): It compares the number of mutations in a group of variants between samples of affected and unaffected unrelated individuals. So, it identifies an excess of alternative alleles in the affected individuals. It computes a genetic score in the region of interest, where variants are weighted by the rare allele frequency, assuming a relationship between allele frequency and effect size.

- C-ALPHA: in a gene harbouring phenotypically relevant variation many variants will be phenotypically neutral. This test is more robust to deviation from assumption of homogeneity of effects of rare variants. It is not strictly a burden test, but a variance test. This test can thus detect a mixture of protective and deleterious variants. Under the null hypothesis of no association between the variants and the phenotype, C-alpha assumes that the distribution of counts (copies of an observed variant) should follow a binomial distribution. For each variant in the region of interest, it tests for unusual departure of the parameter p0 of the binomial distribution, thereby detecting variants in the tail of the distribution (Neale et al, 2011).

- Fisher Hybrid: the performance of each of above tests depends upon the underlying assumption of the relationship between rare variants and the trait. Fisher's hybrid test statistics is proposed to

combine evidence of association from the complementary burden (WSS) and variance (C-alpha) test. Statistical significance of the hybrid statistics is obtained using the same permutation-based method required for the two above tests (Derkach et al, 2013).

In each test the significance value was calculated by applying the permutation of the case control status. As we were using pooled samples the pool and not the single individual was employed as a statistical unit: 10,000 permutations of disease status across pools were used to empirically estimate p-values for all three tests.

2.6 Gene expression analysis

The expression of the two splicing isoforms of *TNFSF14* on two different cohorts of patients and controls (Cohort 1 consisting of frozen PBMC from 64 HC and 45 MS; Cohort 2 consisting of whole blood from 16 HC and 39 MS) was determined by quantitative Real Time PCR with SYBR Green method using the GoTaq 2-step RT-qPCR system (Promega). Its components allow the synthesis of cDNA using GoScript Reverse Transcription System and the subsequent quantification by GoTaq qPCR Master Mix. The qRT-PCR reaction was conducted in the instrument C1000 Thermal Cycler CFX96 Real Time System (Bio-Rad). Each sample was tested in triplicate for the two *TNFSF14* splicing isoforms and for *β-actin* (housekeeping gene). A calibrator RNA was generated from a pool of RNA obtained from pellets of PBMC of two healthy controls and added to all experiments. The analysis of gene expression data was performed with the CFX Manager™ Software Bio-Rad. From our initial experiments we observed that the ΔCT of calibrator RNA was much similar among the different experiments, for this reason we have calculated and compared only ΔCT. qRT-PCR primers: *TNFSF14* full-length isoform forward: GGTGGGTCTGGGTCTCTT; *TNFSF14* full-length isoform reverse: AGACCTTCGCTCTTGTATCAGC; *TNFSF14* ΔTM isoform forward: AGTGTGGCCCGGGACGGA; *TNFSF14* ΔTM isoform reverse: GCTGGAGTTGGCCCCTGTGA; *β-actin* forward: CGCCGCCAGCTCACCATG; *β-actin* reverse: CACGATGGAGGGGAAGACGG.

2.7 eQtl data and allelic imbalance

We looked at eQTL data as available data from 5 public resources: Geuvadis project (Lappalainen et al., 2013) which performed mRNA sequencing on 465 lymphoblastoid cell line samples from 5 populations of the 1000 Genomes Project: the CEPH (CEU), Finns (FIN), British (GBR), Toscani (TSI) and Yoruba (YRI); the Brain eQTL Almanac (Braineac) (Ramasamy et al., 2014) which is a

web-based resource to access the UK Brain Expression Consortium (UKBEC) dataset; the Gtex portal (Melé et al., 2015) which collect RNA sequencing data from 1641 samples across 43 tissues from 175 individuals; the eQTL blood browser (Westra et al., 2013) and the SNPexp (a web tool for calculating and visualizing correlation between HapMap genotypes and gene expression levels in lymphoblastoid cell lines) (Holm et al., 2010).

Finally, we performed allelic imbalance analysis usingWilcoxon paired-samples test on 97 heterozygous EBV cell lines individuals from Geuvadis consortium and from Gtex (version phs000424.v6.p1) in EBV cell lines (33 samples), in PBMC cells (129 samples), in lung (23 samples), liver (37 samples), adipose visceral (56 samples) and esophagus muscularis (20 samples). Results from the various datasets were meta-analyzed with CMA software under a fixed effect model.

2.8 Cell analysis

We performed flow cytometry experiments (BD FACSCAlibur 2 Laser, Marshall Scientific) on eparin blood samples of healthy donors to detect the LIGHT/TNFSF14 transmembrane protein expression (Human Allophycocyanin Mab, Clone 115520, R&D System) and HVEM in different immune cell types: T cells CD8+, T cells CD4+, B cells (CD19+), myeloid dendritic cells (CD11c+), NK cells (CD56dim/CD16bright, CD56-/CD16bright, CD56dim/CD16-) and monocytes (CD14+). We compared the expression of LIGHT in monocyte-derived DC (MDDC) obtained by culturing monocytes for 5 days with GM-CSF+IL-4 or GM-CSF+IFNβ or IL-3 alone or IL-3+IFNβ or GMCSF+IL-15, which are different MDDC types described in the literature (Banchereau and Palucka, 2005). Finally, we stimulated them with lipopolysaccharide (LPS) for 2 days to obtain mature (activated) MDDC. The soluble LIGHT was detected in cell supernatant of dendritic cells by ELISA technique (Human LIGHT/TNFSF14 Quantikine ELISA Kit, R&D System). The production of different cytokines (IL-6, TNFα, IL-10, IL-23) in the supernatants of mature MDC obtained with GM-CSF+IL-15 was evaluated by ELISA technique (Human IL-6 -Kit Elisa- Ready-SET-Go, Prodotti Gianni; Human TNF-alpha DuoSet ELISA, R&D System; Human IL-10 and IL-6 DuoSet ELISA Kit, R&D System).

2.9 Statistical analysis

Genotype association analysis was conducted with PLINK software (Purcell) (Purcell et al., 2007). Conditional analysis was performed fitting a logistic regression model, incorporating sex as covariate and conditioning onone SNP at a time. Meta-analysis of odds ratios was conducted with PLINK software under a fixed-effect model. Power-analysis was conducted with Quanto software. For RNA

expression data we performed two linear regression analyses: 1) we tested the association between expression levels and genotypes assuming an additive model and using sex, cohort, and individual status (case or control) as covariates; 2) we tested the association between expression levels and case-controlstatus using genotype, sex and cohort as covariates. Regarding the cell analysis, in each group, the normal distribution of values of LIGHT expression and cytokines production was verified by Kolmogorov-Smirnov test and differences in mean expression levels were tested by Student's T-test or by Mann-Whitney test as appropriate, with MedCalc Software.

2.10 Gene-gene interaction analysis

Epistatic interactions were tested in four cohorts, with available genotyped subjects both from our laboratory and San Raffaele Hospital:

1) OmniQuad cohort (OQ) (GWAS data set 1), with individuals genotyped on Illumina HumanOmni1-Quad BeadChip (~660k markers);

2) OmniExpress cohort (OE) (GWAS data set 2), with individuals genotyped on Illumina HumanOmniExpress-12 BeadChip and HumanOmni-2.5 BeadChip (~ 550k markers in overlap);

3) ImmunoChip cohort (IC), with individuals genotyped on Immunochip custom array (~190k markers);

4) MSChip cohort (MSC), with individuals genotyped on Mschip custom array (~180k markers).

After QC, the number of subjects included in analyses were: OQ (N_{MS}=736, N_{HC}=1262), OE (N_{MS}=1269, N_{HC}=360), IC (N_{MS}=961, N_{HC}=962), MSC (N_{MS}=921, N_{HC}=934). We tested pairwise interacting SNPs, extracting markers within region of each gene, with a flanking window of ±10kb, to account both for coding SNPs and for variants that can affect transcriptional regulation. Overall, we tested 561 interactions among *TNFSF14* and *TNFRSF14* and 370 genes. Pairwise SNP interaction analyses were conducted using the logistic regression model as implemented in PLINK (Purcell et al., 2007). Logistic regression models were fitted incorporating the two SNPs' additive marginal effects and a multiplicative interaction term, according to additive coding, on which Wald test was performed to detect departure from additivity on the log-odds scale.

3. Chapter 1: The Multiple Sclerosis Genomic Map: role of peripheral immune cells and resident microglia in susceptibility

In this chapter we will present briefly the last work conducted by International Multiple Sclerosis Genetics Consortium (IMSGC) (under second revision on Science), in which our laboratory took part and that has allowed an important advance of the etiopathogenetic knowledges of MS, explaining about the 48 percentage of the hereditability of the disease. Thanks to this study it has been possible to prioritize up to 551 potentially associated MS susceptibility genes, that implicate multiple innate and adaptive pathways distributed across the cellular components of the immune system. Furthermore, using expression profiles from purified human microglia, it was found an enrichment for MS genes in these brain-resident immune cells. This project is the follow-up of the GWAS (IMSGC, 2011) and Immunochip (IMSGC, 2013) IMSGC projects. At the beginning, for the discovery study they organized available and newly genotyped genome-wide data derived from 14,802 subjects with MS and 26,703 controls in 15 data sets. As a result, they identified 1,961 non-MHC autosomal regions that included 4,842 presumably statistically independent SNPs. These were named as "*effects*", assuming that these SNPs tag a true causal genetic effect. Of these, 82 effects were genome-wide significant in the discovery analysis, and another 125 had a p-value $< 1\times10^{-5}$. The following step was to design the MS Chip to directly replicate each of the prioritized effects and this analysis was performed on 20,360 MS subjects and 19,047 controls, organizing in 9 data sets (including our Italian sample set). They found 139 regions with at least one genome-wide effect, and, overall, 200 prioritized effects reached a level of genome-wide significance (GW) in these regions. The odds ratios (ORs) of these genome-wide effects ranged from 1.06 to 2.06. The analysis confirmed prior MHC susceptibility variants and extended the association map to uncover a total of 31 statistically independent effects at the genome-wide level. Regarding of sex chromosome variants, has been identified a SNP as genome-wide significant, within an enhancer peak specific for T cells and downstream of the RNA U6 small nuclear 320 pseudogene (*RNU6-320P*), a component of the U6 small nuclear ribonucleoprotein (snRNP) that is part of the spliceosome and is responsible for the splicing of introns from pre-mRNA. No variant was found in the Y chromosome with a p-value lower than 0.05 in either the discovery or replication sets. Although chromosome X associations cannot be the sole explanation for the preponderance of women among MS patients, the discovery of an MS locus on the X chromosome is a first step towards understanding the genetic contributions of this strong sex bias.

Significant enrichment for MS susceptibility loci was found in many different immune cell types and tissues, whereas there was an absence of enrichment in tissue-level central nervous system (CNS) profiles. An important finding is that the enrichment was observed not only in immune cells that have long been studied in MS, e.g. T cells, but also in B cells whose role has emerged more recently (Bourdette and Yadav, 2008). Furthermore, they demonstrated that many elements of innate immunity, such as natural killer (NK) cells and dendritic cells also displayed strong enrichment for MS susceptibility genes. Interestingly, at the tissue level, the role of the thymus is also highlighted, possibly suggesting the role of genetic variation in thymic selection of autoreactive T cells in MS. Furthermore, they extended the annotation analyses by analysing new data generated from human iPSC-derived neurons as well as from purified primary human astrocytes and microglia. This analysis showed an enrichment for MS genes in human microglia but not in astrocytes or neurons, suggesting that the resident immune cells of the brain may also play a role in MS susceptibility. Cis e-QTL analysis conducted in naive CD4+ T cells and monocytes from 415 healthy subjects as well as peripheral blood mononuclear cells (PBMCs) from 225 remitting relapsing MS subjects, showed that 36 out of the 200 GW MS effects (18%) had at least one tagging SNP (r2>=0.5) that altered the expression of 46 genes (FDR<5%) in CD4+ naïve T cells and 36 MS effects (18%; 10 common with the CD4+ naïve T cells) influenced the expression of 48 genes in monocytes. Since MS is a disease of the CNS, they also investigated a large collection of dorsolateral prefrontal cortex RNA sequencing profiles from two longitudinal cohort studies of aging (n=455), which recruit cognitively non-impaired individuals. This cortical sample provides a tissue-level profile derived from a complex mixture of neurons, astrocytes, and other parenchymal cells such as microglia and occasional peripheral immune cells. They found that 66 of the GW MS effects (33% of the 200 effects) were *cis*-eQTLs for 104 genes. They hypothesized that the effect of a SNP with a cell type-specific cis-eQTL would be stronger if they adjusted for the proportion of the target cell type, so they adjusted each cis-eQTL analysis for the proportion of neurons, astrocytes, microglia, and oligodendrocytes estimated to be present in the tissue. From this analysis was found that the SNP in *SLC12A5* locus was significantly stronger when they accounted for the proportion of neurons and the *CLECL1* locus emerged when they accounted for the proportion of microglia. *SLC12A5* is a potassium/chloride transporter that is known to be expressed in neurons, and a rare variant in *SLC12A5* causes a form of pediatric epilepsy (Puskarjov et al., 2014; Stodberg et al., 2015). On the other hand, *CLECL1* represents a simpler case of a known susceptibility effect that has previously been linked to altered *CLECL1* RNA expression in monocytes (Raj et al., 2014; Wallace et al., 2012); its enrichment in microglial cells, which share many molecular pathways with other myeloid cells, is more straightforward to understand. In conclusion this work represents a milestone in the investigation of

MS and share a roadmap for future work: the establishment of a map with which to guide the development of the next generation of studies with high-dimensional molecular data to explore both the initial steps of immune dysregulation across both the adaptive and innate arms of the immune system and secondly the translation of this auto-immune process to the CNS where it triggers a neurodegenerative cascade. Beyond the characterization of the molecular events that trigger MS, this map will also inform the development of primary prevention strategies since we can leverage this information to identify the subset of individuals who are at greatest risk of developing MS. While insufficient by itself, an MS Genetic Risk Score has a role to play in guiding the management of the population of individuals "at risk" of MS (such as family members) when deployed in combination with other measures of risk and biomarkers that capture intermediate phenotypes along the trajectory from health to disease (De Jager et al., 2009b).

4. Chapter 2: Genomic and functional evaluation of the role of *TNFSF14* gene in the susceptibility to multiple sclerosis

4.1 Introduction

Thanks to GWAS, 32 human leukocyte antigen (HLA) and 200 non-HLA genetic risk factors have been identified in MS susceptibility ((ANZgene), 2009); (Patsopoulos et al., 2011); (Sawcer et al., 2011); (Beecham et al., 2013); (Moutsianas et al., 2015); (IMSGC, 2017). Although several signals are near genes involved in immunologic processes, the effector mechanisms for most associations remain unknown. Until now, very few fine-mapping analyses have been conducted on MS susceptibility loci in order to identify the primary causal variant or gene. The efforts necessary for these complex experiments are particularly relevant to identify the molecular mechanism underlying the association signal. As a recent example, for the Sardinian population, a GWAS analysis followed by a fine mapping approach led to the identification of a variant in *TNFSF13B* locus (BAFF-var), primarily associated with the regulation of BAFF transcription. This variant creates an alternative polyadenylation signal that generates a shorter 3′UTR transcript lacking a miRNA binding site, which leads to increased levels of soluble BAFF, an higher number of B cells and immunoglobulins, reduced levels of monocytes, and an increased risk of autoimmunity (Steri et al., 2017). In another member of the TNF receptor superfamily (*TNFRSF1A*, encoding for tumor necrosis factor receptor 1 (TNFR1)), a SNP influencing the expression of a soluble form of TNFR1 that can block TNF was identified as the causal variant (Gregory et al., 2012).

In the present study, taking advantage of a large cohort of individuals genotyped at genome-wide level, we performed a GWAS in the Italian population and identified the known MS locus in the *TNFSF14* region as the most associated signal in the Italian cohort. Our aim was to perform a fine mapping of this locus and to functionally characterize the primarily associated variant.

4.2 Results

In the framework of a fine mapping study of MS loci in the Italian population, we selected the region showing the strongest association among all the known non HLA MS loci (Sawcer et al., 2011) (Beecham et al., 2013) in a large Italian sample set (1711 patients and 2234 controls) representing a meta-analysis of two sample sets genotyped at genome-wide level using the Human610-Quad (750 MS,1272 HC) or the Immunochip (961 MS, 962 HC) platforms. Indeed, the marker showing the

strongest association in this meta-analysis (p=1.73x10-8) among all the known non HLA MS loci was rs1077667 mapping in the intron 1 of *TNFSF14* (figure 10).
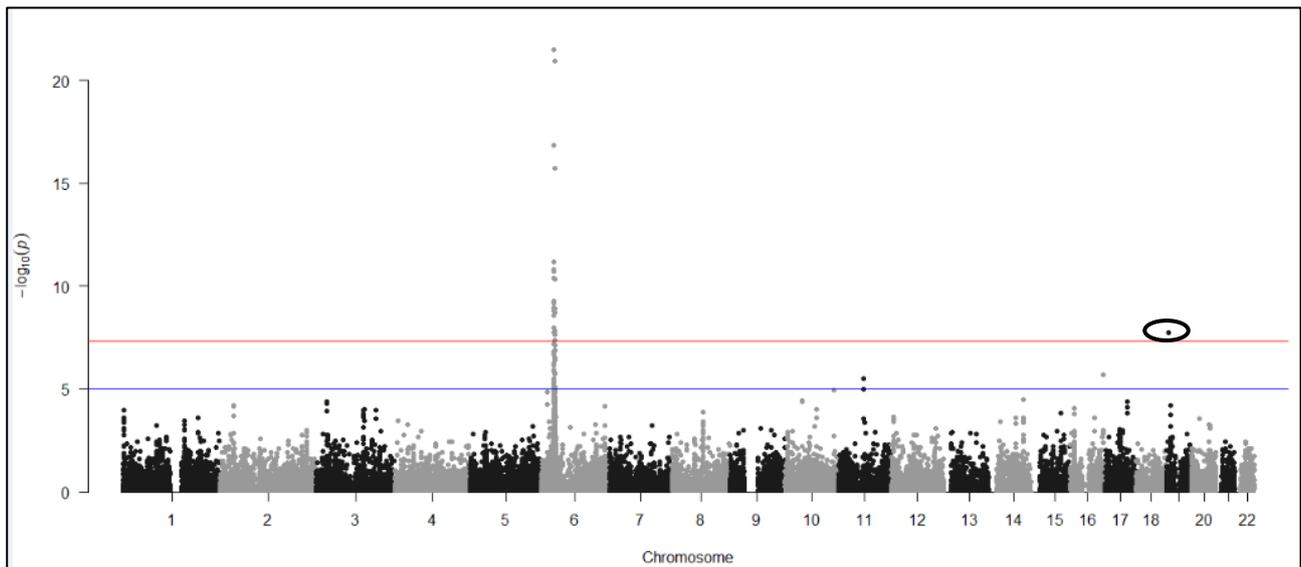

Figure 10: Manhattan plot representing the association results of meta-analysis of GWAS dataset 1 and Immunochip dataset on our Italian cohort. The signal corresponding to the intronic variant (rs1077667) in *TNFSF14* gene is indicated with a circle. The two horizontal lines indicate the thresholds of statistical significance corresponding to genome wide significant association (p=5*e-8, upper line) and suggestive association (p=1*e-5, lower line), respectively.

4.2.1 NGS sequencing of the *TNFSF14* region in the Italian population

The analyses were performed in cooperation with the laboratory of Human Genetics of Neurological Disorders at San Raffaele Hospital headed by professor Filippo Martinelli Boneschi and Federica Esposito and by Institute of biomedical technologies at CNR institute headed by professor Gianluca De Bellis. To identify all the variants in the *TNFSF14* region present in the Italian population and in MS patients, we sequenced the whole genomic region encompassing the *TNFSF14* gene (17,500 bp including exons, introns and 5,000 bp flanking regions) on a sample set of 588 MS patients and 408 controls, pooled in groups of 12 individuals. As we previously published (Anand et al., 2016), we validated an approach which allows to provide an accurate estimate of the allele frequencies in the pool, and thus to perform a preliminary association analysis in the pool of the variants identified in the sequencing experiment. After QC, we identified 112 variants in the *TNFSF14* locus. Among these, 6 variants were in the coding region, 38 with a MAF>1%; only 11 variants (9 with MAF >1%) were already present in the genotyping platforms (Beecham et al., 2013; IMSGC, 2018; IMSGC 2017; Sawcer et al., 2011); 43 were not present in public databases, only 2 of them had a MAF>1% in our population. The comparison of the allele frequencies estimated in the pools of MS patients and

controls showed a statistically significant (p<0.05) association with MS for 15 variants (table S1, appendix), 13 of them present on databases. Only 2 of the variants showing statistically significant association were already present in the genotyping platforms used for the discovery phase. The *TNFSF14* intronic variant (rs1077667) (figure 11 A) showed the strongest signal (p=1.47e-5). We also identified a synonymous variant (rs2291668) in linkage disequilibrium with rs1077667 (r^2=0.808), not present in the genotyping platforms. This SNP is located in exon 1, near the site involved in the alternative splicing which leads to the ΔTM transcript isoform encoding a *TNFSF14* protein lacking the transmembrane domain (short isoform, see expression analysis) (Granger et al., 2001).

4.2.2 Association study of the SNPs selected from the *TNFSF14* sequencing in an independent cohort

To perform a fine-mapping of the *TNFSF14* locus on an independent sample set, we individually genotyped 867 MS and 878 HC individuals (after QC) for 62 variations (2 in/del and 60 SNVs), including: 1) all variants significantly associated (p < 0.05) in the above NGS cohort; 2) all coding variants; 3) LD pruned ($r^2 > 0.9$) common variants (AF>1%); 4) LD pruned ($r^2 > 0.9$) variants reported in the 1000 genome database and not covered in NGS experiment. After QC, we analyzed 42 variants and observed a significant (p < 0.05) MS association for 6 variants (figure 11 B), confirming 5 of the associations observed in the pools. The intronic rs1077667 variant showed the strongest association (p=3.2e-5). The remaining 5 variants showed different LD values (r^2 range: 0.76-0.16) with the intronic rs1077667 variant and did not maintain a statistically significant association after conditioning for rs1077667 (table S2, appendix).

To increase the statistical power of the fine mapping, we enlarged the above dataset with two additional sample sets including 734 MS and 1250 HC (GWAS dataset 1) and 1236 MS and 370 HC (GWAS dataset 2 provided by prof. Filippo Martinelli Boneschi and Federica Esposito group of San Raffaele Hospital), respectively, both imputed on the 1000 genomes dataset for a total of 2837 MS patients and 2498 HC. After meta-analysis of these three sample sets (22 SNPs in common) we observed a significant association for 14 SNPs (p<0.05), confirming the 6 associations observed in the previous analysis. The intronic rs1077667 was still the highest associated signal also in the meta-analyzed dataset (p= 1.363e-010), followed by the exonic variant rs2291668 (p= 6.199e-07) (figure 11 C). After conditioning for rs1077667, the association of this exonic variant was no longer statistically significant (p=0.1623), while 3 other SNPs, with different LD values with rs1077667, still showed a nominally statistically significant association, namely rs142044586 (p=0.04637,

r^2=0.16), rs1862509 (p=0.03893, r^2=0.76), rs344558 (p=0.0023, r^2=0.028) (figure 11 D). After conditioning for each of these 4 variants, rs1077667 maintained the strongest statistical significance (p-value = 6.202e-05, 1.23e-06, 3.19e-06, 4.4e-09 respectively) (table S2, appendix).

Altogether, these analyses indicate that rs1077667 showed the strongest association signal independently of any other tested common (AF>1%) variant in the *TNFSF14* region including those with a nominal independent p value of association. Accordingly, we considered rs1077667 as the primarily associated variant with MS in the *TNFSF14* region and designed functional assays on this variant.
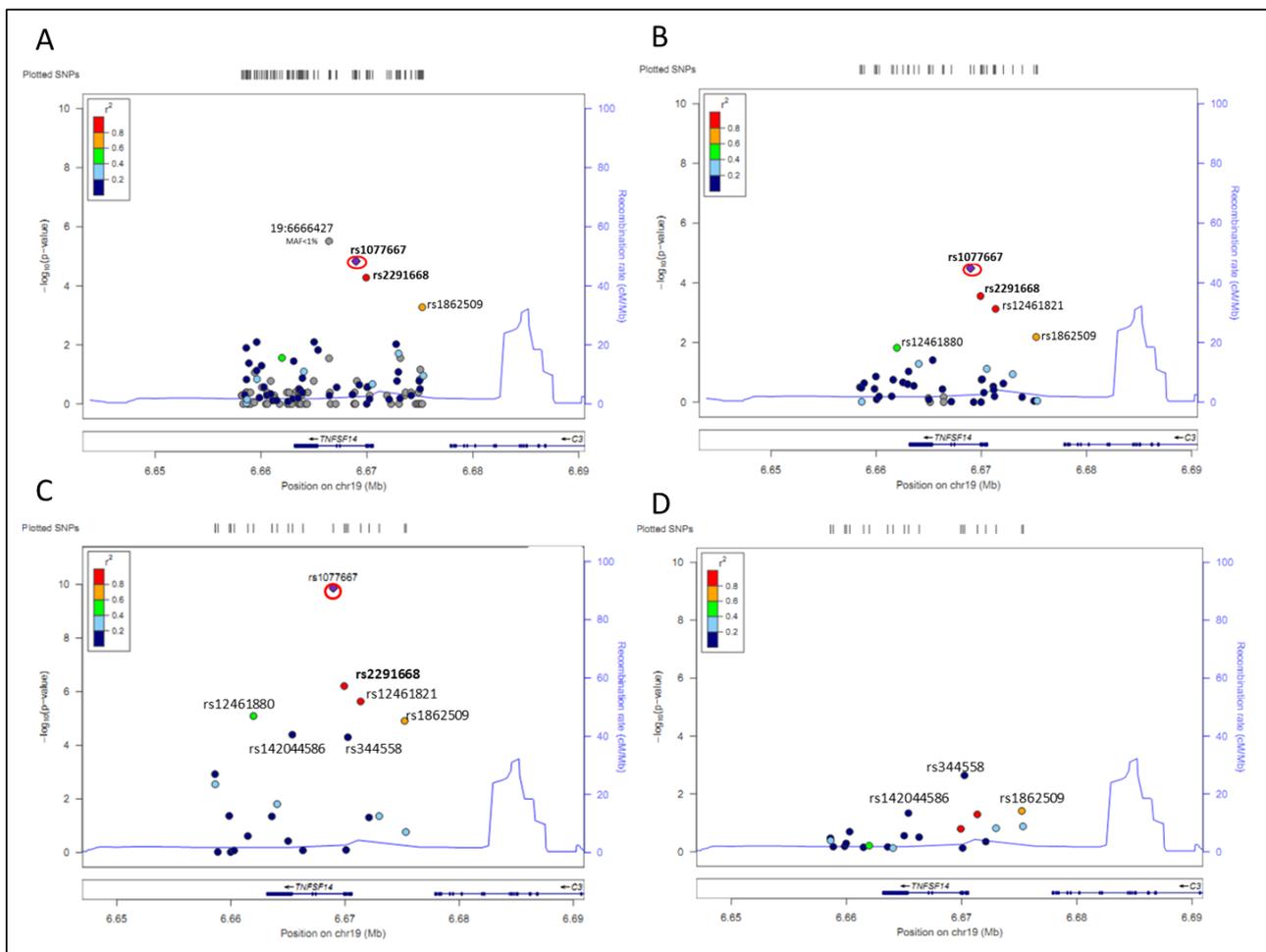
Figure 11: Regional association plots from resequencing data on *TNFSF14* gene (A), from target genotyping platform analysis (B), from meta-analysis of target genotyping platform, GWAS data set 1 and GWAS data set 2 (C), and from meta-analysis conditioning for rs1077667 (D). The two bold signals are the intronic variant (blue dot, indicated with a circle) and exonic variant (red dot). In the A section, the grey dot with the highest association signal is a rare (MAF in MS patients <1%) intronic variant not reported in public databases showing no statistically significant association in target genotyping platform analysis (p value = 0.68, Supplementary table S2).

4.2.3 Rare variants in *TNFSF14* gene

To investigate the role of rare variants (MAF<0.01) in the *TNFSF14* coding region, we sequenced the coding regions of the gene through two NGS experiments analyzing a total of 1092 MS patient and 912 controls pooled in groups of 12 individuals (as described in materials and methods).

We found 10 variants in coding regions (8 with MAF<0.01) and 3 of these were non synonymous (2 with MAF<0.01 in MS patients) (table S3, appendix). No splicing or nonsense variation was observed. To increase the statistical power, we tested the cumulative effect of rare variants performing two burden tests which consider only missense variants or synonymous and missense variants, respectively. We did not observe any statistically significant difference in the cumulative allele counts between MS patients and controls in both sequencing experiments (table S4 supplementary materials).

4.2.4 eQTL data

We investigated eQTL data as available in several public databases (Geuvadis consortium (Lappalainen et al., 2013), Blood eQTL browser (Westra et al., 2013), Bioportal (Holm et al., 2010), Gtex portal (Melé et al., 2015) and Brain eQTL Almanac (Ramasamy et al., 2014), as shown in table 1. Carriers of MS risk allele (C) were consistently found to have a lower *TNFSF14* expression in EBV-transformed lymphoblastoid cell lines, in peripheral whole blood cells, and in peripheral blood mononuclear cells (PBMCs). No significant eQTL association was found on brain cells apart from hippocampus (table 1).

	Geuvadis consortium	Bioportal	Gtex portal			BRAINeac	Blood eQTL browser
Tissue (n samples)	EBV-cells (465)	EBV-cells (270)	EBV-cells (140)	PBMC (338)	Hippocampus (81)	Brain cells (134)	Whole blood (5,311)
Effect	Beta value -0.3	Beta value -0.08	Beta value -0.33	Beta value -0.11	Beta value -0.34	Beta value < 0	Z-score -14.41
p-value	1.22E-09	0.034	0.0081	0.0026	0.012	Not significant	4.362E-47

Table 1: Association between *TNFSF14* expression levels and genotype of the primarily associated variant (rs1077667 intronic variant) across different databases, in EBV-cells, in PBMC, in whole blood and in brain. The C allele (which is

the risk allele) is significantly associated with a decrease in LIGHT expression (beta <0), except for BRAINeac database. This association withstands Bonferroni correction for Geuvadis consortium and Blood eQTL browser.

These data are consistent with the allelic imbalance analysis for the exonic variant rs2291668 (in high LD with rs1077667, $r^2 = 0.808$), performed on RNAseq data from EBV cell lines of 97 heterozygous individuals (Geuvadis consortium). The allelic expression for rs2291668 is unbalanced to the detriment of the allele in phase with the rs1077667 allele with lower eQTL expression (C allele, percentage of individuals with C<T: 68%, Wilcoxon paired-samples test: p<0.0001). This same trend was observed in EBV cell lines (33 samples, C<T: 60.7%) and in PBMC cells (129 samples, C<T: 53.2%) from Gtex consortium although these differences were not statistically significant. The meta-analysis in EBV cell lines from the two above data sets (Geuvadis consortium and Gtex consortium) confirms this allelic imbalance effect (p=0.001). Conversely, a significant opposite trend was observed for other tissues from Gtex database (lung, 23 samples, C<T: 21.7%, p= 0.0264; liver, 37 samples, C<T: 16.2 %, p<0.0001; adipose visceral tissue, 56 samples, C<T: 17.9%, p<0.0001; esophagus muscularis, 20 samples, C<T: 20%, p=0.0020).

4.2.5 *TNFSF14* expression analysis in MS patients and controls

TNFSF14 transcript exists in two different isoforms: a full-length transcript (consisting of 1491 nt) and a small transcript (named as ΔTM isoform, of 1169 nt) (Granger et al., 2001). This alternative transcript is generated by joining the cryptic splice donor in exon 1, at nucleotide position 111, to the splice acceptor that defines the beginning of exon 2, at nucleotide position 218, resulting in the removal of 107 nucleotides (comprising the exonic SNP rs2291668) including the region encoding the transmembrane domain.

To confirm eQTL results, derived from publicly available datasets, we evaluated the *TNFSF14* expression by SYBR green qPCR of full-length and ΔTM isoforms in two different cohorts: 1) PBMCs (frozen pellet) from 64 healthy controls and 45 patients, and 2) whole blood of 16 controls and 39 patients. In both sample sets, MS patients were treatment-naïve. The expression levels of the two isoforms were highly correlated (r = 0.89, p <0.0001), (figure 12) indicating that they are regulated by shared transcriptional mechanisms.
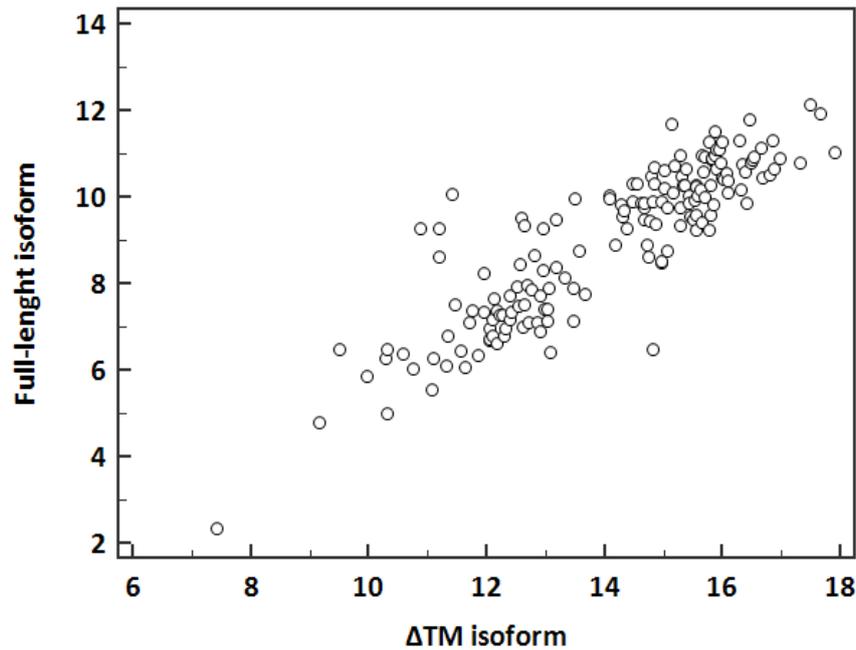
Figure 12: Scatter plot showing the correlation between the two isoforms derived from *TNFSF14* alternative splicing. Values of ΔCT for ΔTM isoform (X-axis) and of the full-length isoform (Y-axis) have been plotted. Each dot corresponds to an individual. Coefficient of correlation: r = 0.89, p <0.0001.

We performed two linear regression analyses (tables 2 and 3) in order to test the association of rs1077667 genotypes with *TNFSF14* expression, accounting for possible confounders (cohort and disease status effect). We observed that the presence of the rs1077667 risk allele is associated with lower levels of *TNFSF14* transcript for both isoforms (full-length: p= 0.0045, ΔTM: p=0.0021), paralleling what previously observed for in eQTL data. Moreover, when using the genotype as a covariate, patients always were the minor producers compared to healthy controls (full-length: p=0.011, ΔTM: p=0.0242).

Table 2: Association between *TNFSF14* expression levels and genotypes of the primarily *TNFSF14* associated variant.

Variant	Allele	*TNFSF14* Isoform	effect (beta value)	P-value
rs1077667	T	full-length	-0.3872	**0.0045**
	T	ΔTM	-0.5347	**0.0021**

A linear regression analysis correlating the number (0, 1 or 2) of the minor (protective) T allele and PCR real time ΔCT values, covariated for sex, cohort and disease status, was performed. The higher the number of protective allele (0,1,2) the lower the PCR real time ΔCT values (beta minor than 1). This means that the higher the number of susceptibility (common) C alleles the lower the TNFSF14 expression levels.

Table 3: Association between *TNFSF14*
expression levels and disease status**.**

TNFSF14 Isoform	Conditional variant	effect (beta value)	P-value
full-length	rs1077667	0.3906	**0.011**
ΔTM	rs1077667	0.4391	**0.0242**

Linear regression analysis covariated for sex, cohort and genotypes of the associated variant rs1077667. ΔCT values are significantly higher in MS patients (thus expression levels are significantly lower) for both isoforms.

Altogether, these results showed that *TNFSF14* RNA expression levels in blood cells were lower in MS patients than in controls, and this is consistent with the lower expression levels observed in carriers of the risk allele of the primarily MS associated variant.

4.2.6 LIGHT protein expression in blood cells according to the genotypes at the intronic variant

To evaluate the association of rs1077667 C/T variant with LIGHT transmembrane protein expression, we performed flow cytometry analyses on peripheral blood of 20 healthy controls with different rs1077667 genotypes (10 CC, 8 CT, 2 TT), testing the percentage of LIGHT positive cells in CD4+ T cells, CD8+ T cells, myeloid dendritic cells (DC) CD11c+, monocytes (CD14+), NK cells (CD56dim/CD16bright, CD56-/CD16bright, CD56dim/CD16-), and B cells (CD19+) (figure 13 A). These flow cytometry analyses were performed thanks to the cooperation with the Immunology laboratory headed by professor Umberto Dianziani at University of Eastern Piedmont, department of Health Sciences, in Novara. The mean percentage of LIGHT positive cells among the analyzed populations ranged between 17.4% (CD4+ T cells) and 0.3% (B cells, the only cell population with less than 1% of LIGHT positive cells). Among cell populations with a percentage of LIGHT positive cells higher than 1%, we observed a significant association with rs1077667 genotype in myeloid dendritic cells (CD11c+). Indeed, the mean percentage of LIGHT positive cells in homozygous individuals for the MS risk allele (CC) was significantly higher compared to individuals with other genotypes (CT+TT) (Student's t-test, p= 0.02), (figure 13 A).

The association was maintained (Student's t-test, p=0.035) when we extended flow cytometry analysis in this cell population (CD11c+) to additional controls, leading to a total of 37 healthy individuals (18 CC, 15 CT, 4 TT) (figure 13 B).
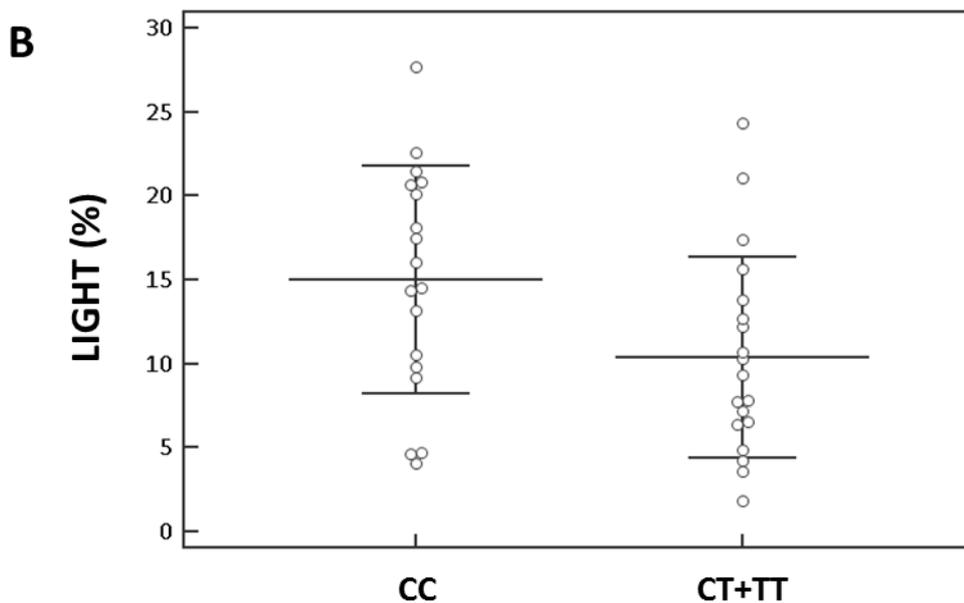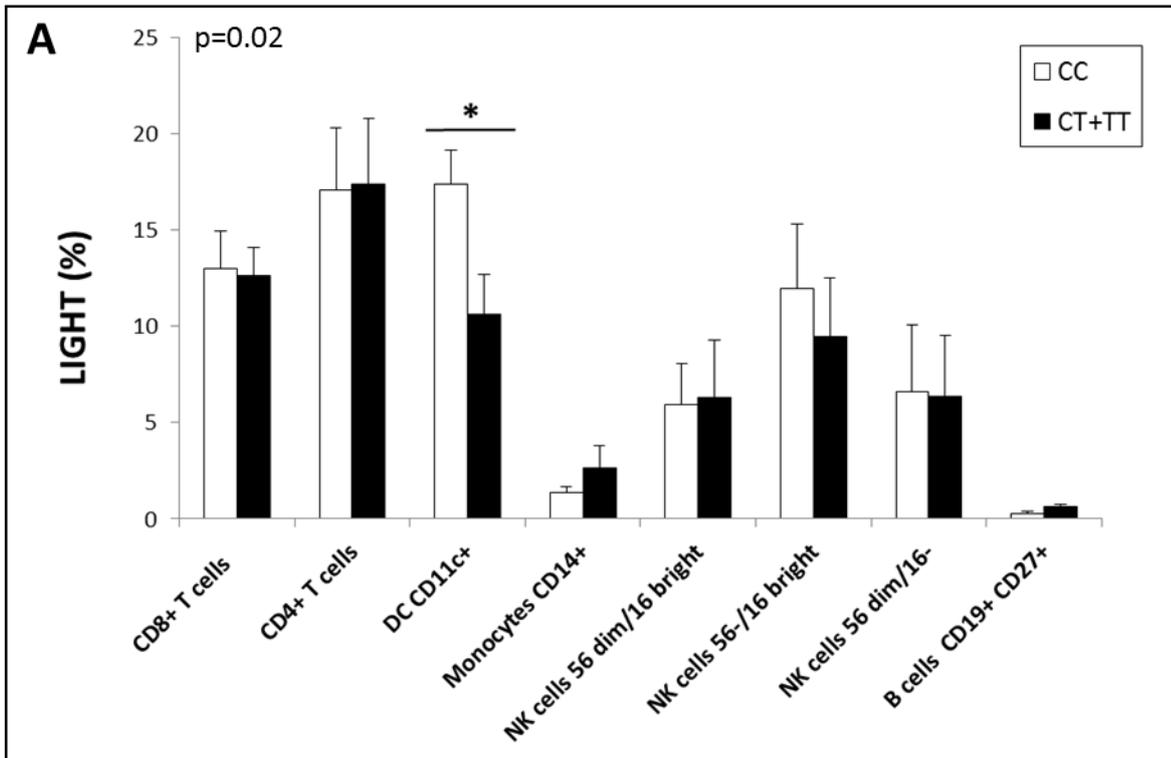
Figure 13: Mean percentage (+/- Standard Error, SE) of LIGHT positive cells detected by flow cytometry in CD4+ T cells, CD8+ T cells, myeloid dendritic cells CD11c+, monocytes (CD14+), NK cells (56dim/16bright, 56-/16bright, 56dim/16-), and B cells (CD19+), in blood from 20 healthy controls stratified according to the rs1077667 genotype (CC vs CT+TT) (A). Percentage of LIGHT positive myeloid dendritic cells (CD11c+) of 37 individuals stratified according to the rs1077667 genotype (CC vs CT+TT). Each dot represents an individual, the lines in the graph represent the mean and standard deviation (B).

To choose the best in vitro model of dendritic cells to replicate the data observed ex-vivo from blood, we compared LIGHT expression in five different monocyte-derived dendritic cell (MDDC) populations, obtained by culturing monocytes from 5 healthy donors for 5 days with GM-CSF+IL-4, or GM-CSF+IFNβ, or IL-3 alone, or IL-3+IFNβ, or GM-CSF+IL-15 (figure 14).

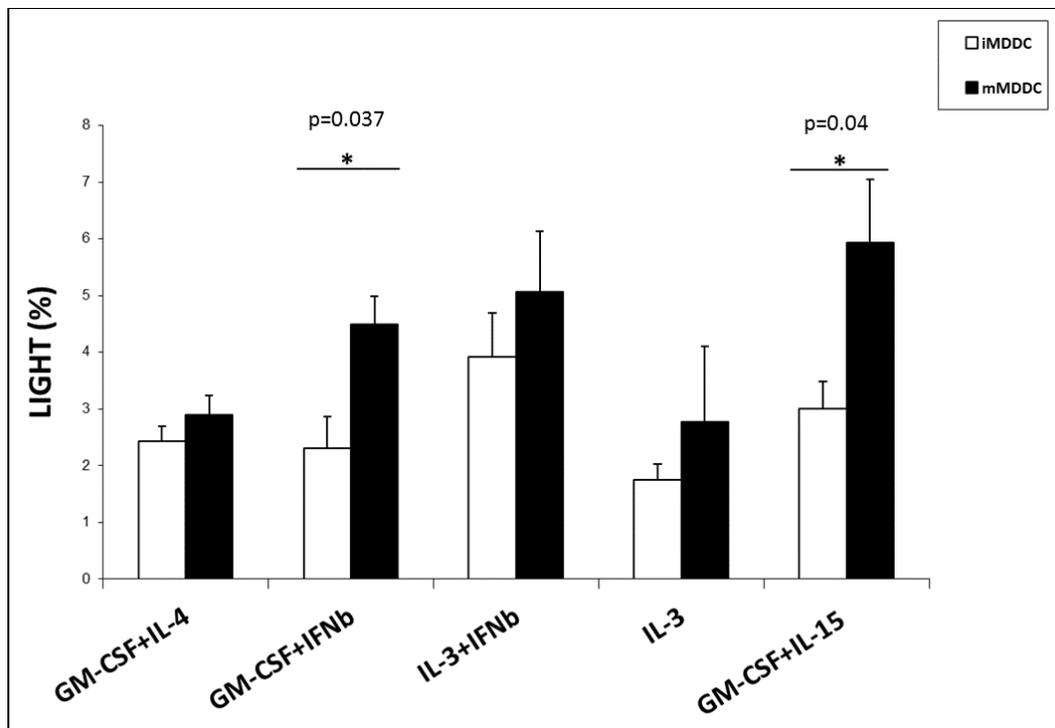**Figure 14**: Mean percentage (+SE) of LIGHT positive immature (iMDDC) and mature (mMDDC, LPS-activated) monocyte-derived DC (MDDC) obtained by culturing monocytes with GM-CSF+IL-4, GM-CSF+IFNβ, IL-3, IL-3+IFN β, GM-CSF+IL-15 detected by flow cytometry, from 5 healthy donors.

Then, surface expression of LIGHT was analyzed by flow cytometry in these MDDC cultured for 2 days in the presence (activated/mature MDCC) or absence (resting/immature MDCC) of LPS. Results showed that LIGHT expression was relatively weak in all types of immature MDDC and tended to be upregulated in mature MDDC, particularly in MDDC obtained with GM-CSF+IL-15 (Student's t-test p= 0.04) or GM-CSF+IFNβ (Student's t-test p=0.037) where the upregulation was statistically significant (figure 14).

Then, we selected the MDDC population showing the highest percentage of LIGHT positive cells, which was that obtained by culture with GM-CSF+IL-15 (MDDC^{IL15}), and we increased the sample set to a total of 22 healthy controls with different genotypes for the rs1077667 C/T variant (12 CC, 9 CT, 1 TT). This analysis showed that individuals carrying the MS risk genotype (CC) had a higher percentage of LIGHT positive cells compared to the other individuals (CC+CT), with a statistically significant difference in mature MDDC^{IL15} cells (Mann-Whitney test, p=0.04) (figure 15).

Figure 15: Mean percentage (+SE) of LIGHT positive immature (iMDDC) and mature (mMDDC, LPS-activated), monocyte-derived DC (MDDC) obtained by culturing monocytes with GM-CSF+IL-15 from 22 healthy donors (12 CC, 9 CT, 1 TT), (detected by flow cytometry) stratified according to the rs1077667 genotype.

Since LIGHT can be produced in soluble form, we also analyzed LIGHT protein released in the culture supernatants of the immature and mature MDDC^{IL15} from the set of 22 healthy donors and compared it with the LIGHT surface expression. Results confirmed that LIGHT surface expression was significantly higher in mature than in immature MDDC^{IL15} (paired Student's T-test p=0.0002) (figure 16 A). By contrast, soluble LIGHT levels displayed an opposite behavior since high levels were detected in the supernatants of immature MDDC^{IL15} and they were down-modulated in the mature counterpart (paired Student's T-test p=0.007) (figure 16 B). This modulation of LIGHT surface and soluble expressions upon MDDC^{IL15} maturation was not influenced by the rs1077667 intronic variant, since no significant differences have been observed according to the genotype of rs1077667 the intronic variant (data not shown).
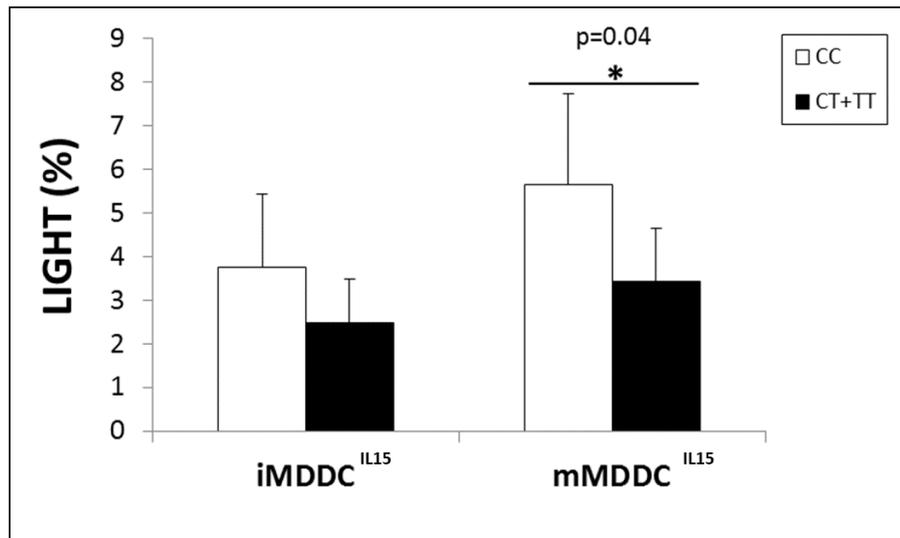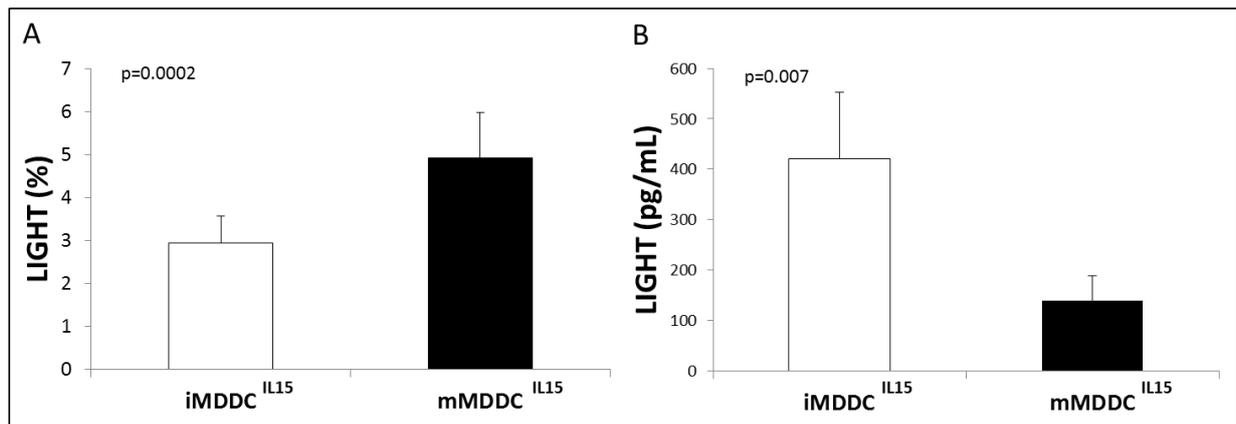
Figure 16: Mean percentage (+SE) of LIGHT positive immature (iMDDC) and mature (mMDDC, LPS-activated), monocyte-derived DC (MDDC) (detected by flow cytometry) (A) and soluble LIGHT (detected by ELISA) (B) obtained by culturing monocytes with GM-CSF+IL-15 from 22 healthy donors.

Finally, we evaluated the production of different cytokines (IL-6, TNFα, IL-10, IL-23) in the supernatants of mature MDC cells obtained with GM-CSF+IL-15 (N=22: 12 CC, 9 CT, 1 TT) but no significant differences have been observed according to the genotype of the intronic variant.

4.3 Discussion

The genetic association between *TNFSF14* locus and MS was reported in IMSGC GWAS in 2011 (Sawcer et al., 2011), and subsequently confirmed in international genomic studies (Beecham et al., 2013; International et al., 2017).

Through a massive sequencing and a fine mapping approach in the Italian population, we were able to identify the rs1077667 SNP located in intron 1 of the *TNFSF14* gene, as the primarily MS associated variant at the *TNFSF14* locus. The intronic SNP was identified as the primarily associated variant in this region after a fine mapping which has taken advantage of the resequencing of the region performed in a large cohort of Italian patients and controls. According to computationally integrated ChIP-seq data using a Hidden Markov Model (HMM) (36) by UCSC genome browser, the rs1077667 SNP is located in a 2600 bp region showing an enrichment of the H3K4Me3 histone mark and identified as active promoter, suggesting a regulative role of the SNP in *TNFSF14* RNA expression. This hypothesis is supported by cis-eQTL data from different databases, showing that carriers of MS risk allele have a lower *TNFSF14* RNA expression in EBV-transformed lymphoblastoid cell lines (Geuvadis, Bioportal, Gtex), in PBMCs (Gtex) and whole blood (Blood eQTL browser). These data are consistent with the imbalance against the risk allele observed in heterozygous individuals in EBV-transformed lymphoblastoid cell lines (Geuvadis). We confirmed these data by RT-PCR expression

analysis on PBMC of 84 Italian MS and 80 healthy controls, in which we tested the expression of both *TNFSF14* isoforms (full length and ΔTM). We observed that both patients and controls with risk genotype (CC) produced lower levels of both *TNFSF14* isoforms than the carriers of the other genotypes. This downmodulation was reported for the first time at the RNA level, and is consistent with the results previously reported at LIGHT protein level in the serum (Malmeström et al., 2013). Moreover, we observed that patients were minor producers compared to controls for both isoforms independently of the TNFSF14 genotype. A similar trend was observed on smaller sample sets without any analysis of TNFSF14 genotype (Jernås et al., 2013; Romme Christensen et al., 2013).

In silico predictions on TRANSFAC (Matys et al., 2006) and MatInspector (Cartharius et al., 2005) suggest that rs1077667 variant can modify the binding of the Aryl Hydrocarbon Receptor (AhR) transcription factor. Particularly, the consensus of this transcription factor is predicted only in the presence for the MS risk allele. AhR was initially discovered and well characterized as a transcription factor responsible for the activation of genes encoding different enzymes involved in the metabolism of xenobiotics (Vogel et al., 2014). Further studies indicated that activation of AhR plays different roles also in other cellular functions, including the regulation of the immune system (Singh et al., 2007), such as in the differentiation of T-lymphocytes, in particular regulatory T cells and T helper 17 (Th 17) (Kimura et al., 2008). The involvement of AhR transcription factor in the differentiation of Th 17 lymphocytes is particularly intriguingly since the expansion of T helper lymphocytes (Th 17) in peripheral blood is associated with the active phase of multiple sclerosis (MS) (Durelli et al., 2009).

Inflammation is one of the key pathogenic mechanisms in multiple sclerosis, at least in the early stages (Frischer et al., 2009). A downmodulation of gene expression in peripheral blood of MS patients was reported for genes involved in regulation of NF-kB pathway, such as NR4A2 which presents a key role in protecting neurons from inflammation induced neurotoxicity mediated by NF-kB pathway signalling) (Navone et al., 2014). LIGHT involvement in NF-kB pathway activation is well known, and our findings about LIGHT expression levels on peripheral blood parallels what reported for NR4A2, so we can speculate that *TNFSF14* can play a similar role. LIGHT, through the link with LTβR, in particular contexts is able to induce cell death due to the recruitment of TRAF 3 and the activation of caspases (Granger and Rickert, 2003). Furthermore Shui et al., in 2011, highlighted a dual role of LIGHT receptor HVEM: the interaction between LIGHT and HVEM has a costimulatory effect, on T cell activation, while the interaction between HVEM and BTLA has an opposite function, leading to inhibition of the activation of T lymphocytes (Shui et al., 2011). Altogether, these data suggest that an imbalance in the modulation of LIGHT production may result in a predisposition to the development of inflammatory conditions and neuropathy.

The key role of LIGHT in determining MS pathogenesis is clear in MS murine model, experimental autoimmune encephalomyelitis (EAE). In fact, LIGHT-deficient mice developed severe EAE resulting in an atypically high mortality rate (Maña et al., 2013) . The same authors demonstrated that in EAE mice LIGHT expression was crucially involved in controlling activated macrophages/microglia during autoimmune CNS inflammation (Maña et al., 2013).

LIGHT has also a known crucial role in the maturation of dendritic cells by cooperating with CD154 (CD40 ligand) (Morel et al., 2001) and interacting with licensed natural killer cells (Holmes et al., 2014). Interestingly, we showed that, LIGHT is expressed in peripheral blood myeloid dendritic cells (CD11c+) and the MS risk allele of rs1077667 is associated with an increased percentage of LIGHT positive cells. These ex-vivo results were confirmed using MDDC differentiated in vitro by culturing monocytes in different cytokine milieus. This analysis showed that immature MDDC express low levels of LIGHT which are up-modulated in mature MDDM, particularly in MDDC^{IL15}. Moreover, an expanded analysis of MDDC^{IL15} showed that the MS risk allele is associated with increased percentages of LIGHT positive cells in mature MDDC.

Our data do not confirm those from a previous work showing that immature MDDC differentiated with GM-CSF+IL-4 express high levels of LIGHT, which are downmodulated upon treatment with LPS (Tamada et al., 2000). The inconsistency with our data might be ascribed to technical differences since, in that work, LIGHT was detected by indirect immunofluorescence using a polyclonal rabbit antibody raised against a small synthetic peptide derived from LIGHT conjugated to KLH. By contrast, we used an anti-LIGHT mAb detected by direct immunofluorescence, which would increase the analysis specificity (Tamada et al., 2000).

The association detected in DC between the MS risk allele and increased percentages of surface LIGHT positive cells is in apparent contrast with our observation that this allele is associated with decreased LIGHT mRNA levels in peripheral blood cells (whole blood or PBMCs) and EBV-derived B cell lines. In peripheral blood cells, the LIGHT mRNA levels may be substantially influenced by changes in the distribution of the different white cell types, which basally express different amounts of LIGHT. For instance, T cells express high levels of LIGHT, whereas B cells and monocytes express low levels, so that unbalance between these cell types would profoundly influence the overall blood LIGHT expression. By contrast, changes of LIGHT expression in myeloid dendritic cells, which represent a small minority of white blood cells (1-3%), would have a minimal impact on the overall blood LIGHT expression. However, this explanation does not apply to EBV-derived B cell lines, in which the risk allele might influence LIGHT mRNA expression. At this light, it is intriguing that we detected a minimal percentage of peripheral blood B cells expressing LIGHT, but this percentage was significantly lower in subjects carrying the risk allele than in the other subjects (0.3±

0.3% vs 0.6±0.4%, p=0.033), which is in line with the EBV-derived B cell lines RNA data. Notably, our analyses of eQTL databases reinforced the hypothesis that the eQTL effect of rs1077667 on *TNFSF14* is tissue-dependent, since in some tissues we detected an effect size displaying an opposite direction of effect than that displayed by PBMCs. However, no data specifically focused on dendritic cells were available in these eQTL databases.

Since LIGHT can be produced in soluble form, we also analyzed LIGHT protein released in the culture supernatants of MDDC[IL15] and found an inverse relationship between the soluble and surface form since immature MDDC[IL15] expressed high levels of the former and low levels of the latter, which was specular as compared to mature DC. However, no significant association was found between production of soluble LIGHT and the risk allele. These data are referred to the production of soluble LIGHT in DC and, thus, are not comparable with the significant association reported by Malmeström (Malmeström et al., 2013) between rs107667 genotypes and serum levels of LIGHT that may reflect the activity of multiple cell and tissue types, with only a minor contribution of dendritic cells.

The involvement of DCs in MS has been well established in humans and in mice. In particular, Serafini et al. (Serafini et al., 2006) detected both immature and mature DCs in the meninges and parenchymal lesions of patients with primary and secondary progressive MS. MDDC from the cerebral spinal fluid of patients with MS have been found to display a mature phenotype since they express high levels of HLA-DR, CD80, CD86 and CD40, and produce high levels of the pro-inflammatory cytokine IL-6. (Pashenkov et al., 2001)

LIGHT has a crucial role in the maturation of DC by cooperating with CD154 (CD40 ligand) (Morel et al., 2001) and interacting with licensed natural killer cells (Holmes et al., 2014). Interestingly, a previous work showed that LIGHT expressed by DC supports IFNγ production by T cells (Tamada et al., 2000), which is relevant since IFNγ plays a key role in the cell mediated autoimmune response of MS. This fits with our data highlighting the role of LIGHT in MDDC[IL15], which have been reported to be particularly efficient in supporting cytotoxic T lymphocytes (CTLs) playing a key role in cell-mediated autoimmunity of MS (Banchereau and Palucka, 2005).

In conclusion, our work suggests that an altered TNFSF14 expression in immune cells driven by an intronic variant may contribute to MS pathogenesis. Particularly, the MS associated variant seems to be associated with a low *TNFSF14* RNA expression in a mixed population of PBMCs and with an increased percentage of LIGHT positive cells in DC, which may influence the cell-mediated autoimmunity in MS.

5. Chapter 3. Rare variants analysis

5.1 Introduction: state of art of the project

Despite these last discoveries, the mechanisms at the basis of MS susceptibility are not completely understood, and for this reason the research is still involved in the identification of new MS susceptibility genetic markers, including rare variants, which have not been so deeply studied so far. A recent paper (IMSGC, 2018) suggest that part of the missing heritability may be due to rare variants. The second aim of the project was to identify rare genetic variants involved in MS genetic susceptibility, in particular in those genes already known to be associated to the disease.

The genes and genetic regions that were analysed for rare variants were selected thanks to our participation to the two international efforts: IMSGC Genome Wide Association Study and Immunochip project, as previously described. The regions associated in both the International and in the Italian population, were selected for Next Generation sequencing on the Italian sample set, performed on pools of individuals.

One of the reasons for which rare variants have been poorly investigated in complex diseases is the need of very large sample sets to achieve sufficient power to perform association analysis on single variants. Consequently, we tested the effect of rare (MAF < 1%) and at low frequency (MAF 1-5%) variants as burden test comparing the total numbers of alternative and reference alleles in patients and controls. The rationale of this analysis, which tests the cumulative effect of rare variants in the susceptibility to the disease, arises from the fact that if a gene is involved in disease pathogenesis, it is possible that it may be affected by rare mutations, even different, in different patients. The analysis of the cumulative effect of these variants requires much smaller sample sizes than the ones needed in order to investigate the effect of the single rare variants.

The analysis was performed in two phases:

1) *Discovery phase*: we selected a list of 90 regions for a total of 1.9 Mb associated in both the International and in the Italian population: 17 regions containing 27 genes (including introns, exons, 5'UTR and 3'UTR) and 73 genes (only coding and 5'UTR, 3'UTR). We sequenced 600 MS patients and 408 healthy controls grouped in pools of 12 individuals each. Patients were selected with the attribution of a "risk score" specifically designed on the sequenced regions, which allowed to associate to each patient a score proportional to its genetic risk and so to identify the patients with the highest number of risk alleles for polymorphism associated to the disease, present in the regions selected for the sequencing. The risk score had been calculated based on the model proposed from De Jager et al (De Jager et al., 2009a). Instead controls were individuated through a matching phase

with patients through statistical analysis that allowed to match to each single patient one genetically similar control, with the Principal Component Analysis (PCA, Principal Component Analysis).

2) *Replication phase:* The genes that showed a significant burden were sequenced on a second independent sample set (504 MS and 504 healthy controls), pooled in groups of 12 individuals and following the same pipeline used for the discovery phase. The collection of MS patients was unbiased, not influenced by the genetic risk score. These analyses were performed thanks to the contribution of the laboratory of Human Genetics of Neurological Disorders at San Raffaele Hospital headed by professor Filippo Martinelli Boneschi and Federica Esposito and of Institute of biomedical technologies at CNR institute headed by professor Gianluca De Bellis.

It was necessary to conduct appropriate quality controls for the methodology used (sequencing analysis in pools of individuals) in order to avoid incorrect allele frequencies and false positives. The 600 MS patients in the first sequencing experiment have been previously individually genotyped either with the Illumina 660Q chip or with the Immunochip platform and AF comparison with these platforms demonstrated a high correlation with AF in the pools ($R^2=0.987$). Similarly, we observed an high correlation also between pooled AF and frequencies reported in public databases (1000 genomes_EUR $R^2=0.980$, ExAC $R^2=0.970$) (Anand et al., 2016). This comparison highlighted that the NGS pooled method can estimate the allelic frequencies in patients and controls in an accurate way, suitable for the subsequent analysis.

In order to remove false variants due to possible sequencing errors, we defined a threshold on the observed frequency of the alternative allele within the single pool (0.026 for the discovery, 0.024 for the replication), empirically determined as described in Anand et al., and we recalculated the AF applying this threshold (Anand et al., 2016). The number of alternative reads present in each pool was considered, for the calculation of the allelic frequency, only if the frequency of the alternative allele exceeded the established threshold in the same pool.

With this analysis we achieved two main goals:

1) we performed a burden test on rare and low frequency variants on the selected genes.

2) for some genes that were fully sequenced also in non coding regions (such as *TNFSF14* and *TNFRSF14*), frequency of common variants was used as first-pass fine mapping of the region (as described in chapters 2 and 4).
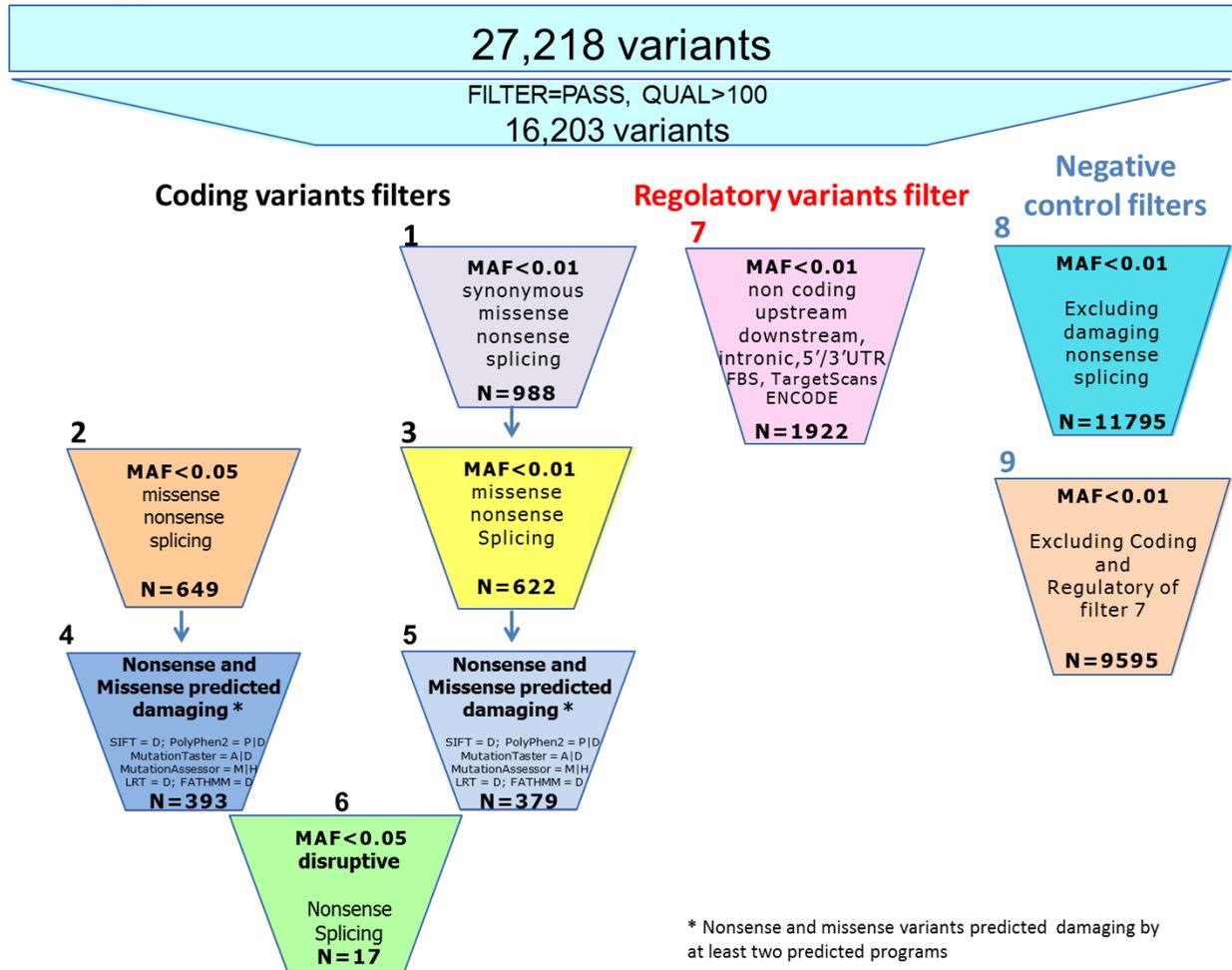
5.2 Discovery phase results

Figure 17: Description of the filters used for the burden test analysis in the discovery phase. Filters 1,2, 3,4,5, and 6 were used for coding variants; filter7 for noncoding variants. Filters 8 and 9 contain all rare variants excluding those with a suggested functional role, and they were used as negative controls.

We sequenced 84 pools (12 individuals per pool) on Illumina GaIIx sequencer producing 2×85 bp read lengths and we analysed our data by an ad-hoc bioinformatics pipeline (as described in materials and methods). One of the pools did not pass quality controls, thus it was re-sequenced without success and was discarded from further analysis. We generated 13.96 million reads per pool on average, the mean depth was 351.9x with more than 85% of the target regions covered by NGS reads in each pool. On average, 75% of the target regions were covered at least 50x and 69% were covered at least 100x. From the initial sequencing with NGS technology we obtained 27,218 variants (of these 54.93% were novel variants not found in public database and 98.38% of them are rare variants with AF<0.01). Variants were then filtered for quality controls, retaining 16,203 variants.

We then performed 3 groups of functional based filters, for a total of 9 filters: coding variants filters (from 1 to 6), regulatory variants filter (7), negative control filters (8-9) (as reported in figure 17). We filtered the variants on the basis of their frequency (MAF<0.05, MAF<0.01), on their effect on the

protein sequence (missense, nonsense, splicing, synonymous), on the prediction of the effect of missense variations that we tested with 6 prediction programs (SIFT; PolyPhen2; MutationTaster; MutationAssessor; LRT; FATHMM), and, for non coding variants with the annotation based on TFBS, TargetScans, ENCODE. In particular in the group of coding variants (filters 1-6), we identified 988 variants for filter 1 (synonymous, missense, nonsense, splicing variants with a MAF<0.01); from this filter we derived the filter 3 (missense, nonsense, splicing variants with a MAF<0.01) in which we found 622 variants and the filter 5 (nonsense and missense variants predicted damaging by at least two prediction programs with a MAF<0.01) with 379 variants. For filter 2 (missense, nonsense, splicing variants with a MAF<0.5), we identified 649 variants, which become 393 in the derived filter 4 (nonsense and missense variants with a MAF<0.05 predicted damaging by at least two prediction programs). We also created in this category a filter for disruptive variants (filter 6) which included nonsense and splicing variants for a total of 17 variants (8 of which located in *EFCAB13* gene). Furthermore, we created a for variants with a possible regulatory role with a MAF <0.01 (filter 7) and we found 1922 variants. Finally, we created two negative control filters: filter 8 as control of filters for coding variants with 11,795 variants and filter 9 as control of filter 7, with 9,595 variants.

For each gene, the cumulative frequency of the variants selected from each of the applied filters, was compared between MS patient and control pools with three different statistical tools for burden test (WSS, C-ALPHA and Fisher Hybrid Test). The results of burden test identified 17 genes, showing a significant difference in the number of rare or low frequency variants between patients and controls in at last one of the 7 filters with at last one of the three statistical tests used for the analysis (figures 18-25). In particular, the highest signal was observed for a gene on chromosome 17, *EFCAB13* (EF-Hand Calcium Binding Domain 13) (p<1.0e-4), coding for a protein of unknown function. In fact, this gene showed a statistically significant burden test for all the 6 filters which contain variants that alter the coding sequence. In particular, *EFCAB13* showed 12 nonsynonymous, 5 stop-gain, 3 splicing and 10 synonymous variants. It was the gene with the highest number of variants with a probable severe effect on protein function ("disruptive" variants) accounting for 8 variants (5 stop-gain and 3 splicing variants). Seen these results, these 17 genes were selected for the replication phase in an independent cohort of MS patients and controls.
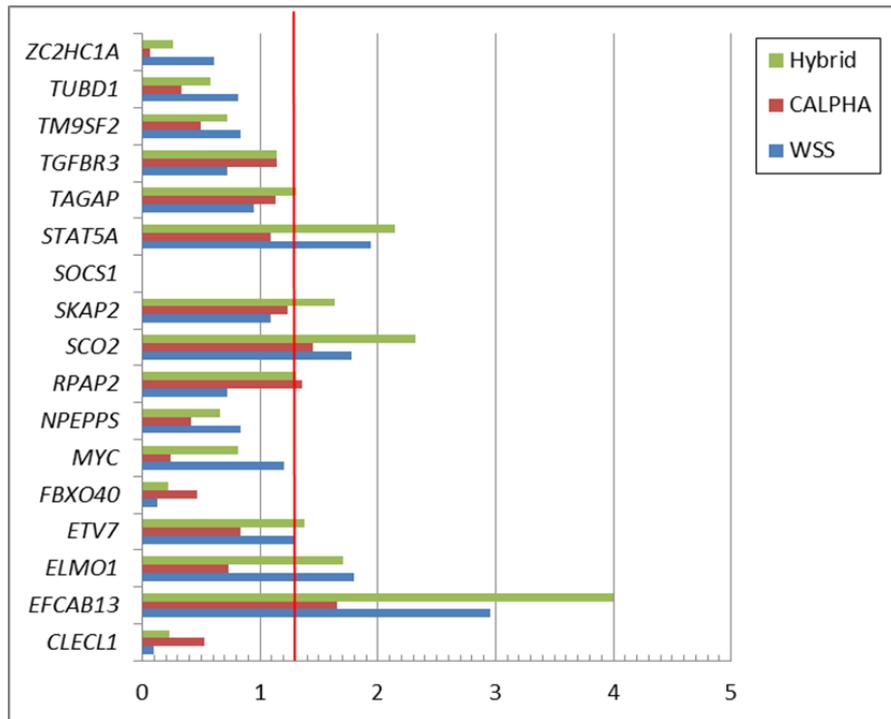
Figure 18: Results of Burden test after filter 1 (synonymous, missense, nonsense, splicing variants with a MAF<0.01). The bar-plots show the –log10 of the p-value obtained with the 3 tests. The red line corresponds to p=0.05
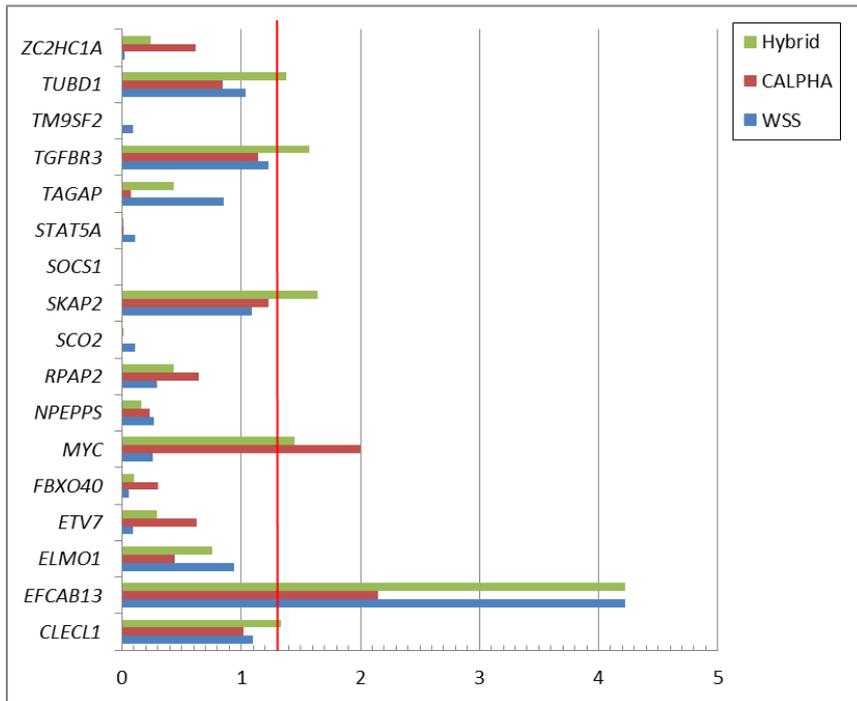
Figure 19: Results of Burden test after filter 2 (missense, nonsense, splicing variants with a MAF<0.5). The bar-plots show the –log10 of the p-value obtained with the 3 tests. The red line corresponds to p=0.05.
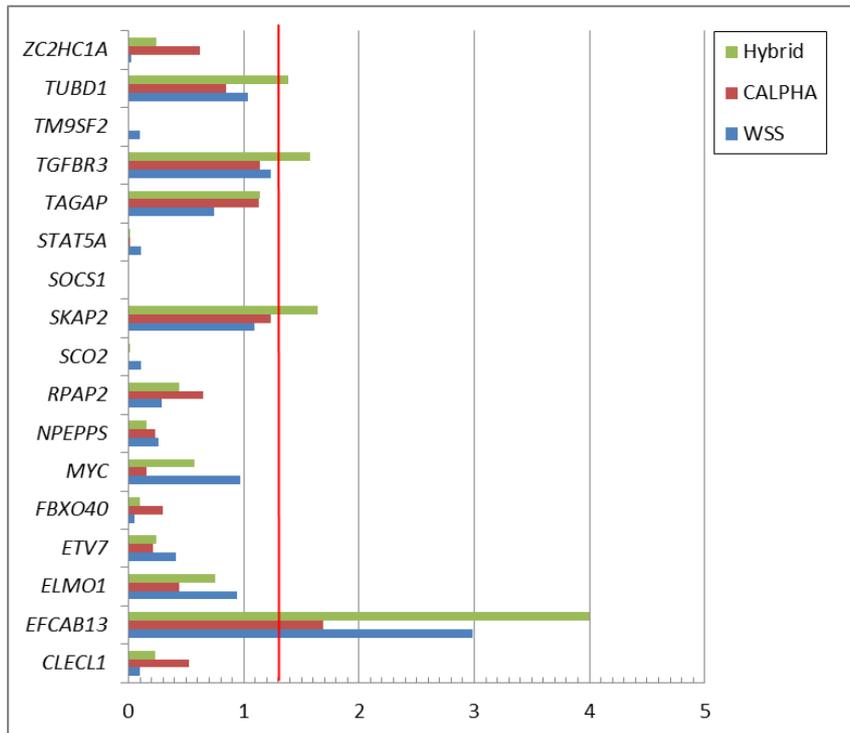
Figure 20: Results of Burden test after filter 3 (missense, nonsense, splicing variants with a MAF<0.01). The bar-plots show the –log10 of the p-value obtained with the 3 tests. The red line corresponds to p=0.05.
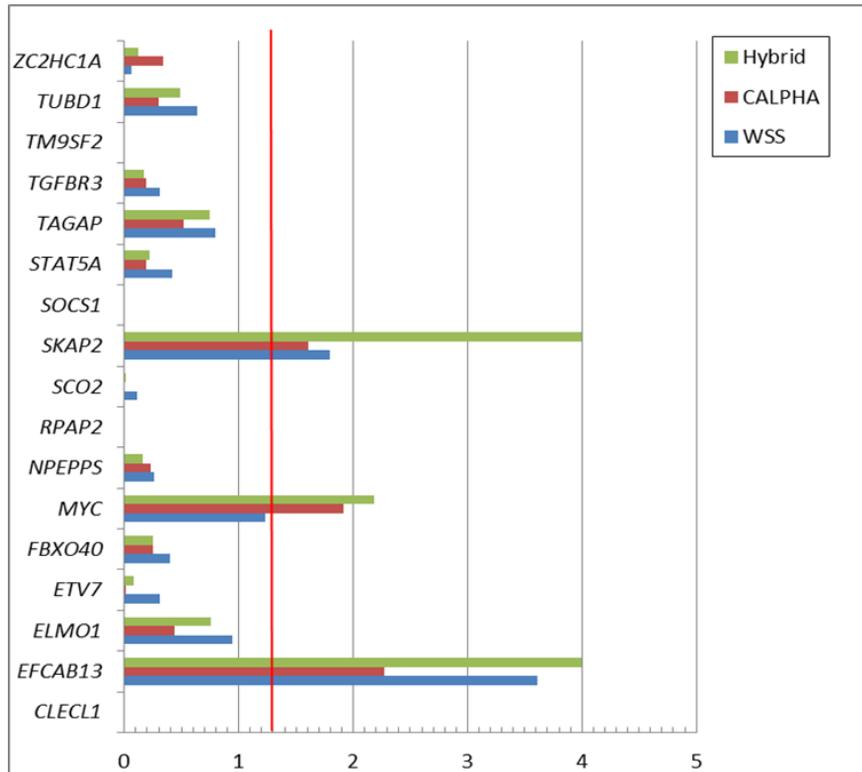
Figure 21: Results of Burden test after filter 4 (nonsense and missense variants with a MAF<0.05 predicted damaging by at least two predicted programs). The bar-plots show the –log10 of the p-value obtained with the 3 tests. The red line corresponds to p=0.05
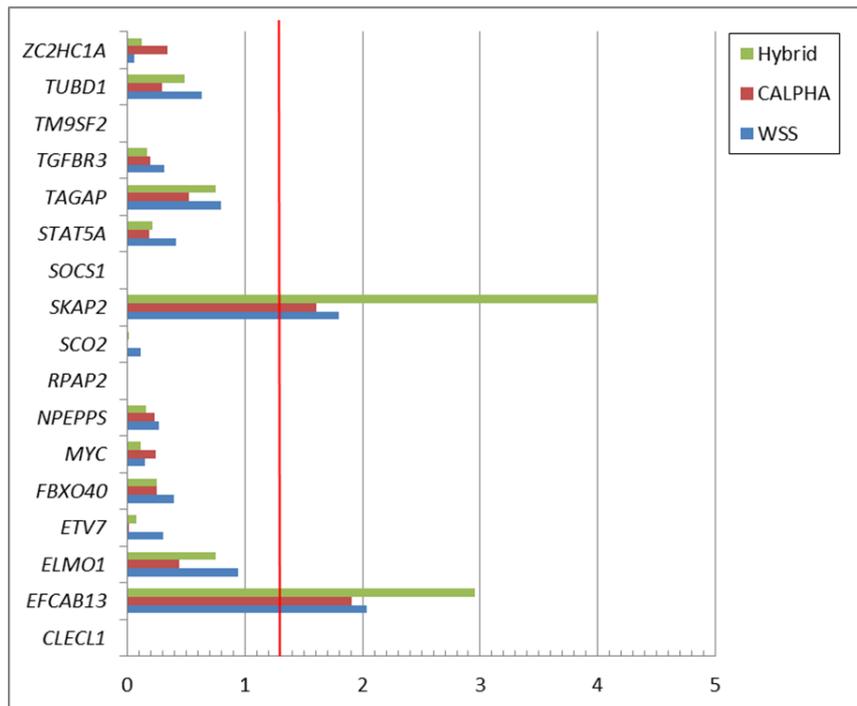
Figure 22: Results of Burden test after filter 5 (nonsense and missense variants predicted damaging by at least two predicted programs with a MAF<0.01). The bar-plots show the –log10 of the p-value obtained with the 3 tests. The red line corresponds to p=0.05
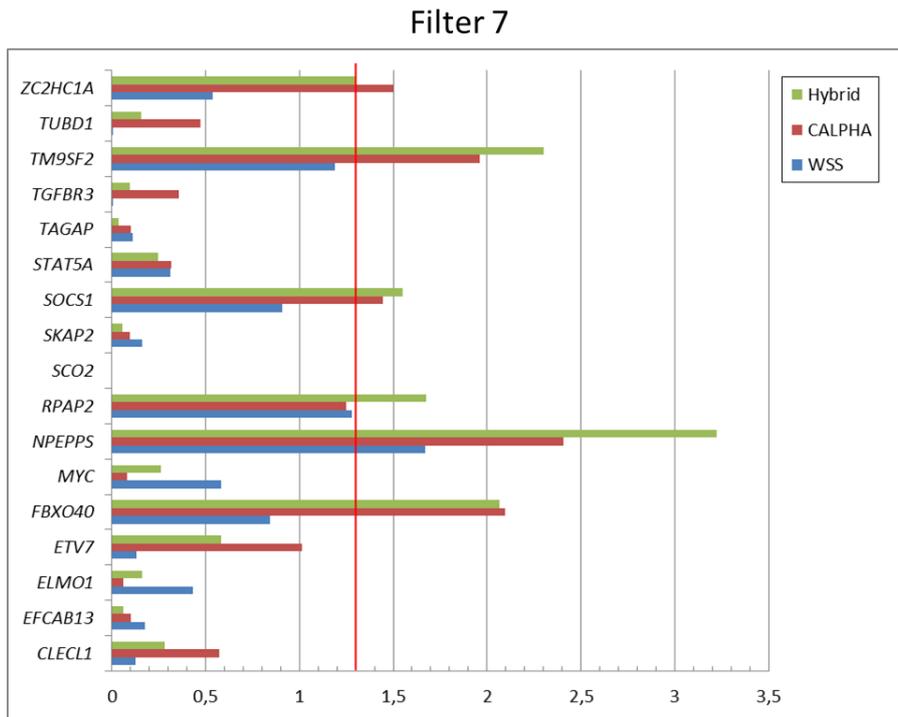
Figure 23: Results of Burden test after filter 7 (variants with a possible regulatory role with a MAF <0.01). The bar-plots show the –log10 of the p-value obtained with the 3 tests. The red line corresponds to p=0.05
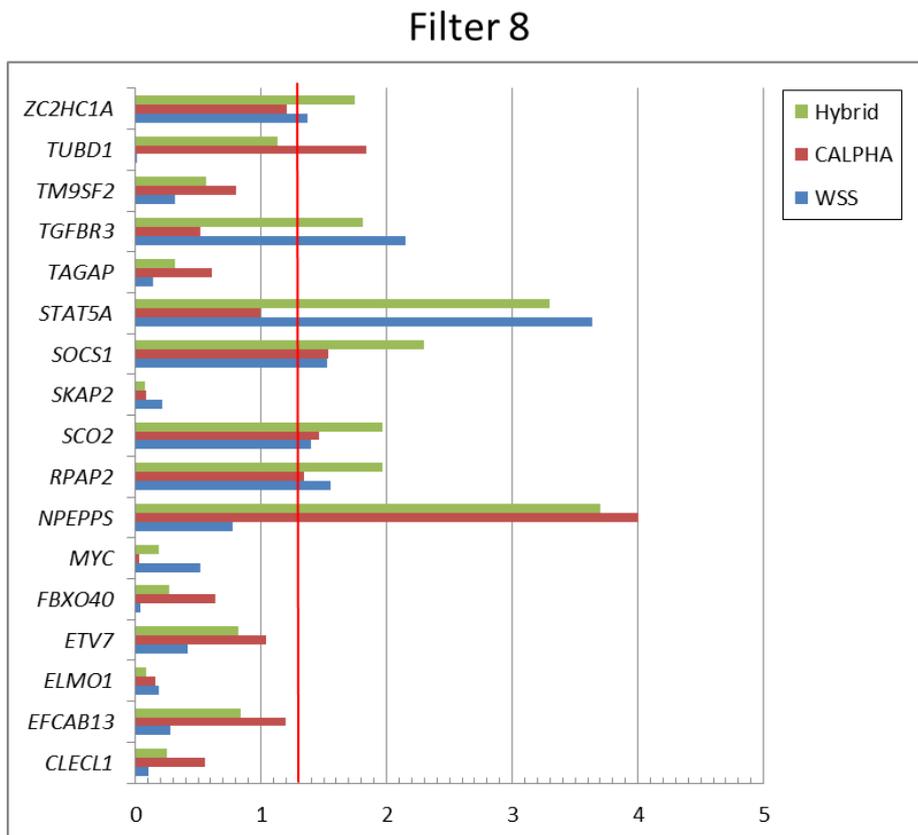
Figure 24: Results of Burden test after filter 8. The bar-plots show the –log10 of the p-value obtained with the 3 tests. The red line corresponds to p=0.05
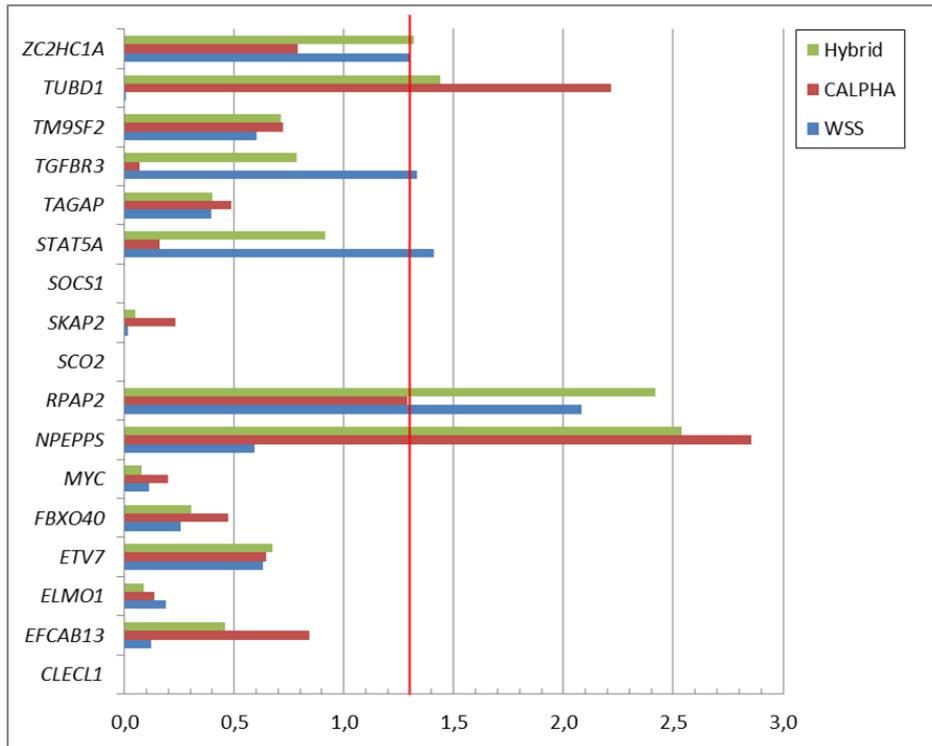
Figure 25: Results of Burden test after filter 9. The bar-plots show the –log10 of the p-value obtained with the 3 tests. The red line corresponds to p=0.05.

5.3 Replication phase results
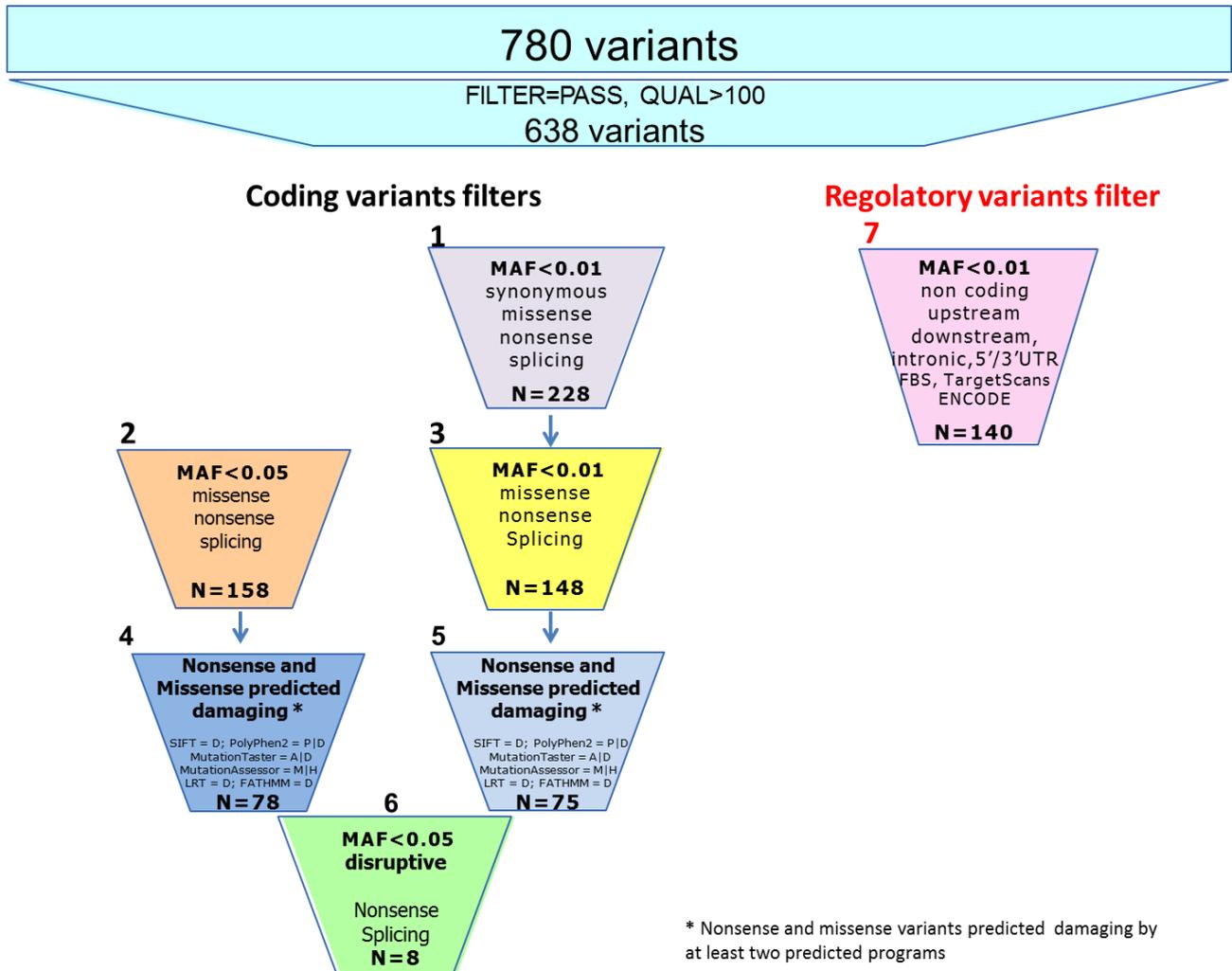
Figure 26: Description of the filters used for the burden test analysis in the replication phase. Filters 1,2, 3,4,5, and 6 were used for coding variants; filter 7 for noncoding variants.

We performed NGS target resequencing for the coding regions of the 17 genes selected from the discovery phase, in 504 MS patients and 504 controls pooled in group of 12 individuals, as in the discovery phase. Paired end multiplexed sequencing was performed on the Illumina NextSeq 500 (Illumina San Diego) platform, producing 2x150 bp read length and we analysed our data like in the discovery phase (as described in materials and methods). The MS cohort was selected regardless the genetic risk score of common susceptibility variants and was enriched in patients with positive family history of MS: in particular, for the Novara patients 108 were familial MS and 144 non-familial patients. Among the non-familial MS cases and the healthy controls, we selected individuals which have previously been analyzed with SNP genotyping platforms (Immunochip, Replication chip) (Beecham et al., 2013) (IMSGC, 2017) in order to avail of genotyping data at single individual level that will be used, during the quality control analysis, to compare the pooled allelic frequencies with

the real ones. The mean coverage obtained in total was 98,87%, with a total target of 430,870 bps. From the NGS analysis we obtained 780 variants, and after quality controls (call quality>100, filter=pass), we obtained 638 variants: 555 of them with MAF <0.05 (152 nonsynonymous, 4 nonsense, 2 splicing, 85 synonymous, 118 UTR3'/5', 172 intronic, 17 upstream+downstream, 2 ncRNA_intronic, 3 intergenic).

The significant data observed in the discovery phase was replicated after meta-analysis of the two cohorts (discovery and replication) for *TUBD1* (Tubulin Delta 1) and *MYC* (MYC Proto-Oncogene) for filters involving coding variants, for *NPEPPS* (Aminopeptidase Puromycin Sensitive) for the filter involving regulatory variants and for *EFCAB13* genes if we consider the "disruptive" filter (filter 6: stop-gain, stop-loss, splicing variants) (Table 4 and figures 27-28). In particular, for *MYC* filters 2 and 4 (both containing low-frequency missense variants) were significant both in the discovery (with c-alpha test) and in the replication and meta-analysis cohorts (with WSS test); for *TUBD1* gene all filters containing rare or low frequency missense variants were significant after meta-analysis.

			Cohort		
			p-value Discovery	p-value Replication	p-value Meta-analysis
MYC	Filter 2*	WSS	0.5549	0.0043	0.0392
		C-Alpha	0.0101	0.901	0.4411
	Filter 4*	WSS	0.0583	0.0259	0.0065
		C-Alpha	0.0121	0.8944	0.2389
TUBD1	Filter 2*	WSS	0.0919	0.1022	0.0331
		C-Alpha	0.143	0,2403	0.1051
	Filter 4*	WSS	0.2311	0.0343	0,0353
		C-Alpha	0.5059	0.2403	0.5448
	Filter 3*	WSS	0.0919	0.1022	0.0331
		C-Alpha	0.143	0.2403	0.1051
	Filter 5*	WSS	0.2311	0.0343	0.0353
		C-Alpha	0.5059	0.2403	0.3127
NPEPPS	Filter 7*	WSS	0.0213	0.4152	0.0565
		C-Alpha	0.0039	0.2553	0.0095
EFCAB13	Disruptive variants filter*	WSS	0.0001	0.3292	0.0016
		C-Alpha	0.0023	0.0062	8.9e-5

Table 4: Meta-analysis Between Discovery and Replication cohorts which displays the results for the 3 genes that showed a significant burden with at least 1 of the filters after meta-analysis. In particular, this analysis used 2 tests: Weighted-Sum Statistic, and C-alpha.

* Filter 2: MAF<0.05, Missense, nonsense, splicing; Filter 4: MAF<0.05, nonsense and damaging missense; Filter 3: MAF<0.01, Missense, nonsense, splicing; Filter 5: MAF<0.01, nonsense and damaging missense; Filter 7: variants with a possible regulatory role with a MAF <0.01; Disruptive variants filter: stop-gain, stop-loss, splicing variants.
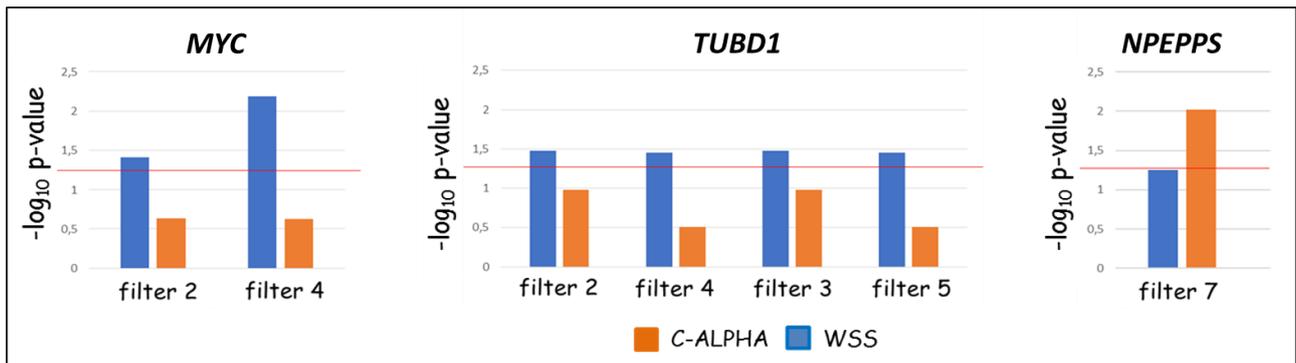
Figure 27: Meta-analysis between discovery and replication cohorts: results for the 3 of 4 genes that showed a significant burden with at least 1 of the filters after meta-analysis. The bar-plots show the –log10 of the p-value obtained with 2 tests (Weighted-Sum Statistic, blue and C-alpha, orange). The red line corresponds to p=0.05.
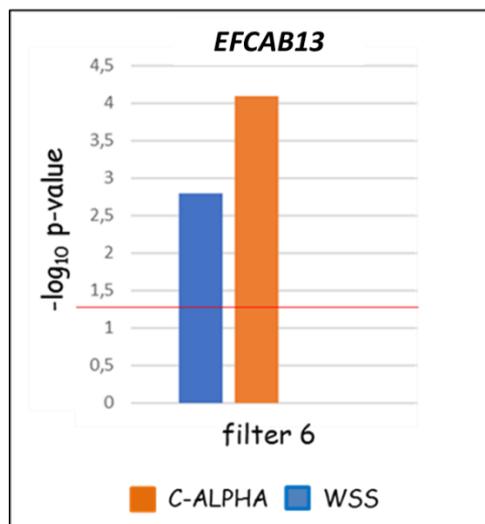
Figure 28: Meta-analysis between discovery and replication cohorts: results for the disruptive variants in *EFCAB13*. The red line corresponds to p=0.05.

Again, the strongest result was observed for *EFCAB13,* which is also the more enriched in disruptive variants in both sequencing experiments (discovery and replication): in fact, there is an overlap of 6 disruptive variants in *EFCAB13* between the discovery and the replication: 5 stop-gain variants, 1 splice acceptor variant. In addition to these 6 variants, in the discovery we found 2 additional disruptive variants in *EFCAB13*: 1 splice acceptor and 1 splice donor variant (figure 29). Among the two NGS experiments, 4 variants (SNV 2, 4, 5, 7) showed a similar trend in the distribution of allele counts between MS patients and controls (figure 30).
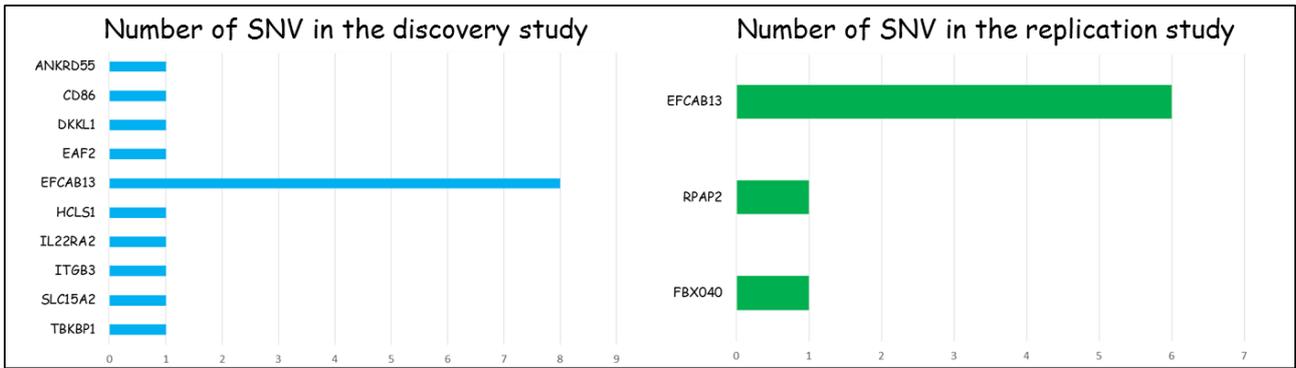
Figure 29: Number of disruptive variants (stopgain, stoploss, splicing) in discovery and replication cohorts.
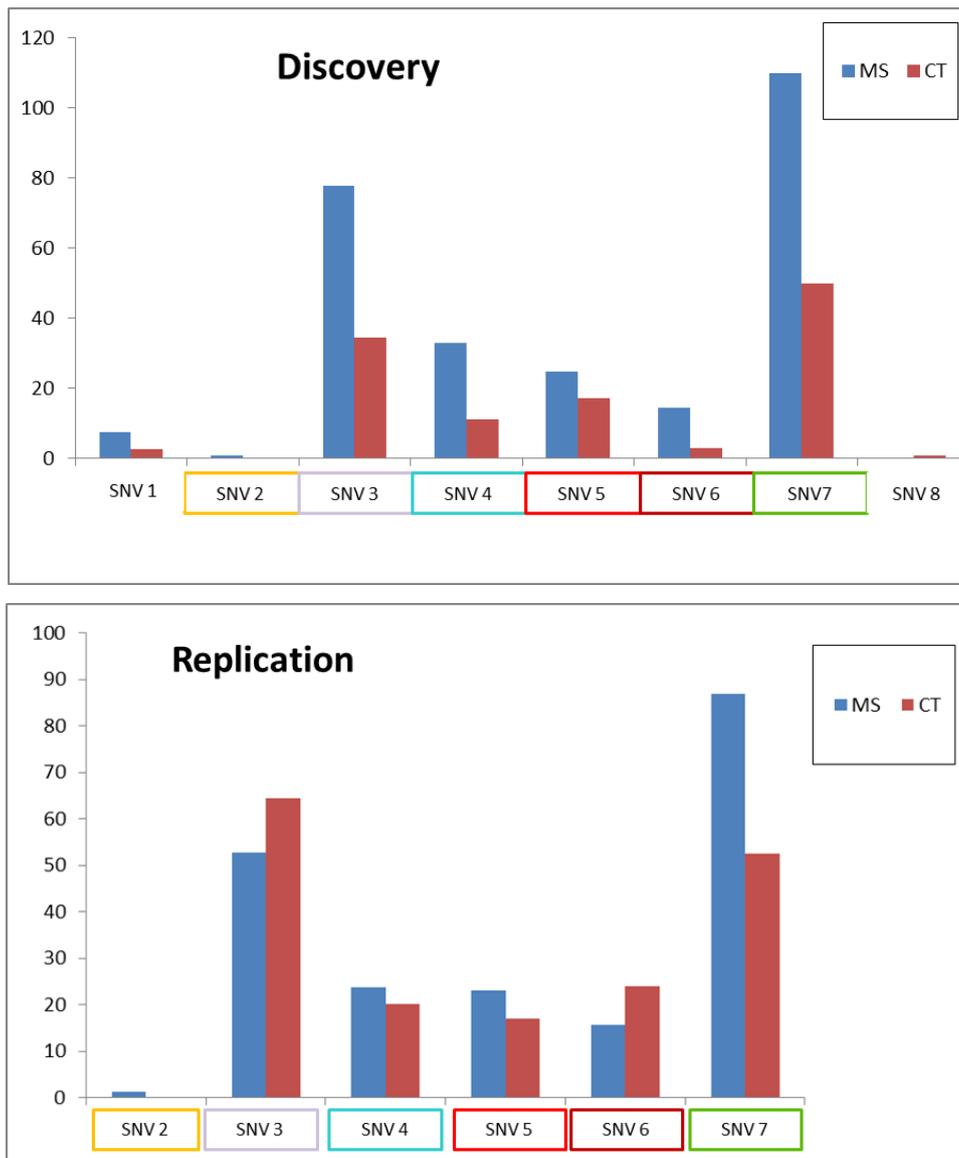
Figure 30: Distribution of minor allele counts for disruptive variants in *EFCAB13* in discovery and replication cohorts.

5.4 Discussion and conclusions

In these years Multiple Sclerosis research is focused on the research of susceptibility genes and variants. In particular the discovery of rare variants associated with Multiple Sclerosis is still controversial, in fact it is difficult to identify these signals on one side because of the need of very huge cohorts, on the other side because it is difficult to discriminate these rare variants with the available techniques. Therefore, the research of rare susceptibility variants in MS susceptibility is still not so deeply studied. Recently, a paper published from International Multiple Sclerosis Genetics Consortium on Cell journal (IMSGC, 2018), analyzing 120,991 low-frequency non-synonymous coding variants in 68,379 cases and controls, identified 7 low-frequency variants in 6 genes outside the HLA region. Two of these variants were in regions identified by MS GWAS and showed linkage disequilibrium with the common-variant associations previously reported (Sawcer et al., 2011), while the remaining signals were novel and did not show linkage disequilibrium to common variant association signals in GWASs. This work concludes that nearly 5% of heritability is explained by coding low-frequency variants and that more low-frequency and rare-variant associations remain to be discovered and it will be necessary larger sample sizes to increase statistical power. The laboratory of Human Genetic of Health Science Department is involved in the research of rare genetic variants involved in MS susceptibility using cumulative evaluation of the different distribution in MS patients and healthy controls of all the rare and low frequency variants in genes mapping in MS associated regions (Burden Test analysis). This approach was performed through a Next Generation Sequencing (NGS) analysis on a high number of pooled patients and controls and aimed to identify possible genes and so possible variants involved in the MS susceptibility. The Next Generation Sequencing on pools of individuals technique is a method that has been developed in the recent years to analyze extended portion of the genome on an elevated number of individuals with reduced costs compared to the classic analysis of sequencing on single individuals (Christian Schlötter, 2014). For example, different studies of genic associations and Burden Test analysis with Next generation sequencing on pools had been effectuated: one study on Rheumatoid Arthritis by Diogo et al (Dorothée Diogo et al, 2013) in which 10 pools of patients and 13 of healthy controls (each of 50 individuals) were used. Moreover, it is possible to account also the study of Yohei Kirino et al (Yohei Kirino et al, 2013) on Behçet disease and the study of Hongsheng Gui et al (Hongsheng Gui et al, 2014) on the Hirschsprung disease. However, pooling of DNA creates new problems and challenges for accurate variant call and allele frequency (AF) estimation. In particular, sequencing errors confound with the alleles present at low frequency in the pools possibly giving rise to false positive variants. Thus, it is necessary to remove sequencing errors during the analysis phase to obtain an accurate AF estimation as described

by Anand S. (Anand S. et al, 2016). During our discovery phase, a specific pipeline was developed and it was decided, after quality controls and quality analysis (comparison of pooled sequencing data with single individual level genotype data), to apply a quality filter (Quality filter ≥100) and a threshold (2.6%) on the alternative (ALT) allele frequency observed in the single pools: this value of threshold allowed to eliminate from data the possible variants due to artificial problem of the NGS technique, and to maintain the variants that can be with more probability true. In the discovery phase we examined the data from the Burden Test analysis with 9 different filters with the use of three statistical programs: in particular 2 filters has been used as a negative control filter to verify if the significant burden observed in the discovery filters was actually due to the effect of functional variations or to a general unbalance of all rare variants of the region between patients and controls. In total 17 genes resulted statistically significant and in particular *EFCAB13, MYC, SKAP2, TGFBR3* appears in more than one filter for variants selection. Seen the results of the Discovery phase, we selected these genes for the Replication in the new cohort of patients and controls. In the Replication phase we found 3 genes (*MYC, TUBD1* and *EFCAB13*) that showed a significant burden in at least one of the 6 filters for coding variants with at least one of the statistic programs. The meta-analysis between the two studies confirmed a significant burden with one of the statistical test for *MYC* and *TUBD1* genes for filters involving coding variants, *NPEPPS* for the filter involving regulatory variants and for *EFCAB13* for the "disruptive" filter (filter 6: stop-gain, stop-loss, splicing variants). In particular *EFCAB13* seemed to show the most promising result. It was the only gene resulted to have a statistically significant p-value with 2 statistic test (C-ALPHA and WSS) both in discovery and replication study and in meta-analysis among the two data sets. Among the two studies, it is the gene most enriched in disruptive variants (totaling 8 between discovery and replication). Six of these variants were observed both in discovery and in replication subset, and among them, 4 variants (3 stop-gained and 1 splice acceptor) showed a concordant trend for minor allele count between patients and controls. This gene encodes for two main transcripts: a transcript variant A of 117,351 bp (NM_152347) which leads to the formation of a protein of 973 aa, and a transcript variant B of 106,750 bp (NM_001195192) lacking 3 of the 25 exons, which leads to the formation of a protein of 784 aa. The function of EFCAB13 protein is still unknown, however this protein has 6 different EF-hand domains, that are helix-loop-helix structural domains that usually bind calcium ions. The observed 8 disruptive variants could impact the gene function at various levels: 5 stop-gained (a variation falls into a EF-hand 2 domain (524-559 aa) which is a structural domain involved in calcium ions binding) some of them isoform-specific (2/5 only for A isoform), two splicing acceptor variants (one not present in public databases) and one splicing donor. *EFCAB13* is a not deeply investigated gene, so for the future we will plan to study in vitro the function of these variants on RNA expression

and protein production, in order to find a possible explanation about the role of this gene in MS pathogenesis.

6. Chapter 4: Genomic and functional analysis on TNFSF14-TNFRSF14 pathways

6.1 Introduction

Until now, our effort allowed us to identify the primary associated variant in *TNFSF14* gene in the susceptibility to Multiple Sclerosis. Furthermore, we propose a possible functional role of this variant in the gene regulation and protein production. Our second task was to pose our attention on possible interactors of this gene. *TNFSF14* gene encodes for LIGHT, a protein that binds 2 different receptors: HVEM (herpes virus entry mediator) on T lymphocytes and natural killer cells and LTβR (lymphotoxinβ receptor) on stromal cells and monocytes. In previous international studies, different SNPs in the region of the gene that encodes for HVEM (*TNFRSF14* gene) were reported as associated to MS, in details: rs4648356 in IMSGC-WTCCC2 GWAS (Sawcer et al., 2011) (p = 1E-14, OR= 1.14), rs3748817 in the Immunochip project (IC) (Beecham et al., 2013) (p =1.33E-12, OR=1.14) and rs6670198 in MS replication chip (MS chip, the follow-up of the Immunochip project and the IMSGC-WTCCC2 GWAS) (IMSGC, 2017) (p=2,09E-13, OR= 1.15). These variants map in neighboring genes (*MMEL1* and *FAM213B*) within a region with high LD (r^2>0.8, chr1:2473821-2711009, SNAP tool Broad institute, data from 1000 genomes) including the *TNFRSF14* gene. These variants were associated also in the Italian population (IC: 961 MS, 962 HC; rs3748817: p=1.73E-03, OR= 1.28; MS chip: 941 MS, 950 HC, rs6670198: p=0.001869, OR=1.27). Paralleling what observed for *TNFRSF14* locus, other MS associated variants fall in genetic regions containing genes coding for other interactors of TNFSF14/TNFRSF14 or other protein part of TNFSF14 pathway, such as *LTβR* and *STAT3*. Although these genes are confirmed MS-associated loci, the genetic variants primarily responsible of the association and the disease molecular mechanisms controlled by these genes are unknown. As reported in literature, it is known that immunological response plays a key role in MS pathogenesis and given the role of *TNFSF14* in immunological process, it is possible that also the molecules that interact with LIGHT could take part in the MS etiopathogenesis. The receptor of LIGHT, HVEM plays important roles in the immune system such as T-cell co-stimulation, regulation of dendritic cells (DC) homeostasis, autoimmune-mediated inflammatory responses, as well as host defense against pathogens. Based on these assumptions, the second aim of our project was to focus our attention on *TNFRSF14* locus since the product of these gene (HVEM) is known the be the main interactor of LIGHT. We wanted to better investigate the association signals in this region

in order to find the primary associated variant through a sequencing and a fine-mapping approach, following the same strategy previously applied for the *TNFSF14* locus.

Parallel to this, we aimed to expand the analysis to other possible interactors of *TNFSF14* following different strategies:

a) we analyzed rare variants in the genes of TNFSF14 pathway constructing a series of burden tests;

b) we performed a gene-gene interaction analysis on couples of interacting genes including TNFSF14 or TNFRSF14;

c) we built a weight genetic risk score on interactors of *TNFSF14* and *TNFRSF14*.

6.2 Sequencing of *TNFRSF14* region
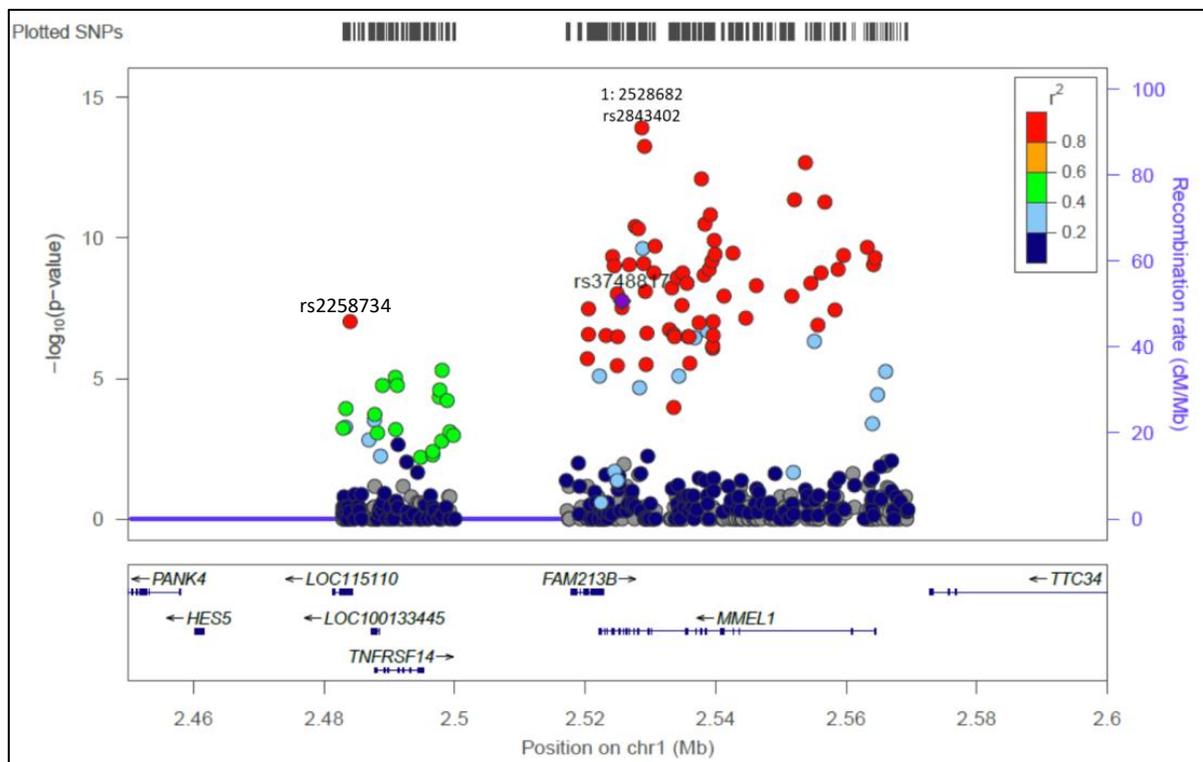
Figure 31: Regional association plot from resequencing data on TNFRSF14 region.

The *TNFRSF14* gene maps in a region characterized by high linkage disequilibrium (LD) spanning for about 100 kb and containing 5 genes: *MMEL1*, *FAM213B*, *TNFRSF14*, *LOC115110* and *LOC100133445*. We sequenced the whole region of 100 kb containing *TNFRSF14* gene by NGS

method in 588 MS and 408 HC pooled in groups of 12 individuals, following the same strategy used for *TNFSF14* gene (see materials and methods). After quality controls, we observed 491 variants in *TNFRSF14* gene region. Among these, 64 were in the coding regions and 143 with a MAF >0.01. Significant association (p<0.05) was observed for 123 variations (figure 31). The best hit (rs2843402, p=1.23e-14, r2=1 with rs3748817, r2=0.96 with rs6670198) was an intronic variant in the *MMEL1* gene. In details, we found 141 variants in a region that includes *TNFRSF14* gene and two long non-coding genes (*LOC115110* and *LOC100133445*) which partially overlap with it. In this region 27 variants were statistically associated (p<0.05) and the best hit was rs2258734 (p=9.37e-08, r2=0.82 with rs3748817), an exonic variant of the long non-coding gene *LOC115110,* mapping in the promoter of *TNFRSF14*.

The next step was to perform a conditional analysis of this region in order to identify the primarily associated variant. To perform this analysis, we availed of 4 independent datasets (listed in table 6), totaling 3314 MS and 3272 healthy controls (HC) of Italian origin: IMSGC GWAS (IG), Immunochip (IC), MSchip (MSC), all 3 available thanks to the involvement in international IMSGC projects (Beecham et al., 2013; IMSGC, 2017; Sawcer et al., 2011), and OminiExpress GWAS (OE) provided by prof. Filippo Martinelli Boneschi and Federica Esposito group of San Raffaele Hospital (HSR).

Sample set	Abbreviation	MS	HC
IMSGC GWAS	IG	734	1250
Immunochip	IC	961	962
MSchip	MSC	938	952
OmniExpress GWAS	OE	1269	360
Total sample set		3314	3272

Table 6: list of data sets used for the conditional analysis on Italian population.

IC and MSC platforms have a deep coverage for the *TNFRSF14* region, while the other 2 datasets were imputed with the "1000Genome ALL" reference panel using MACH (Li et al., 2010) and Minimac (Fuchsberger et al., 2015). The most associated SNPs within *TNFRSF14* region were determined in the large international sample set, then conditional association analysis was performed with PLINK software (Purcell et al., 2007), using logistic regression covariated for sex. Analysis was performed on each sample set separately, then the 4 datasets were meta-analysed to increase statistic

power. There were 102 SNPs in the MSC, 154 SNPs in the IC, 253 in the IG and 108 in the OE platforms mapping in this region. After meta-analysis of the 4 sample sets (66 SNPs) we observed a strong association (p<10-7) for a cluster of variants in high LD, located mainly in *MMEL1* gene and including rs3748817. The best associated SNP (p=1e-15) in this region in the international MSchip dataset was an intronic variant within *MMEL1* gene (rs3748817), which was also significantly associated in our cohort (p=6.27e-8). After conditioning for this variant, we still observed nominal significance (p=0.003) for 2 intronic SNPs in *MMEL1* (rs72646016, rs12138909) in complete LD (r2=1) that may underlie an independent signal (figure 32).
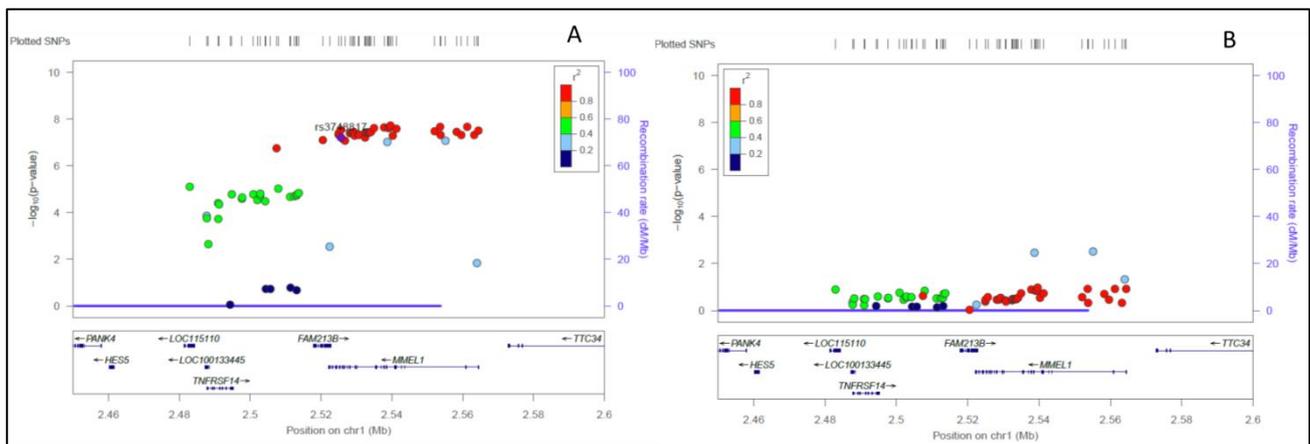
Figure 32: Regional association plot from meta-analysis before (A) and after (B) the conditioning for rs3748817.

Our data confirmed that association in this region is mainly driven by some marker in *MMEL1* gene, part of a LD cluster that comprise at least 97 SNPs (among the variants reported in the public database 1000 genomes as reported in figure 33, there are 96 SNPs with r2 > 0.8 with rs3748817). Therefore, the genetic association analysis alone is not able to distinguish which of them is the functional variation.
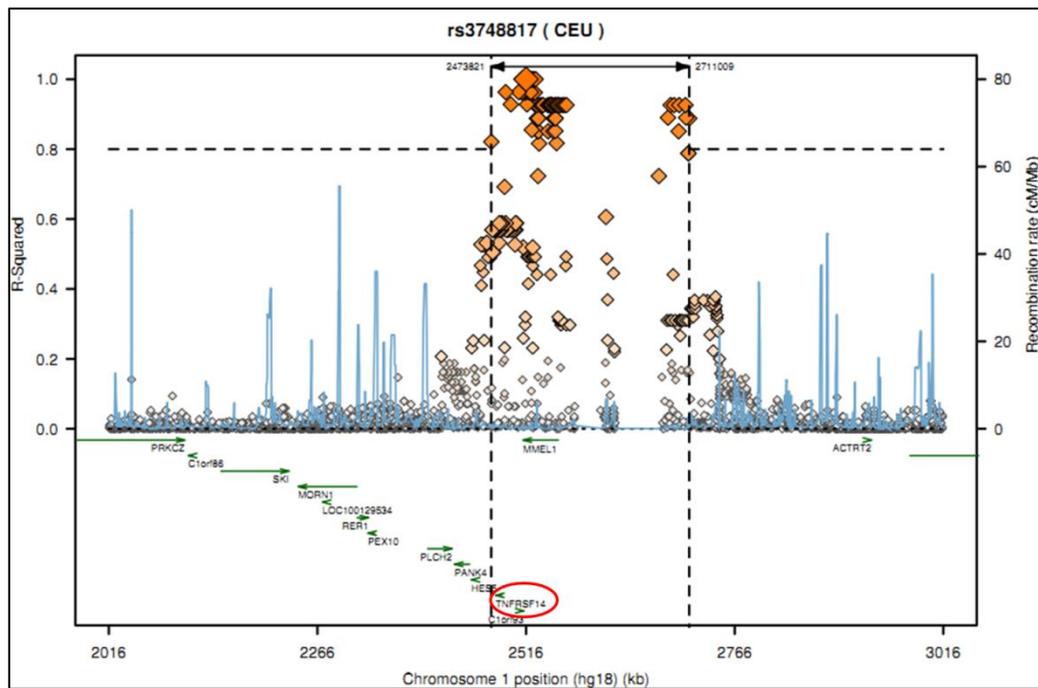
Figure 33: The LD within the TNFRSF14 region by SNAP proxy tool of Broad Institute.

So, we tried to prioritize these variants using different in silico approaches and prediction tools aimed to identify putative effects at the expression and/or protein levels. Moreover, we performed a haplotype analysis on two datasets to better define the association signal in this region.

6.3 Haplotype analysis on *TNFRSF14* gene region.

In parallel with these analysis, we conducted a haplotype analysis of TNFRSF14 gene region in order to find a possible risk haplotype that could be associated to the disease more than the single variants in this region. We performed this analysis both on MSchip (MSC) and Immunochip (IC) sample sets. Haplotype blocks were defined using two parallel different strategies and then we performed association analysis on all haplotypes showing a frequency > 0.01. For one strategy, we defined haplotype blocks according to Gabriel rule using Haploview software (Barrett et al., 2005). Due to high LD, only 1 block, was generated for both platforms. In details the program recognized 15 haplotypes with frequency < 0.01 in IC and 6 in MSC. The second strategy was based on analysis of all haplotypes generated by sliding windows of 3, 4, 5 and 6 SNPs by PLINK software (Purcell et al., 2007). Overall, 390 windows were analyzed for MSC and 602 for IC. We observed only 10 haplotypes, in MSC subset, that retained statistical significance when Bonferroni correction was applied (correction for the number of SNPs in this region), and they all contain the same core of three SNPs. The p-value for the haplotype (p=0.0005) is more significant than those of the single variants

comprising it (p≥0.030). The three variants are not significant in the IC sample set, so we are not powered to replicate this association in IC platform. None of the variants retains significance after conditioning for the best SNP (rs3748817), so we can conclude that this signal is probably secondary to the main one.

6.4 Annotation of functional elements in *TNFRSF14* gene region.
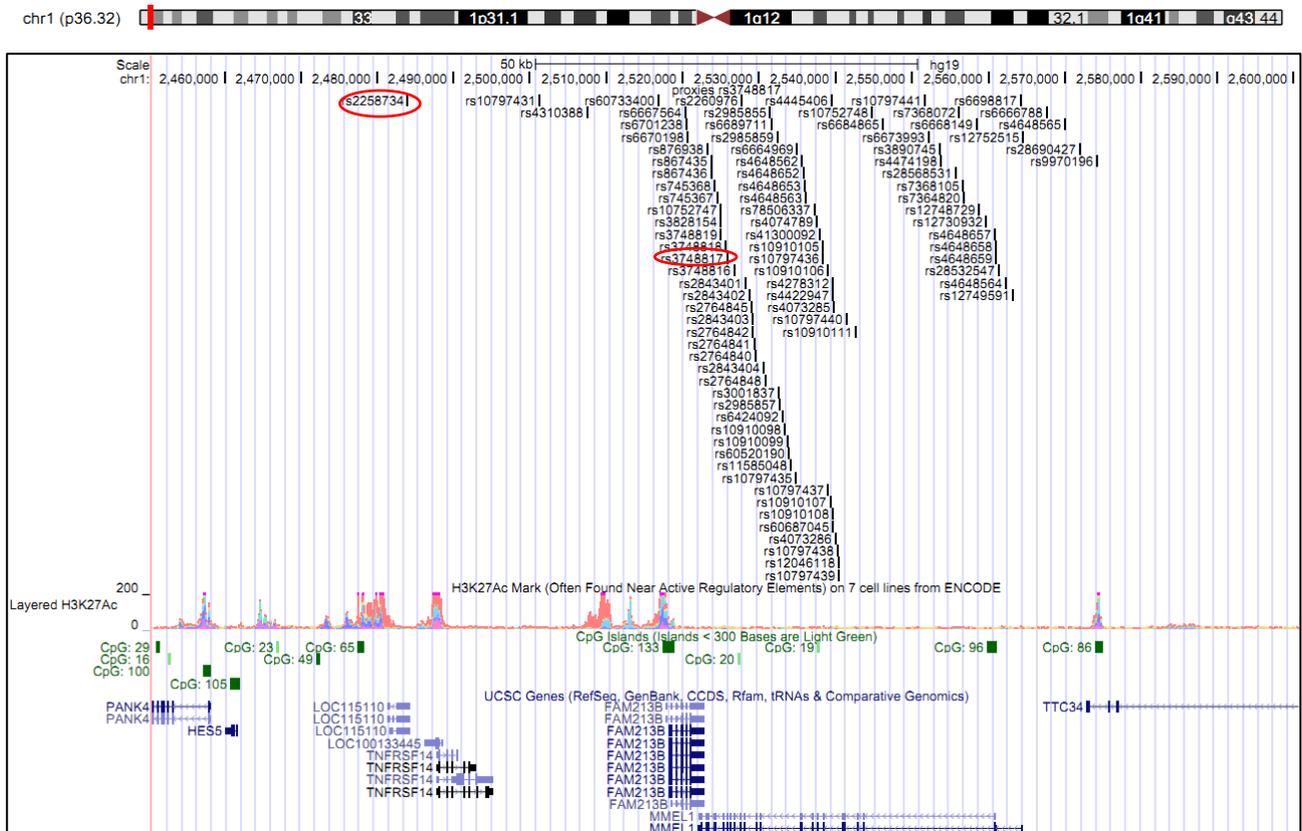
Figure 34: UCSC track of 97 SNPs of the TNFRSF14 LD block. The best associated SNP in MMEL1 (rs3748817) and in TNFRSF14 promoter (rs2258734) are displayed within a circle.

We have evaluated if the 97 SNPs of the *TNFRSF14* LD block fall in functionally relevant regions, such as coding regions, splicing sites, UTRs or promoters using the annotation software wANNOVAR (web ANNOVAR) (Wang et al., 2010) and the tracks in the UCSC genome browser (figure 34). We found that 79 of these SNPs are intronic variants, and most of them (72) map within *MMEL1* gene. There are 5 exonic variations, all mapping in the *MMEL1* gene: 4 synonymous and one missense variation (M518T, rs3748816, r2=0.928). This variant is predicted as benign according to 6 prediction algorithms (SIFT; PolyPhen2; MutationTaster; MutationAssessor; LRT; FATHMM).

We did not observe coding variants with a clear impact on protein function such as nonsense (stop-gain, stop-loss), splicing variants or missense annotated as damaging. So, we rather focused on the possible regulative effect of these variations.

One SNP (rs2258734) (r2=1.00 with the best signal, rs3748817) maps in the promoter of *TNFRSF14* gene and within a long non-coding RNA (*LOC115110* also named as *TNFRSF14-AS1*) annotated as a transcript partially overlapping the *TNFRSF14* gene. There are 2 intergenic SNPs (rs4310388, r2=0.928 and rs60733400, r2= 0.963) falling in a peak of high acetylation of the histone H3 (H3K27AC mark), correlating with an open conformation of the chromatin, thus with active transcription. Furthermore, one intergenic SNP (rs60733400) binds several transcription factors, including *ELF1* in Jurkat T cells (data from RegulomeDB). Several SNPs fall in CpG islands, in particular three intronic variants: rs9970196 (r2=0.926) mapping in the *TTC34* gene, and rs4648657 (r2=0.852) and rs4130092 (r2=0.926) both mapping in *MMEL1*, and two synonymous variants (rs4648658 (r2=0.888, A14A) and rs4648659 (r2=0.888, P7P)) in exon 2 of *MMEL1* gene.

6.5 SNP prioritization

We performed a prioritization analysis using three tools: RegulomeDB (Boyle et al., 2012), SPOT (SNP Prioritization Online Tool) (Saccone et al., 2008) and Variant Ranker (Alexander et al., 2017). RegulomeDB is a database that annotates SNPs within known and predicted regulatory elements in the intergenic and promoter regions, including DNAase hypersensitivity regions, transcription factors binding sites (from TRANSFAC and JASPAR CORE databases), and promoter regions that have been biochemically characterized for transcription regulation. Source of these data include public datasets from GEO, the ENCODE project, and published literature. Thanks to this analysis we found the highest score (1f) for two intronic variants in *MMEL1* gene (rs10910099, rs6684865), followed by 4 variants with a score of 2b (in details: intronic SNP rs10910111in *MMEL1*, intergenic SNP rs4648565, synonymous SNP rs4648659 in *MMEL1*, intergenic SNP rs2100574).

SPOT is a tool based on the Genomic Information Network (GIN) method, designed to provide a systematic method for incorporating specific biological hypotheses into the design of a genetic association study. The highest score was observed for an intergenic SNP (rs897628), followed by 4 exonic variants in *MMEL1* (rs4648659, rs4648658, rs4648562, rs10797440).

Variant Ranker is a tool that implements and aggregates multiple prediction algorithm scores, conservation scores, allelic frequencies and clinical informations. The highest scores were reported for 3 exonic variants in *MMEL1* gene (rs4648562, rs3748816, rs4648659).

The variant in the *TNFRSF14* promoter (rs2258734), the SNP reported as best signal in the international dataset (rs3748817) and the 2 variants that maintain significance after conditioning for the best SNP (rs72646016, rs12138909) had medium-low priority scores with all the tools. We did not observe a SNP or a group of SNPs with high prediction scores with all used tools, so we considered conclusions from this analysis to be inconsistent.

6.6 eQTL data from online databases

To predict the biological relevance of SNPs of the *TNFRSF14* LD block and hence to identify the most likely causal variant, we performed an in silico eQTL analysis merging multi omics data derived from several public databases like Gtex (Melé et al., 2015), Geuvadis (Lappalainen et al., 2013) and eQTL Blood Browser (Westra et al., 2013). Expression Quantitative Trait Loci (eQTL) analysis performed by Gtex portal on 44 tissues for 449 healthy donors have showed that there are 3 genes (*MMEL1*, *FAM213B* and *TNFRSF14*) in the region with a significant eQTL with the best associated SNP (rs3748817). In details for *MMEL1*, we observe that the MS risk allele (T) is associated with a statistically significant increase in gene expression in 18 different tissues, particularly whole blood (p=1,4e-13, beta= 0,496), and brain-nucleus accumbens (basal ganglia) (p=3,2 e-12, beta= 0,71). The same trend was observed for *FAM213B* in 16 tissues, particularly whole blood (p=5,2e-7, beta= 0,217), and brain-caudate (basal ganglia) (p=5,1 e-4, beta = 0,3). In contrast, there is an opposite trend for *TNFRSF14* where the MS risk allele correlates with a statistically significant decrease in gene expression especially in brain-cerebellum (p-value=2,2e-04, beta=-0,4) and in EBV- transformed lymphocytes (p=2,3e-3, beta=-0,15). Similar results have been observed for further two SNPs (rs6670198 and rs2258734), interestingly, one maps in the *TNFRSF14* promoter. Consistent data were observed performing eQTL analysis on EBV transformed cell lines with other two cohorts: Geuvadis consortium (RNA sequencing data of 465 lymphoblastoid cell lines from 5 populations of the 1000 genomes project), and Blood eQTL browser (expression quantitative trait locus (eQTL) meta-analysis in peripheral blood samples from 5,311 individuals) where we observed that the presence of the risk alleles (T) of rs3748817 was correlated to a lower *TNFRSF14* expression in both datasets (Geuvadis data set: p = 3,03e-06; Blood eQTL browser p = 1,534e-13). Altogether, these data suggest that at least three SNPs within the *TNFRSF14* LD bock are potentially involved in the regulation of the *TNFRSF14* gene expression. In particular, one SNP maps in the *TNFRSF14* promoter. Another interesting finding is the identification of a long non-coding RNA (*TNFRSF14-AS1*) annotated as a transcript partially overlapping the *TNFRSF14* gene, containing one SNP of the

LD block. The three identified SNPs might be potentially involved also in the expression of this long non-coding RNA.

6.7 Functional analysis on HVEM protein.

We analyzed HVEM expression by flow cytometry in several types of immune cells in the peripheral blood: CD8+T cells, CD4+T cells, B cells (CD19+), myeloid dendritic cells (mDC Lin-/HLA-DR+/CD11c+), three subsets of NK cells (CD56dim/CD16bright, CD56-/CD16bright, CD56dim/CD16-) and monocytes CD14+. These flow cytometry analyses were performed thanks to the cooperation with the Immunology laboratory headed by prof. Umberto Dianziani at University of Eastern Piedmont, department of Health Sciences, in Novara. We collected the blood from 22 healthy controls: 10 homozygous for the risk allele (T) of rs3748817, 10 heterozygous and 2 homozygous for the protective allele (C). The mean percentage of HVEM positive cells among the analyzed populations was found to range between 12.69% (NK 56-/16bright cells) and 45.87% (NK 56 dim/16 bright). No difference was observed according to the genotype of the associated variant (rs3748817). We compared the expression of HVEM in myeloid derived dendritic cells (MDDC) obtained by culturing monocytes for 5 days with GM-CSF+IL-4 or GM-CSF+IFNβ or IL-3 alone or IL-3+IFNβ or GM-CSF+IL-15, which are different MDDC types described in the literature. Results showed that HVEM expression was quite different in the different types of MDDC: it was low in immature MDDC and upregulated in mature MDDC using MDDC derived with GM-CSF+IL-4 or GM-CSF+IFNβ or IL-3+IFNβ, whereas it was high in immature MDDC and downregulated in mature MDDC using MDDC derived with IL-3 alone or GM-CSF+IL-15 (Figure 35).
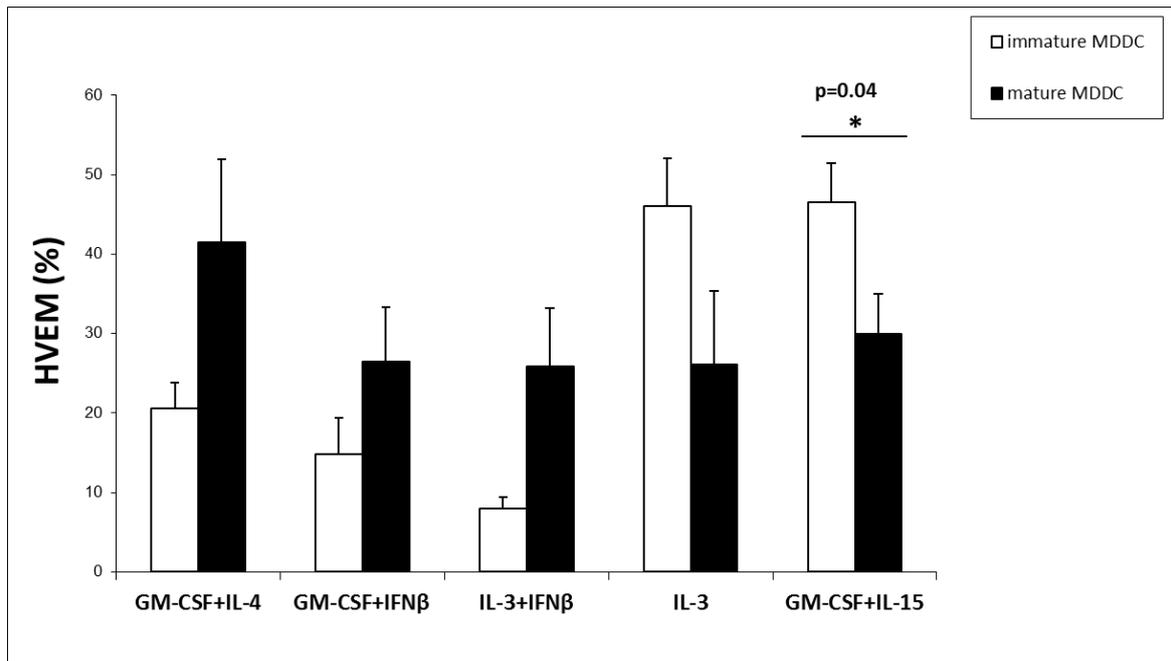
Figure 35: Mean percentage (+SE) of HVEM positive immature and mature (LPS-activated) monocyte-derived DC (MDDC) obtained by culturing monocytes with GM-CSF+IL-4, GM-CSF+IFNβ, IL-3, IL-3+IFN β, GM-CSF+IL-15 detected by flow cytometry, from 5 healthy donors.

As previously seen for LIGHT, we also focused our attention toward MDDC obtained with GM-CSF+IL-15 since express high levels of surface HVEM. The same analysis in this dendritic population was performed for HVEM in 21 HC (8 TT, 10 TC, 3 CC): we confirmed a down modulation of this protein after maturation (paired T-test: p= 0.0007) as previously observed, while no difference was found according to the genotype of the associated variant (rs3748817) (figure 36).
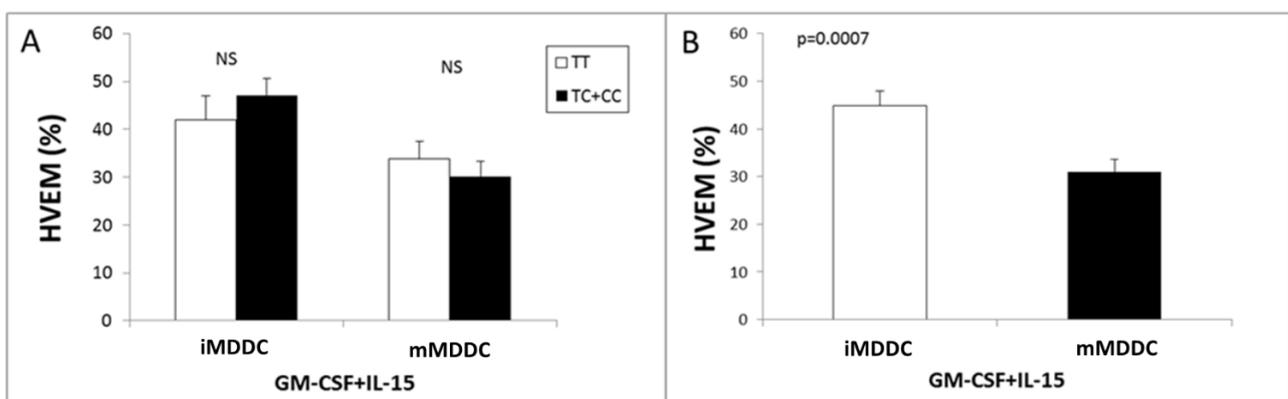
Figure 36: Mean percentage (+SE) of HVEM positive immature (iMDDC) and mature (mMDDC, LPS-activated), monocyte-derived DC (MDDC) (detected by flow cytometry) obtained by culturing monocytes with GM-CSF+IL-15 from 21 healthy donors.

6.8 TNFSF14-TNFRSF14 pathways analysis

As a secondary task, we investigated the other genes belonging to the TNFSF14-TNFRSF14 pathway with three aims:

1) we performed sequencing analysis in these genes to investigate the role of rare (MAF < 1%) and low frequency (MAF 1-5%) variants by burden test analysis with the aim to analyze the cumulative burden of rare functional variations of the genes belonging to TNFSF14 pathway. It is in fact possible that MS patients present an enrichment in rare variants for several genes involved in the same functional pathway.

2) We conducted a gene-gene interaction analysis for pairwise interacting SNPs, testing interactions between SNPs located in TNFSF14 or TNFRSF14 genes and 370 in silico selected interactors, with the aim to search for strong associations with interactions within TNFSF14 pathway.

3) We constructed a weighted Genetic Risk Score (wGRS) using MS susceptibility variants located in chromosome regions containing interactors of TNFSF14/TNFRSF14, in order to define the cumulative effect of common susceptibility variants within this pathway.

These analyses were performed thanks to the contribution of the laboratory of Human Genetics of Neurological Disorders at San Raffaele Hospital headed by professor Filippo Martinelli Boneschi and Federica Esposito.

6.8.1 Burden test on *TNFSF14-TNFRSF14* pathway

We identified on the basis of literature data and of 4 protein-protein interaction analysis algorithms (PINA (Protein Interaction Network Analysis) v2 (Cowley et al., 2012), STRING v10(Jensen et al., 2009), GPS-Prot (Fahey et al., 2011) and IID (Integrated Interaction Database) (Kotlyar et al., 2016) 370 genes in total belonging to the TNFSF14-TNFRSF14 pathway. 31 of them (*LTBR, TRAF3, GRB2, MAP3K14, TNFSF14, TNFRSF14, STAT3, APP, LTB, TRAF2, BIRC2, TNFRSF6B, DIABLO, BTLA, GDI2, SLC25A22, NDUFS2, ZBTB48, PFDN2, CDC37, RUVBL2, DRAP1, PSPH, EIF3I, PSMB1, TNFSF13, PTPN11, RAD21, HSPE1, EIF3E* and *ATXN10*), were prioritized for sequencing analysis on the basis of the following criteria:

-the interaction was based on or supported by experimental data,

-the gene was located within 500 kb upstream/downstream a SNP showing a genome-wide significance level association (p<5*10-8) with MS in the MSchip international cohort,

-the gene was located within 500 kb upstream/downstream a SNP showing a suggestive association (p<1*10-5) with MS in the MSchip international cohort.

The sequencing analysis was performed in 504 MS patients and 504 healthy controls pooled in group of 12 individuals on the Illumina NextSeq 500 (Illumina San Diego) (as described in materials and methods and following the same strategy adopted for rare variants analysis as described in chapter 2). For the NGS analysis we sequenced only the coding regions (+/-50 bp from the exons) and after quality controls, we obtained 657 variants, from which 131 non-synonymous, 147 synonymous, 307 intronic, 3 ncRNA exonic, 5 ncRNA intronic, 1 splicing, 41 UTR3'/5').

Our laboratory in collaboration with a group in San Raffaele institute, leaded by Prof. Filippo Martinelli-Boneschi and dr. Federica Esposito have already conducted an analysis on rare variants involved in the susceptibility to MS (as seen in chapter 2).

For the current task, we defined new criteria for variants filtering compared to the analysis described in chapter 2. In fact, thanks to our past experience we observed that the most consistent results were found for variants with a strong functional impact on protein production, such as disruptive variants (stop-gain, stop-loss, splicing) and r missense variants predicted as damaging. Based on these evidences, for the present analysis on genes belonging to TNFSF14 pathways we focused our attention on non-sense, splicing and missense variants predicted as damaging. Furthermore, we also investigated non-coding variants with a possible regulatory role, paralleling the same analysis already performed with the previous experiments (filter 7 in chapter 2). Compared to the annotation performed for the study described in chapter 2, the number of available predictors of missense variants effect increased (21 vs 60). Therefore, we made a literature search in order to choose the best strategy for the filtering of missense variants, and we decided to use the strategy suggested by (Ghosh et al., 2017), who used 5 predictors (Polyphen, SIFT, CADD, PROVEAN and MutationTaster). According to the authors, who tested 18 prediction algorithms, this strategy resulted in higher concordance against ClinVar variants, with 79% concordance for pathogenic variants. So, we filtered variants on the basis of the following criteria:

- Filter 1 for disruptive filter which includes all the stop-gain, stop-loss and splicing variants;
- Filter 2 which includes disruptive variants and non-synonymous variants with a MAF < 0.01 predicted as damaging with at last one of the 5 prediction programs (SIFT; Polyphen; MutationTaster; PROVEAN; CADD with a score >=20);
- Filter 3 which includes disruptive variants and non-synonymous variants with a MAF < 0.01 predicted as damaging with at last 4 of the 5 prediction programs (SIFT; Polyphen; MutationTaster; PROVEAN; CADD with a score >=20);
- Filter 4 for downstream/upstream, UTR3'/5', intronic variants with a MAF < 0.01 predicted as regulatory with the annotation based on TFBS, TargetScans, ENCODE;

- Filter 5 for synonymous variants and non-synonymous with a MAF < 0.01 predicted as benign in all the 5 predictor programs as control filter.

We found 97 variants for filter 2, 40 for filter 3, 60 for filter 4 and 484 for filter 5. We did not observe any stop-gain or stop-loss variants for the analysed genes, so we did not apply the filter for disruptive variants. For each gene, the cumulative frequency of the variants selected from each of the applied filters was compared between MS patient and control pools with three different statistical tools for burden test (WSS, C-ALPHA and Fisher Hybrid Test). We observed a significant burden for rare regulative variants (filter 4) for *EIF3E* (p=0,043 with WSS test, p= with Fisher Hybrid Test) and *RUVBL2* (p=0.008 with C-ALPHA test, p=0.037 with Fisher Hybrid Test) and for rare missense variants (filter 2) for *CDC37* (p=0.0278 with WSS test). The contribution of these results were mainly due to 2 variants in *EIF3E* (rate of minor allele counts in HC: 0.000496, rate of minor allele counts in MS: 0.00248), 2 variants in *RUVBL2* (rate of minor allele counts in HC: 0.002976, rate of minor allele counts in MS: 0.001488) and to 5 missense variants in *CDC37* that we found only in patients (rate of minor allele counts in MS: 0.000992).

6.8.2 Gene-gene interaction analysis for *TNFSF14* and *TNFRSF14* interactors.

We adopted a "candidate-interactions" strategy, leveraging information from freely available protein-protein interaction (PPI) and pathways resources:

a) selecting physical and validated interactors of genes TNFSF14 and TNFRSF14 from various PPI resources (STRING, Reactome, GPS-PROT, PINA), bona-fide experimental interactions (n=165 interacting gene pairs).

b) genes annotated to three 3 KEGG pathways in which TNFSF14 and TNFRSF14 were present: Cytokine-cytokine receptor interaction pathway, Herpes Simplex Infection Pathway and NF-kappaB Signaling Pathway.

c) HLA class I and class II genes, reported on coordinate file available at https://www.cog-genomics.org/plink/1.9/resources#genelist URL (N=24 genes: *HLA-A, HLA-B, HLA-C, HLA-E, HLA-F, HLA-G, HLA-H, HLA-J, HLA-L, HLA-DMA, HLA-DMB, HLA-DOA, HLA-DOB, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DRA, HLA-DRB1, HLA-DRB4, HLA-DRB5, HLA-DPB2, HLA-DRB6*).

Epistatic interactions were tested in the four cohorts (IG, IC, MSC, OE). We tested pairwise interacting SNPs, extracting markers within region of each gene, with located in a flanking window of ±10kb for each gene, to account both for coding SNPs and for variants that can affect transcriptional regulation. Overall, we tested 561 interactions among *TNFSF14* and *TNFRSF14* and

370 genes. Pairwise SNP interaction analyses were conducted using logistic regression model (PLINK software). Logistic regression models were fitted incorporating the two SNPs' additive marginal effects and a multiplicative interaction term, according to additive coding, on which Wald test was performed to detect departure from additivity on the log-odds scale. Only for 71 pairs of SNPs has been found a significant interaction after Bonferroni correction (p<10-4) but in only one data-set. These results belong to the interaction of *TNFSF14* with 5 genes: *C3, PLCG2, PTPN11* and 2 *HLA* genes.

6.8.3 Calculation of a weighted Genetic Risk Score on interactors of *TNFSF14* and *TNFRSF14*.

We constructed a genetic score to analyse the cumulative effect of strongly validated MS risk variants located in genes involved in the TNFSF14 pathway. We used only physical and validated interactors of *TNFSF14* and *TNFRSF14*, (156 genes), that were prioritized basing on the reported association on the international MSchip cohort. We selected only genes for which an association signal (p<1*e-5) was reported in the region of ±50kb containing the gene. Thanks to these criteria, we selected for this analysis 13 genes which are: *LTBR, TNFRSF14, TNFSF14, STAT3, EIF3I, TNFRSF6B, BAD, CDC37, DRAP1, GRB2, CASP8, PSPH, PTPN11*.

We calculated wGRS both on MSC (13 SNPs; 938 MS, 952 HC) and IC (10 SNPs; 961 MS, 962 HC) data sets, following a model developed by De Jager et al (De Jager et al., 2009a). The effect of each variant was weighted basing on the ln of the OR of the international MSchip sample set. The mean wGRS value was higher in MS compared to HC for both populations (p<0.0001, T-test). As expected the area under the Receiver Operating Characteristic (ROC) curve was small (<0.6) although significantly different from the null value (0.5), indicating that these variants account only for a small fraction of total MS genetic susceptibility (figures 37-38).
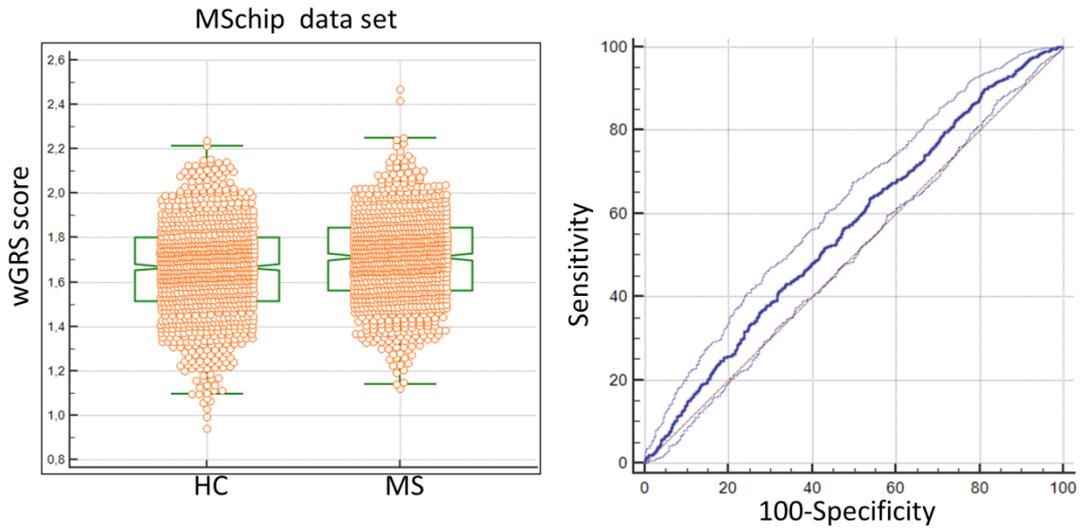
Figure 37: Distribution of the wGRS on MS chip data set of 13 SNPs in 938 MS patients and 962 healthy controls (on the left) and the graph of the ROC curve (on the right). The comparisons between the wGRS of MS patients and of controls was statistically significant (p-value<0,0001).
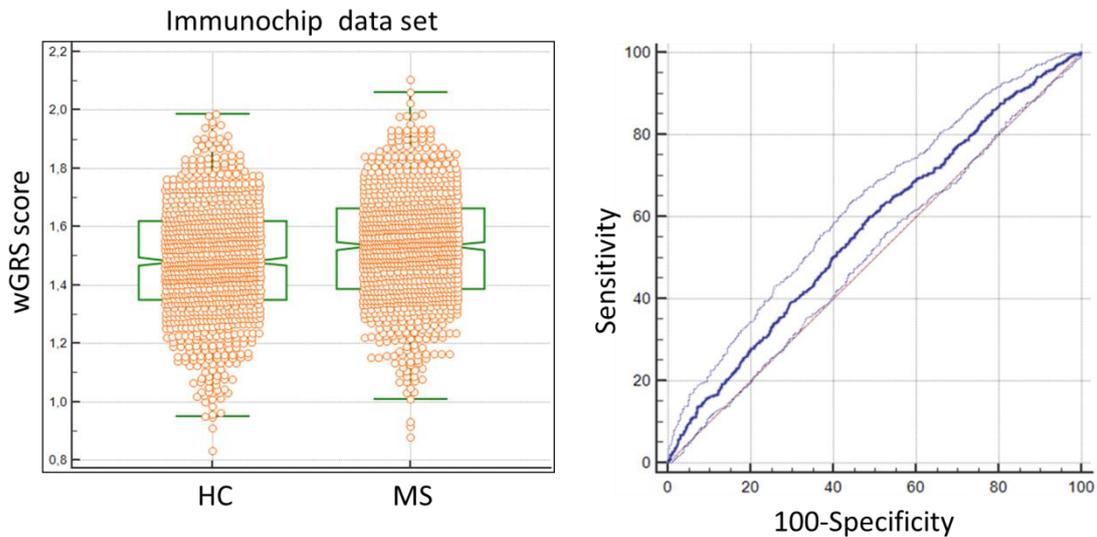
Figure 38: Distribution of the wGRS on Immunochip data set of 10 SNPs in 961 MS patients and 962 healthy controls (on the left) and the graph of the ROC curve (on the right). The comparisons between the wGRS of MS patients and of controls was statistically significant (p-value<0,0001).

6.9 Discussion and conclusions

In this chapter we reported the analysis conducted on *TNFRSF14* gene which encodes for HVEM (LIGHT receptor) and our attempt to look for the primary associated variant in this region as previously done for *TNFSF14* gene. Unfortunately, after sequencing analysis and fine mapping approach on a large sample set of 3314 MS and 3272 healthy controls we were not able to distinguish

which is the primary associated variant due to high LD present in this region. So, we tried to purse various approaches in order to prioritize the cluster of variants in high LD with the most associated variant (rs3748817) in the international cohort and also confirmed in our data set, and the most consistent data were observed for eQTL analysis. In fact, we found in different databases an eQTL effect for the associated variant rs3748817 and its proxy rs2258734 (r2= 1), a variant that maps in the promoter of *TNFRSF14* gene. In details the risk alleles of both these variants significantly correlated with a decrease expression of *TNFRSF14* in different data sets, especially in brain-cerebellum (p-value=2,2e-04, beta=-0,4) and in EBV-transformed lymphocytes (p=2,3e-3, beta=-0,15) from Gtex and Geuvadis data set (p = 3,03e-06) and in whole blood from Blood eQTL browser (p = 1,534e-13). No association was found for the expression of surface HVEM protein and the risk genotype of rs3748817 by flow cytometry analysis in whole blood in different cell types, as well as no association was reported for its expression on in vitro MDDC cells. Parallel to this analysis we explored the interactome of TNFSF14 and TNFRSF14 pathway, at the purpose to conduct an analysis for rare variants following the same pipeline conducted for other MS risk loci (as described in chapter 3) in order to investigate the cumulative effect of rare variants in the susceptibility to MS. We selected for this analysis genes which showed an interaction with *TNFSF14* or *TNFRSF14* genes experimentally valuated or predicted from online tools. Among these, we prioritized those genes which showed a SNP with an already reported association in the international MSchip project. With these criteria we prioritized 31 genes for NGS analysis. We found a significant burden for three genes: *EIF3E* for rare regulative variants, *RUVBL2* and *CDC37* for rare missense variants. *EIF3E* (eukaryotic translation initiation factor 3, subunit E) is a component of the eukaryotic translation initiation factor 3 complex, which is required for several steps in the initiation of protein synthesis. In a recent work this gene was found down-regulated in systemic lupus erythematosus after a meta-analysis of gene expression profiles of peripheral blood cells (Bae and Lee, 2018).

RUVBL2 gene encodes for the second human homologue of the bacterial RuvB gene, a protein belonging to the AAA+ family (**A**TPases **A**ssociated with diverse cellular **A**ctivities). Functional analysis showed that this gene product has both ATPase and DNA helicase activities. This protein is also a component of the NuA4 histone acetyltransferase complex which is involved in transcriptional activation of select genes principally by acetylation of nucleosomal histones H4 and H2A. This modification may both alter nucleosome-DNA interactions and promote interaction of the modified histones with other proteins which positively regulate transcription. In previous works was identified a high level of autoantibodies against RuvB protein in sera of patients with autoimmune diseases such as polymyositis/dermatomyositis and autoimmune hepatitis (Makino et al., 1998), also Anti-RuvBL1/2 antibody was detected in patients with systemic sclerosis (Pauling et al., 2018). Finally,

gene encodes for a protein similar to Cdc 37, a cell division cycle control protein of Sacchromyces cerevisiae. This protein is a molecular chaperone with specific function in cell signal transduction. It has been shown to form complex with Hsp90 and a variety of protein kinases including CDK4, CDK6, SRC, RAF-1, MOK, as well as eIF2 alpha kinases. It is thought to play a critical role in directing Hsp90 to its target kinases. Heat shock protein 90 (Hsp90), a 90-kDa molecular chaperone, is responsible for biological activities of key signalling molecules such as protein kinases, ubiquitin ligases, steroid receptors, cell cycle regulators, and transcription factors regulating various cellular processes, including growth, survival, differentiation, and apoptosis. Recent in vitro and in vivo studies have shown that Hsp90 is also involved in activation of innate and adaptive cells of the immune system (neutrophils, dendritic cells, macrophages, B and T lymphocytes) and that anti Hsp90 therapies have a clinical used in several autoimmune diseases (Tukaj and Węgrzyn, 2016). Since epistasis seems to be a ubiquitous phenomenon in complex traits (Moore, 2003), being an important contributor that might account for a substantial proportion of missing heritability (Zuk et al., 2012) we thus conducted gene-gene interaction analysis for the two genes *TNFSF14* and *TNFRSF14,* modelling multiple loci jointly, searching for non-additive effects beyond the single SNP effects. We adopted a "candidate-interactions" strategy, leveraging information from freely available protein-protein interaction (PPI) and pathways resources. Potentially interacting SNPs were hence prioritized in order to narrow down the search space, extracting candidate interacting pairs from three sources: PPI resources (STRING, Reactome, GPS-PROT, PINA), 3 KEGG pathways, HLA class I and class II genes (22 genes) for a total of 561 interactions among *TNFSF14* and *TNFRSF14* and 370 genes. Epistatic interactions were tested in four cohorts (IG, IC, MSC, OE). No significant interaction after Bonferroni correction (p<10-4) was found for pairs of SNPs in common among the 4 data sets. Finally, we calculated the weight genetic risk score (wGRS) on 13 genes belonging to TNFSF14 pathways in two different data sets (Immunochip and MS chip) and we found a significant wGRS for both data sets, confirming that variants in this pathway have a role in MS susceptibility. The area under the ROC curve was however very small (<0.6), indicating that these variants account only for a small fraction of total MS genetic susceptibility.

7. General conclusions and future perspectives

Multiple sclerosis (MS) is a chronic inflammatory demyelinating disease of the central nervous system (CNS) with an etiology not still completely understood. It is known that both environmental and genetics factors contribute to its etiology. In this thesis we disclosed about genetic factors involved in the susceptibility of the disease. In fact, in the last decade, thanks to the efforts of genome-wide association studies (GWAS) and the international collaboration of different countries, analyzing 47,429 MS subjects and 68,374 controls, 233 loci have been associated to MS susceptibility to date. Among them, 32 maps in HLA region, one on chromosome X and 200 outside the HLA region (GWAS –IMSGC 2011; IC-IMSGC, 2013; IMSGC 2017). Significantly these studies have confirmed the role of adaptive and innate immune cells and pathways steering the risk of developing MS. The results also suggest functional responses of brain-resident cells such as microglia and astrocytes affecting susceptibility. For most loci, however, the specific DNA variation causative of the association statistical signal and the mechanism linking susceptibility with brain inflammation and autoimmunity remain unknown. Our genetics laboratory aimed to investigate genetic markers involved in the susceptibility to MS specific for the Italian population. At this purpose we focused on known MS associated regions showing a significant association in the Italian population in order to find the primary associated genes or variants and then to try to define their functional role. Thanks to our analysis, we identified an intronic variant (rs1077667) in the *TNFSF14* gene (encoding for LIGHT protein) as the primary associated one and we were able to define its functional role in the regulation of gene transcription and protein production. In details this variant seems to be associated with a low *TNFSF14* RNA expression in a mixed population of PBMCs and with a higher percentage of LIGHT positive cells in myeloid dendritic cells, suggesting a cell specific influence of this variant on LIGHT expression at the protein level. Furthermore, we showed that patients in general displayed a lower expression level of this gene compared to controls as reported in previously studies (Jernås et al., 2013; Romme Christensen et al., 2013). Our discoveries also are in line with the role of LIGHT in determining MS pathogenesis in MS murine model of experimental autoimmune encephalomyelitis (EAE), In fact, LIGHT-deficient mice developed severe EAE resulting in an atypically high mortality rate (Maña et al., 2013). Seen the involvement of dendritic cells in MS pathogenesis (Serafini et al., 2006), our results might suggest a key role of this gene in the etiology of the disease and, in future, will be interesting to better investigate the role of LIGHT in dendritic cell signaling in the immune response in order to unveil new possible disease mechanisms and therapeutic targets. In this thesis we also showed our analyses on interactors of LIGHT, and first among all, *TNFRSF14* gene which encodes for HVEM receptor (the main receptor of LIGHT in the immune cells) and one among the

MS risk loci. We tried to define the most associated variant in the *TNFRSF14* gene region but without success, due to high linkage disequilibrium (LD). Despite this, we observed a cis-eQTL effect for different variants in this region on *TNFRSF14* gene expression. So, based on these evidences, we proposed for these variants a possible role in gene regulation (especially for a SNP in the gene promoter, in high LD with the associated variant in the international studies). Although we did not confirm this effect on protein production in a specific cell population, further analysis will be required to confirm our hypothesis and to try to better investigate the regulative role for the most interesting variants inside the region of *TNFRSF14* gene region. Gene-gene interaction analysis, burden test and weight genetic risk score on TNFSF14-TNFRSF14 pathway seemed to confirm our hypothesis that also genes which interact with *TNFSF14*, can also play a role in MS pathogenesis. Further analysis will be required to better investigate the causative variant in these genes and to study in deeper the role of this pathway in MS pathogenesis.

Parallel to these analysis, we conducted a research of rare functional variants in MS associated loci in order to assess if the genes in these regions showed an imbalance of rare variant frequencies (burden) between MS patients and healthy controls. *EFCAB13* was the gene that seemed to show the most promising result especially for disruptive variants (stop-gain, stop-loss and slicing). *EFCAB13* encodes for EF-hand calcium-binding domain-containing protein 13 which is a poorly characterized calcium binding adaptor protein. Further analysis will be required in order to define its functional role in mediating calcium induced cytoskeletal remodelling, mechanisms of endocytosis and recycling of membrane receptors, and modulating gene expression profile by specific shRNA and overexpression using lentiviral constructs in cell lines. Our preliminary RT-PCR experiments showed that *EFCAB13* is expressed in PMBC, activated T cells and dendritic cells (both immature and LPS activated), so in future we wanted to conduct immunological analysis of individuals carrying risk allele variants of *EFCAB13* such as in vitro functional analyses of cultured T and B cells and serologic analyses on cytokine levels.

Among the possible future perspectives of this study, the analysis of the role of epigenetics in MS seems particularly attractive. Previous studies have demonstrated that epigenetic modifications such as DNA methylation and histone acetylation can play a role in MS pathogenesis (Rito et al., 2018). Epigenetic changes in MS could explain a portion of missing hereditability as well as provide a contribution to the interpretation of the functional role of MS loci identified by means of GWAS. Accordingly, for the future, it will be interesting to investigate this aspect especially at the light of our results that showed a cell specific regulation in the gene transcription and protein production.

Like many complex diseases, Multiple Sclerosis presents a varied clinical picture. Despite the improvements in the knowledge about disease susceptibility variants, little is known so far about the

genetic variants involved in disease severity and progression, or about the response to therapy. Recently, our laboratory has took part in a European project (MultipleMS) with the aim to develop novel personalised medicine approaches for MS patients. To this end we will identify a combination of evidence-based selection of clinical, biological, and lifestyle features that can predict the clinical course, stratify patients based on their risk and the therapeutic response to the existing DMTs in a real-world setting, and to gain in-depth knowledge of distinct pathogenic pathways to allow identification of targets for novel treatments. Uncovering these aspects can open the way that may lead to personalised medicine.

Table S1: list of 15 associated variants derived from sequencing analysis of *TNFSF14* gene.

					Sequencing data set				Target genotyping platform data set						GWAS 1 data set						GWAS 2 data set						meta-anlysis 3 datasets			
					allelic frequencies		association (not conditioned)		allelic frequencies		association (not conditioned)		association (conditioned for rs1077667)		allelic frequencies		association (not conditioned)		association (conditioned for rs1077667)		allelic frequencies		association (not conditioned)		association (conditioned for rs1077667)		association (not conditioned)		association (conditioned for rs1077667)	
position (hg1...	SNP	ALT	REF	Function a	AF MS	AF HC	P-value	OR	AF MS	AF HC	P-value	OR	P-value	OR	AF MS	AF HC	P-value	OR	P-value	OR	AF MS	AF HC	P-value	OR	P-value	OR	P-value	OR	P-value	OR
19:6658613	rs344570	T	C	intergenic	0.081	0.12	**0.013**	0.68	0.088	0.097	0.33	0.89	0.46	1.10	0.065	0.11	**4.1E-06**	0.54	**0.0022**	0.65	0.091	0.08	0.86	1.03	0.43	1.14	**0.0012**	0.7780	0.3402	0.9233
19:6658863	rs8113119	A	G	intergenic	0.061	0.041	**0.042**	1.53	0.062	0.053	0.24	1.20	0.45	1.13	0.027	0.034	0.20	0.77	0.10	0.72	0.041	0.044	0.85	0.96	0.84	0.96	0.95	1.0069	0.6695	0.9552
19:6659597	rs344566	A	G	intergenic	0.080	0.12	**0.0081**	0.67	NA (*)	NA (*)	NA (*)	NA (*)	NA (*)	NA (*)	0.16	0.19	**0.0026**	0.75	0.088	0.84	0.086	0.075	0.57	1.10	0.22	1.24	NA	NA	NA	NA
19:6661983	rs12461880	C	T	intergenic	0.067	0.094	**0.028**	0.70	0.073	0.095	**0.015**	0.72	0.95	0.99	0.08	0.12	**0.0029**	0.71	0.63	1.08	0.079	0.102	**0.023**	0.71	0.13	0.74	**8.1E-06**	0.7126	0.6154	0.9506
19:6663092	rs148085223	A	G	downstream	0.012	0.025	**0.036**	0.48	0.017	0.0093	0.094	1.74	0.10	1.72	0.0034	0.0084	0.086	0.41	0.074	0.40	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
19:6665020	rs344560	T	C	exonic (missense K214E)	0.041	0.070	**0.0080**	0.57	0.064	0.060	0.82	1.04	0.16	1.26	0.051	0.063	0.15	0.80	0.82	1.04	0.056	0.056	0.76	0.94	0.88	1.03	0.37	0.9191	0.2787	1.1140
19:6665387	rs142044586	T	G	intronic	0.025	0.045	**0.015**	0.54	0.024	0.038	**0.039**	0.65	0.42	0.83	0.020	0.046	**4.4E-05**	0.41	**0.016**	0.56	0.025	0.031	0.43	0.82	0.92	0.97	**4.03E-05**	0.5841	**0.0464**	0.7523
19:6666427	ND	T	G	intronic	0.0036	0.028	**3.1E-06**	0.12	0.0011	0.00058	0.680	1.70	0.76	1.49	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
19:6666428	ND	C	T	intronic	0.054	0.033	**0.029**	1.67	NA (**)	NA (**)	NA (**)	NA (**)	NA (**)	NA (**)	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
19:6668972	rs1077667	T	C	intronic	0.11	0.18	**1.5E-05**	0.56	0.13	0.18	**3.2E-05**	0.66	NA	NA	0.13	0.19	**8.4E-07**	0.62	NA	NA	0.138	0.162	0.088	0.82	NA	NA	**1.4E-10**	0.6810	NA	NA
19:6669934	rs2291668	A	G	exonic (synonymous A49A)	0.10	0.16	**5.3E-05**	0.58	0.11	0.15	**0.0003**	0.67	0.67	0.91	0.11	0.16	**0.0002**	0.68	0.059	1.65	0.121	0.132	0.33	0.88	0.20	1.45	**6.20E-04**	0.7252	0.1623	1.2255
19:6672763	rs77612372	A	G	intergenic	0.037	0.017	**0.0095**	2.27	NA (**)	NA (**)	NA (**)	NA (**)	NA (**)	NA (**)	0.023	0.045	**0.0008**	0.50	**0.0002**	0.46	0.035	0.041	0.25	0.77	0.21	0.75	NA	NA	NA	NA
19:6672995	rs62123257	C	T	intergenic	0.37	0.42	**0.020**	0.80	0.39	0.41	0.12	0.89	0.69	1.04	0.40	0.43	0.078	0.88	0.27	1.10	0.377	0.378	0.98	1.003	0.31	1.11	**0.04349**	0.9139	0.1537	1.0776
19:6673164	rs78637822	C	G	intergenic	0	0.0051	**0.028**	0	NA (*)	NA (*)	NA (*)	NA (*)	NA (*)	NA (*)	NA (***)	NA (***)	NA (***)	NA (***)	NA (***)	NA (***)	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
19:6675230	rs1862509	A	G	intergenic	0.097	0.15	**0.00050**	0.61	0.12	0.14	**0.0066**	0.74	0.28	1.28	0.11	0.15	**0.0004**	0.69	0.078	1.50	0.112	0.125	0.28	0.87	0.49	1.19	**1.233E-05**	0.7494	**0.0389**	1.3249

All the analysis, apart from the association analysis on the pooled sequencing dataset, have been covaried for sex. We could not confirm four variants on the target genotyping platform dataset due to bad quality, but two of them have been replicated on one of the two GWAS datasets. Aminoacid positions are referred to the full-length isoform.

 * *Designed failed*

** Removed after QC for bad clustering

*** Below frequency threshold

Table S2: list of 22 variants derived from the meta-analysis step in *TNFSF14* gene

POSITION	SNP	A1	A2	Sequencing data set (not conditioned)		Target genotyping data set (not conditioned)		Meta-analysis (not conditioned)			Meta-analysis (conditioned for rs1077667)			Meta-analysis (conditioned for rs2291668)			Meta-analysis (conditioned for rs142044586)			Meta-analysis (conditioned for rs344558)			Meta-analysis (conditioning for rs1862509)		
				P-value	OR	P-value	OR	P-value	OR	I	P-value	OR	I	P-value	OR	I	P-value	OR	I	P-value	OR	I	P-value	OR	I
19:6658613	rs344570	T	C	0.013	0.68	0.3293	0.8873	0.001192	0.7780	82.92	0.3402	0.9233	79.22	0.1863	0.8930	81.43	0.0271	0.8366	76.09	0.0003067	0.7544	84.88	0.03743	0.8420	81.45
19:6658636	rs344569	A	G	0.950	1.012	0.9711	0.9966	0.002844	0.8410	64.49	0.4108	0.9500	59.77	0.0593	0.8923	63.54	0.0006107	0.8179	63.29	0.299	0.9328	0.00	0.08124	0.8933	43.55
19:6658863	rs8113119	A	G	0.042	1.53	0.2318	1.203	0.9485	1.0069	35.43	0.6695	0.9552	35.02	0.7087	0.9605	27.73	0.8987	0.9864	36.93	0.9753	1.0033	49.25	0.701	0.9591	29.07
19:6659855	rs12608923	A	G	0.613	1.053	0.3593	1.075	0.04281	1.1016	0.00	0.6486	1.0228	0.00	0.223	1.0622	0.00	0.0605	1.0952	0.00	0.1256	1.0772	0.00	0.1731	1.0700	0.00
19:6659982	rs10410439	C	G	NA	NA	0.1387	1.276	0.9582	0.9943	52.15	0.5139	0.9302	43.49	0.6339	0.9485	44.36	0.8826	0.9838	56.81	0.9214	0.9891	60.77	0.438	0.9164	18.22
19:6660287	rs62123253	G	A	0.270	1.124	0.6477	1.037	0.8417	1.0096	24.45	0.203	0.9393	0.00	0.3104	0.9513	0.00	0.625	0.9765	0.00	0.714	0.9823	20.30	0.4378	0.9624	0.00
19:6661983	rs12461880	C	T	0.028	0.70	0.01513	0.7223	8.126e-006	0.7126	0.00	0.6154	0.9506	10.66	0.02345	0.8210	0.00	1.81e-006	0.6937	0.00	0.0009376	0.7682	0.00	0.06934	0.8191	0.00
19:6663594	rs2279627	C	G	0.310	1.112	0.284	1.088	0.04491	1.0993	0.00	0.6814	1.0203	0.00	0.3468	1.0470	0.00	0.1031	1.0809	0.00	0.1694	1.0683	0.00	0.1954	1.0657	0.00
19:6664054	rs344561	G	A	0.0798	0.84	0.05229	0.8551	0.01561	0.8896	67.66	0.744	1.0178	41.90	0.5021	0.9655	48.06	0.1608	0.9317	55.67	0.02865	0.8985	67.12	0.355	0.9534	52.99
19:6665020	rs344560	T	C	0.0080	0.57	0.8173	1.037	0.3716	0.9191	0.00	0.2787	1.1140	0.00	0.442	1.0813	0.00	0.3132	0.9086	0.00	0.2164	0.8890	0.00	0.5658	1.0592	0.00
19:6665387	rs142044586	T	G	0.015	0.54	0.03922	0.6458	4.03e-005	0.5841	57.73	0.04637	0.7523	21.37	0.01715	0.7043	37.44	NA	NA	NA	2.639e-005	0.5736	62.33	9.484e-005	0.5970	60.18
19:6666316	rs4533396	A	G	NA	NA	0.3684	1.07	0.8299	0.9903	4.92	0.3129	0.9545	0.00	0.3937	0.9614	0.00	0.2836	0.9515	0.00	0.907	1.0054	0.00	0.507	0.9698	0.00
19:6668972	rs1077667	T	C	1.5E-05	0.56	0.0000324	0.6642	1.363e-010	0.6810	38.02	NA	NA	NA	6.202e-005	0.5807	39.35	1.297e-006	0.7281	0.00	4.395e-009	0.6981	36.67	3.19e-006	0.5587	13.40
19:6669934	rs2291668	A	G	5.3E-05	0.58	0.0002796	0.6749	6.199e-007	0.7252	35.02	0.1623	1.2255	42.10	NA	NA	NA	0.001547	0.7958	0.00	1.59e-007	0.7106	37.81	0.01236	0.7863	0.00
19:6670094	rs8112236	A	G	0.262	1.291	0.1715	1.249	0.7978	1.0279	22.28	0.7437	0.9652	20.86	0.8159	0.9751	10.69	0.8621	1.0190	35.33	0.918	1.0112	34.50	0.711	0.9602	0.00
19:6670253	rs344558	C	A	0.685	0.929	0.4815	0.9024	5.012e-005	0.6929	74.78	0.002308	0.7547	73.67	6.412e-006	0.6617	74.47	1.122e-005	0.6697	72.55	NA	NA	NA	0.0003424	0.7132	65.82
19:6670529	rs3760746	G	A	0.213	0.89	0.07676	0.8775	0.0058	0.8854	5.72	0.7811	1.0141	0.00	0.4347	0.9621	0.00	0.03976	0.9109	0.00	0.1504	0.9357	0.00	0.3175	0.9526	0.00
19:6671369	rs12461821	A	G	NA	NA	0.0007534	0.6956	2.324e-006	0.7375	36.97	0.05124	1.3222	21.65	0.002767	0.8052	0.00	NA	NA	NA	4.384e-007	0.7195	43.05	0.04141	0.8218	0.00
19:6672104	rs72988360	T	C	0.7022	1.053	0.2377	1.122	0.04903	1.1256	14.98	0.447	1.0477	44.54	0.2792	1.0686	40.90	0.1432	1.0929	33.69	0.1288	1.0966	33.87	0.3012	1.0661	39.39
19:6672995	rs62123257	C	T	0.020	0.80	0.1165	0.8892	0.04387	0.9139	0.00	0.1537	1.0776	0.00	0.7503	1.0164	0.00	0.2624	0.9496	0.00	0.06038	0.9188	0.00	0.9928	1.0005	0.00
19:6675230	rs1862509	A	G	0.00050	0.61	0.006595	0.7416	1.233e-005	0.7494	0.00	0.03893	1.3249	0.00	0.3134	0.9058	0.00	2.66e-005	0.7562	0.00	0.0002361	0.7795	0.00	NA	NA	NA
19:6675334	rs1874072	G	A	0.1108	0.86	0.9088	0.992	0.1717	0.9426	5.56	0.1332	1.0747	0.00	0.6221	1.0237	0.00	0.4552	0.9674	0.00	0.8397	0.9909	0.00	0.7832	1.0132	0.00

Table S3: rare variants in *TNFSF14*.

Position (hg19)	SNP	AA_Change	Function	1st sequencing (*)		2nd sequencing (**)	
				Af_HC	Af_MS	Af_HC	Af_MS
19:6664955	rs79452416	TNFSF14:NM_172014:exon4:c.C597T:p.F199F,TNFSF14:NM_003807:exon5:c.C705T:p.F235F	synonymous	-	-	0	0.0013
19:6665057	rs145049392	TNFSF14:NM_172014:exon4:c.C495T:p.S165S,TNFSF14:NM_003807:exon5:c.C603T:p.S201S	synonymous	-	-	0.0011	0
19:6665138	rs141976417	TNFSF14:NM_172014:exon4:c.C414T:p.P138P,TNFSF14:NM_003807:exon5:c.C522T:p.P174P	synonymous	-	-	0	0.0010
19:6667151	rs143854617	TNFSF14:NM_172014:exon3:c.G163A:p.E55K,TNFSF14:NM_003807:exon4:c.G271A:p.E91K	nonsynonymous	0	0.0028	0.0012	0
19:6670024	rs138116115	TNFSF14:NM_172014:exon1:c.C57T:p.I19I,TNFSF14:NM_003807:exon2:c.C57T:p.I19I	synonymous	0	0.0027	0	0.0025
19:6670030	rs200256328	TNFSF14:NM_172014:exon1:c.C51T:p.T17T,TNFSF14:NM_003807:exon2:c.C51T:p.T17T	synonymous	-	-	0	0.0011
19:6664984	rs371136658	TNFSF14:NM_003807:exon5:c.C676A:p.R226R,TNFSF14:NM_172014:exon5:c.C568A:p.R190R	synonymous	0.0030	0	-	-
19:6669986	rs2291667	TNFSF14:NM_003807:exon2:c.C95T:p.S32L,TNFSF14:NM_172014:exon2:c.C95T:p.S32L	nonsynonymous	0	0.0011	-	-
19:6665020	rs344560	TNFSF14:NM_172014:exon4:c.A532G:p.K178E,TNFSF14:NM_003807:exon5:c.A640G:p.K214E	nonsynonymous	0.0698	0.0413	0.0561	0.0674
19:6669934	rs2291668	TNFSF14:NM_003807:exon2:c.C147T:p.A49A	synonymous	0.1646	0.1018	0.1366	0.1138

(*) 588 MS patients, 408 HC; (**)504 MS patients, 504 HC.

Table S4: results of the analysis of rare variants in *TNFSF14*

	Missense variants MAF<1%				Synonymous and missense variants MAF<1%			
	variants (n)	Cumulative number of mutated alleles			variants (n)	Cumulative number of mutated alleles		
		MS	HC	P-value		MS	HC	P-value
1st sequencing (*)	2	4	0	0.15	4	7	2	0.32
2nd sequencing (**)	1	0	1	1	6	6	2	0.29
1st and 2nd sequencing (***)	2	4	1	0.38	8	13	4	0.088

588 MS patients, 408 HC; (**)504 MS patients, 504 HC; (***)1092 MS patients, 912 HC.

References

1. (ANZgene), A.a.N.Z.M.S.G.C. (2009). Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. Nat Genet *41*, 824-828.

2. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., McVean, G.A., and Consortium, G.P. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

3. Alexander, J., Mantzaris, D., Georgitsi, M., Drineas, P., and Paschou, P. (2017). Variant Ranker: a web-tool to rank genomic data according to functional significance. BMC Bioinformatics *18*, 341.

4. Anand, S., Mangano, E., Barizzone, N., Bordoni, R., Sorosina, M., Clarelli, F., Corrado, L., Martinelli Boneschi, F., D'Alfonso, S., and De Bellis, G. (2016). Next Generation Sequencing of Pooled Samples: Guideline for Variants' Filtering. Sci Rep *6*, 33735.

5. Andrews, S. (2015). FastQC: A quality control tool for high throughput sequence data.

6. Bae, S.C., and Lee, Y.H. (2018). Meta-analysis of gene expression profiles of peripheral blood cells in systemic lupus erythematosus. Cell Mol Biol (Noisy-le-grand) *64*, 40-49.

7. Ban, M., Goris, A., Lorentzen, A.R., Baker, A., Mihalova, T., Ingram, G., Booth, D.R., Heard, R.N., Stewart, G.J., Bogaert, E.*, et al.* (2009). Replication analysis identifies TYK2 as a multiple sclerosis susceptibility factor. Eur J Hum Genet *17*, 1309-1313.

8. Banchereau, J., and Palucka, A.K. (2005). Dendritic cells as therapeutic vaccines against cancer. Nat Rev Immunol *5*, 296-306.

9. Bansal, V. (2010). A statistical method for the detection of variants from next-generation resequencing of DNA pools. Bioinformatics *26*, i318-324.

10. Baranzini, S.E., and Oksenberg, J.R. (2017). The Genetics of Multiple Sclerosis: From 0 to 200 in 50 Years. Trends Genet *33*, 960-970.

11. Barizzone, N., Pauwels, I., Luciano, B., Franckaert, D., Guerini, F.R., Cosemans, L., Hilven, K., Salviati, A., Dooley, J., Danso-Abeam, D.*, et al.* (2013). No evidence for a role of rare CYP27B1 functional variations in multiple sclerosis. Ann Neurol *73*, 433-437.

12. Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. (2005). Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics *21*, 263-265.

13. Beecham, A.H., Patsopoulos, N.A., Xifara, D.K., Davis, M.F., Kemppinen, A., Cotsapas, C., Shah, T.S., Spencer, C., Booth, D., Goris, A.*, et al.* (2013). Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat Genet *45*, 1353-1360.

14. Blakemore, W.F., and Franklin, R.J. (2008). Remyelination in experimental models of toxin-induced demyelination. Curr Top Microbiol Immunol *318*, 193-212.

15. Bourdette, D., and Yadav, V. (2008). B-cell depletion with rituximab in relapsing-remitting multiple sclerosis. Curr Neurol Neurosci Rep *8*, 417-418.

16. Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S.*, et al.* (2012). Annotation of functional variation in personal genomes using RegulomeDB. Genome Res *22*, 1790-1797.

17. Cai, G., and Freeman, G.J. (2009). The CD160, BTLA, LIGHT/HVEM pathway: a bidirectional switch regulating T-cell activation. Immunol Rev *229*, 244-258.

18. Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M., and Werner, T. (2005). MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics *21*, 2933-2942.

19. chris.cotsapas@yale.edu, I.M.S.G.C.E.a., and Consortium, I.M.S.G. (2018). Low-Frequency and Rare-Coding Variation Contributes to Multiple Sclerosis Risk. Cell.

20. cotsapas@broadinstitute.org, I.M.S.G.C.E.a. (2016). NR1H3 p.Arg415Gln Is Not Associated to Multiple Sclerosis Risk. Neuron *92*, 929.

21. Cowley, M.J., Pinese, M., Kassahn, K.S., Waddell, N., Pearson, J.V., Grimmond, S.M., Biankin, A.V., Hautaniemi, S., and Wu, J. (2012). PINA v2.0: mining interactome modules. Nucleic Acids Res *40*, D862-865.

22. De Jager, P.L., Chibnik, L.B., Cui, J., Reischl, J., Lehr, S., Simon, K.C., Aubin, C., Bauer, D., Heubach, J.F., Sandbrink, R.*, et al.* (2009a). Integration of genetic risk factors into a clinical algorithm for multiple sclerosis susceptibility: a weighted genetic risk score. Lancet Neurol *8*, 1111-1119.

23. De Jager, P.L., Chibnik, L.B., Cui, J., Reischl, J., Lehr, S., Simon, K.C., Aubin, C., Bauer, D., Heubach, J.F., Sandbrink, R.*, et al.* (2009b). Integration of genetic risk factors into a clinical algorithm for multiple sclerosis susceptibility: a weighted genetic risk score. Lancet Neurol *8*, 1111-1119.

24. De Jager, P.L., Jia, X., Wang, J., de Bakker, P.I., Ottoboni, L., Aggarwal, N.T., Piccio, L., Raychaudhuri, S., Tran, D., Aubin, C.*, et al.* (2009c). Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. Nat Genet *41*, 776-782.

25. Delaneau, O., Marchini, J., and Zagury, J.F. (2011). A linear complexity phasing method for thousands of genomes. Nat Methods *9*, 179-181.

26. Derré, L., Rivals, J.P., Jandus, C., Pastor, S., Rimoldi, D., Romero, P., Michielin, O., Olive, D., and Speiser, D.E. (2010). BTLA mediates inhibition of human tumor-specific CD8+ T cells that can be partially reversed by vaccination. J Clin Invest *120*, 157-167.

27. Didonna, A., Isobe, N., Caillier, S.J., Li, K.H., Burlingame, A.L., Hauser, S.L., Baranzini, S.E., Patsopoulos, N.A., and Oksenberg, J.R. (2015). A non-synonymous single-nucleotide polymorphism associated with multiple sclerosis risk affects the EVI5 interactome. Hum Mol Genet *24*, 7151-7158.

28. Durelli, L., Conti, L., Clerico, M., Boselli, D., Contessa, G., Ripellino, P., Ferrero, B., Eid, P., and Novelli, F. (2009). T-helper 17 cells expand in multiple sclerosis and are inhibited by interferon-beta. Ann Neurol *65*, 499-509.

29. Dyment, D.A., Cader, M.Z., Herrera, B.M., Ramagopalan, S.V., Orton, S.M., Chao, M., Willer, C.J., Sadovnick, A.D., Risch, N., and Ebers, G.C. (2008). A genome scan in a single pedigree with a high prevalence of multiple sclerosis. J Neurol Neurosurg Psychiatry *79*, 158-162.

30. Fahey, M.E., Bennett, M.J., Mahon, C., Jäger, S., Pache, L., Kumar, D., Shapiro, A., Rao, K., Chanda, S.K., Craik, C.S.*, et al.* (2011). GPS-Prot: a web-based visualization platform for integrating host-pathogen interaction data. BMC Bioinformatics *12*, 298.

31. Frischer, J.M., Bramow, S., Dal-Bianco, A., Lucchinetti, C.F., Rauschka, H., Schmidbauer, M., Laursen, H., Sorensen, P.S., and Lassmann, H. (2009). The relation between inflammation and neurodegeneration in multiple sclerosis brains. Brain *132*, 1175-1189.

32. Fuchsberger, C., Abecasis, G.R., and Hinds, D.A. (2015). minimac2: faster genotype imputation. Bioinformatics *31*, 782-784.

33. Galarza-Muñoz, G., Briggs, F.B.S., Evsyukova, I., Schott-Lerner, G., Kennedy, E.M., Nyanhete, T., Wang, L., Bergamaschi, L., Widen, S.G., Tomaras, G.D.*, et al.* (2017). Human Epistatic Interaction Controls IL7R Splicing and Increases Multiple Sclerosis Risk. Cell *169*, 72-84.e13.

34. Granger, S.W., Butrovich, K.D., Houshmand, P., Edwards, W.R., and Ware, C.F. (2001). Genomic characterization of LIGHT reveals linkage to an immune response locus on chromosome 19p13.3 and distinct isoforms generated by alternate splicing or proteolysis. J Immunol *167*, 5122-5128.

35. Granger, S.W., and Rickert, S. (2003). LIGHT-HVEM signaling and the regulation of T cell-mediated immunity. Cytokine Growth Factor Rev *14*, 289-296.

36. Gregory, A.P., Dendrou, C.A., Attfield, K.E., Haghikia, A., Xifara, D.K., Butter, F., Poschmann, G., Kaur, G., Lambert, L., Leach, O.A.*, et al.* (2012). TNF receptor 1 genetic risk mirrors outcome of anti-TNF therapy in multiple sclerosis. Nature *488*, 508-511.

37. Gregory, S.G., Schmidt, S., Seth, P., Oksenberg, J.R., Hart, J., Prokop, A., Caillier, S.J., Ban, M., Goris, A., Barcellos, L.F.*, et al.* (2007). Interleukin 7 receptor alpha chain (IL7R) shows allelic and functional association with multiple sclerosis. Nat Genet *39*, 1083-1091.

38. Hansen, T., Skytthe, A., Stenager, E., Petersen, H.C., Kyvik, K.O., and Brønnum-Hansen, H. (2005). Risk for multiple sclerosis in dizygotic and monozygotic twins. Mult Scler *11*, 500-503.

39. Holm, K., Melum, E., Franke, A., and Karlsen, T.H. (2010). SNPexp - A web tool for calculating and visualizing correlation between HapMap genotypes and gene expression levels. BMC Bioinformatics *11*, 600.

40. Holmes, T.D., Wilson, E.B., Black, E.V., Benest, A.V., Vaz, C., Tan, B., Tanavde, V.M., and Cook, G.P. (2014). Licensed human natural killer cells aid dendritic cell maturation via TNFSF14/LIGHT. Proc Natl Acad Sci U S A *111*, E5688-5696.

41. Hunt, K.A., Mistry, V., Bockett, N.A., Ahmad, T., Ban, M., Barker, J.N., Barrett, J.C., Blackburn, H., Brand, O., Burren, O.*, et al.* (2013). Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. Nature *498*, 232-235.

42. Huppke, B., Ellenberger, D., Rosewich, H., Friede, T., Gärtner, J., and Huppke, P. (2014). Clinical presentation of pediatric multiple sclerosis before puberty. Eur J Neurol *21*, 441-446.

43. International Multiple Sclerosis Genetics Consortium, P., N., Baranzini, S.E.,, Santaniello, A., Shoostari, P., Cotsapas, C., Wong, G., Beecham, A.H., James,, and T., R., J, et al. (2017). The Multiple Sclerosis Genomic Map: Role of peripheral immune cells and resident microglia in susceptibility (bioRxiv).

44. Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M.*, et al.* (2009). STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res *37*, D412-416.

45. Jernås, M., Malmeström, C., Axelsson, M., Nookaew, I., Wadenvik, H., Lycke, J., and Olsson, B. (2013). MicroRNA regulate immune pathways in T-cells in multiple sclerosis (MS). BMC Immunol *14*, 32.

46. Kimura, A., Naka, T., Nohara, K., Fujii-Kuriyama, Y., and Kishimoto, T. (2008). Aryl hydrocarbon receptor regulates Stat1 activation and participates in the development of Th17 cells. Proc Natl Acad Sci U S A *105*, 9721-9726.

47. Kingwell, E., Zhu, F., Marrie, R.A., Fisk, J.D., Wolfson, C., Warren, S., Profetto-McGrath, J., Svenson, L.W., Jette, N., Bhan, V*., et al.* (2015). High incidence and increasing prevalence of multiple sclerosis in British Columbia, Canada: findings from over two decades (1991-2010). J Neurol *262*, 2352-2363.

48. Kotlyar, M., Pastrello, C., Sheahan, N., and Jurisica, I. (2016). Integrated interactions database: tissue-specific view of the human and model organism interactomes. Nucleic Acids Res *44*, D536-541.

49. Kurtzke, J.F. (1983). Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology *33*, 1444-1452.

50. Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G*., et al.* (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506-511.

51. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

52. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, G.P.D.P. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

53. Li, Y., Willer, C.J., Ding, J., Scheet, P., and Abecasis, G.R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol *34*, 816-834.

54. Lincoln, M.R., Montpetit, A., Cader, M.Z., Saarela, J., Dyment, D.A., Tiislar, M., Ferretti, V., Tienari, P.J., Sadovnick, A.D., Peltonen, L*., et al.* (2005). A predominant role for the HLA class II region in the association of the MHC region with multiple sclerosis. Nat Genet *37*, 1108-1112.

55. Lublin, F.D., Reingold, S.C., Cohen, J.A., Cutter, G.R., Sørensen, P.S., Thompson, A.J., Wolinsky, J.S., Balcer, L.J., Banwell, B., Barkhof, F*., et al.* (2014). Defining the clinical course of multiple sclerosis: the 2013 revisions. Neurology *83*, 278-286.

56. Makino, Y., Mimori, T., Koike, C., Kanemaki, M., Kurokawa, Y., Inoue, S., Kishimoto, T., and Tamura, T. (1998). TIP49, homologous to the bacterial DNA helicase RuvB, acts as an autoantigen in human. Biochem Biophys Res Commun *245*, 819-823.

57. Malmeström, C., Gillett, A., Jernås, M., Khademi, M., Axelsson, M., Kockum, I., Mattsson, N., Zetterberg, H., Blennow, K., Alfredsson, L*., et al.* (2013). Serum levels of LIGHT in MS. Mult Scler *19*, 871-876.

58. Manolio, T.A. (2010). Genomewide association studies and assessment of the risk of disease. N Engl J Med *363*, 166-176.

59. Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K.*, et al.* (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res *34*, D108-110.

60. Maña, P., Liñares, D., Silva, D.G., Fordham, S., Scheu, S., Pfeffer, K., Staykova, M., and Bertram, E.M. (2013). LIGHT (TNFSF14/CD258) is a decisive factor for recovery from experimental autoimmune encephalomyelitis. J Immunol *191*, 154-163.

61. McFarland, H.F., and Martin, R. (2007). Multiple sclerosis: a complicated picture of autoimmunity. Nat Immunol *8*, 913-919.

62. McFarlin, D.E., and McFarland, H.F. (1982a). Multiple sclerosis (first of two parts). N Engl J Med *307*, 1183-1188.

63. McFarlin, D.E., and McFarland, H.F. (1982b). Multiple sclerosis (second of two parts). N Engl J Med *307*, 1246-1251.

64. Melé, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., Sammeth, M., Young, T.R., Goldmann, J.M., Pervouchine, D.D., Sullivan, T.J.*, et al.* (2015). Human genomics. The human transcriptome across tissues and individuals. Science *348*, 660-665.

65. Moore, J.H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered *56*, 73-82.

66. Morel, Y., Truneh, A., Sweet, R.W., Olive, D., and Costello, R.T. (2001). The TNF superfamily members LIGHT and CD154 (CD40 ligand) costimulate induction of dendritic cell maturation and elicit specific CTL activity. J Immunol *167*, 2479-2486.

67. Moutsianas, L., Jostins, L., Beecham, A.H., Dilthey, A.T., Xifara, D.K., Ban, M., Shah, T.S., Patsopoulos, N.A., Alfredsson, L., Anderson, C.A.*, et al.* (2015). Class II HLA interactions modulate genetic risk for multiple sclerosis. Nat Genet *47*, 1107-1113.

68. Mumford, C.J., Wood, N.W., Kellar-Wood, H., Thorpe, J.W., Miller, D.H., and Compston, D.A. (1994). The British Isles survey of multiple sclerosis in twins. Neurology *44*, 11-15.

69. Navone, N.D., Perga, S., Martire, S., Berchialla, P., Malucchi, S., and Bertolotto, A. (2014). Monocytes and CD4+ T cells contribution to the under-expression of NR4A2 and TNFAIP3 genes in patients with multiple sclerosis. J Neuroimmunol *272*, 99-102.

70. O'Gorman, C., Lin, R., Stankovich, J., and Broadley, S.A. (2013). Modelling genetic susceptibility to multiple sclerosis with family data. Neuroepidemiology *40*, 1-12.

71. Owens, T. (2006). Animal models for multiple sclerosis. Adv Neurol *98*, 77-89.

72. Oya, Y., Watanabe, N., Kobayashi, Y., Owada, T., Oki, M., Ikeda, K., Suto, A., Kagami, S., Hirose, K., Kishimoto, T.*, et al.* (2011). Lack of B and T lymphocyte attenuator exacerbates autoimmune disorders and induces Fas-independent liver injury in MRL-lpr/lpr mice. Int Immunol *23*, 335-344.

73. Pantazou, V., Schluep, M., and Du Pasquier, R. (2015). Environmental factors in multiple sclerosis. Presse Med *44*, e113-120.

74. Pashenkov, M., Huang, Y.M., Kostulas, V., Haglund, M., Söderström, M., and Link, H. (2001). Two subsets of dendritic cells are present in human cerebrospinal fluid. Brain *124*, 480-492.

75. Patsopoulos, N.A., Esposito, F., Reischl, J., Lehr, S., Bauer, D., Heubach, J., Sandbrink, R., Pohl, C., Edan, G., Kappos, L.*, et al.* (2011). Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. Ann Neurol *70*, 897-912.

76. Pauling, J.D., Salazar, G., Lu, H., Betteridge, Z.E., Assassi, S., Mayes, M.D., and McHugh, N.J. (2018). Presence of anti-eukaryotic initiation factor-2B, anti-RuvBL1/2 and anti-synthetase antibodies in patients with anti-nuclear antibody negative systemic sclerosis. Rheumatology (Oxford) *57*, 712-717.

77. Procaccini, C., De Rosa, V., Pucino, V., Formisano, L., and Matarese, G. (2015). Animal models of Multiple Sclerosis. Eur J Pharmacol *759*, 182-191.

78. Pugliatti, M., Sotgiu, S., Solinas, G., Castiglia, P., Pirastru, M.I., Murgia, B., Mannu, L., Sanna, G., and Rosati, G. (2001). Multiple sclerosis epidemiology in Sardinia: evidence for a true increasing risk. Acta Neurol Scand *103*, 20-26.

79. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J.*, et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet *81*, 559-575.

80. Puskarjov, M., Seja, P., Heron, S.E., Williams, T.C., Ahmad, F., Iona, X., Oliver, K.L., Grinton, B.E., Vutskits, L., Scheffer, I.E.*, et al.* (2014). A variant of KCC2 from patients with febrile seizures impairs neuronal Cl- extrusion and dendritic spine formation. EMBO Rep *15*, 723-729.

81. Raj, T., Rothamel, K., Mostafavi, S., Ye, C., Lee, M.N., Replogle, J.M., Feng, T., Lee, M., Asinovski, N., Frohlich, I.*, et al.* (2014). Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science *344*, 519-523.

82. Ramagopalan, S.V., Dyment, D.A., Cader, M.Z., Morrison, K.M., Disanto, G., Morahan, J.M., Berlanga-Taylor, A.J., Handel, A., De Luca, G.C., Sadovnick, A.D.*, et al.* (2011). Rare

variants in the CYP27B1 gene are associated with multiple sclerosis. Ann Neurol *70*, 881-886.

83. Ramasamy, A., Trabzuni, D., Guelfi, S., Varghese, V., Smith, C., Walker, R., De, T., Coin, L., de Silva, R., Cookson, M.R.*, et al.* (2014). Genetic variability in the regulation of gene expression in ten regions of the human brain. Nat Neurosci *17*, 1418-1428.

84. Rito, Y., Torre-Villalvazo, I., Flores, J., Rivas, V., and Corona, T. (2018). Epigenetics in Multiple Sclerosis: Molecular Mechanisms and Dietary Intervention. Cent Nerv Syst Agents Med Chem *18*, 8-15.

85. Rivas, M.A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C.K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N.*, et al.* (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet *43*, 1066-1073.

86. Romme Christensen, J., Börnsen, L., Ratzer, R., Piehl, F., Khademi, M., Olsson, T., Sørensen, P.S., and Sellebjerg, F. (2013). Systemic inflammation in progressive multiple sclerosis involves follicular T-helper, Th17- and activated B-cells and correlates with progression. PLoS One *8*, e57820.

87. Saccone, S.F., Saccone, N.L., Swan, G.E., Madden, P.A., Goate, A.M., Rice, J.P., and Bierut, L.J. (2008). Systematic biological prioritization after a genome-wide association study: an application to nicotine dependence. Bioinformatics *24*, 1805-1811.

88. Sawcer, S., Ban, M., Wason, J., and Dudbridge, F. (2010). What role for genetics in the prediction of multiple sclerosis? Ann Neurol *67*, 3-10.

89. Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C.C., Patsopoulos, N.A., Moutsianas, L., Dilthey, A., Su, Z., Freeman, C., Hunt, S.E.*, et al.* (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature *476*, 214-219.

90. Serafini, B., Rosicarelli, B., Magliozzi, R., Stigliano, E., Capello, E., Mancardi, G.L., and Aloisi, F. (2006). Dendritic cells in multiple sclerosis lesions: maturation stage, myelin uptake, and interaction with proliferating T cells. J Neuropathol Exp Neurol *65*, 124-141.

91. Shui, J.W., and Kronenberg, M. (2013). HVEM: An unusual TNF receptor family member important for mucosal innate immune responses to microbes. Gut Microbes *4*, 146-151.

92. Shui, J.W., Larange, A., Kim, G., Vela, J.L., Zahner, S., Cheroutre, H., and Kronenberg, M. (2012). HVEM signalling at mucosal barriers provides host defence against pathogenic bacteria. Nature *488*, 222-225.

93. Shui, J.W., Steinberg, M.W., and Kronenberg, M. (2011). Regulation of inflammation, autoimmunity, and infection immunity by HVEM-BTLA signaling. J Leukoc Biol *89*, 517-523.

94. Singh, N.P., Hegde, V.L., Hofseth, L.J., Nagarkatti, M., and Nagarkatti, P. (2007). Resveratrol (trans-3,5,4'-trihydroxystilbene) ameliorates experimental allergic encephalomyelitis, primarily via induction of apoptosis in T cells involving activation of aryl hydrocarbon receptor and estrogen receptor. Mol Pharmacol *72*, 1508-1521.

95. Steri, M., Orrù, V., Idda, M.L., Pitzalis, M., Pala, M., Zara, I., Sidore, C., Faà, V., Floris, M., Deiana, M.*, et al.* (2017). Overexpression of the Cytokine BAFF and Autoimmunity Risk. N Engl J Med *376*, 1615-1626.

96. Stodberg, T., McTague, A., Ruiz, A.J., Hirata, H., Zhen, J., Long, P., Farabella, I., Meyer, E., Kawahara, A., Vassallo, G.*, et al.* (2015). Mutations in SLC12A5 in epilepsy of infancy with migrating focal seizures. Nat Commun *6*, 8038.

97. Tamada, K., Shimozaki, K., Chapoval, A.I., Zhai, Y., Su, J., Chen, S.F., Hsieh, S.L., Nagata, S., Ni, J., and Chen, L. (2000). LIGHT, a TNF-like molecule, costimulates T cell proliferation and is required for dendritic cell-mediated allogeneic T cell response. J Immunol *164*, 4105-4110.

98. Tintore, M., Vidal-Jordana, A., and Sastre-Garriga, J. (2018). Treatment of multiple sclerosis - success from bench to bedside. Nat Rev Neurol.

99. Totaro, R., Marini, C., Cialfi, A., Giunta, M., and Carolei, A. (2000). Prevalence of multiple sclerosis in the L'Aquila district, central Italy. J Neurol Neurosurg Psychiatry *68*, 349-352.

100. Tsunoda, I., Kuang, L.Q., Libbey, J.E., and Fujinami, R.S. (2003). Axonal injury heralds virus-induced demyelination. Am J Pathol *162*, 1259-1269.

101. Tukaj, S., and Węgrzyn, G. (2016). Anti-Hsp90 therapy in autoimmune and inflammatory diseases: a review of preclinical studies. Cell Stress Chaperones *21*, 213-218.

102. Tuohy, V.K., Lu, Z., Sobel, R.A., Laursen, R.A., and Lees, M.B. (1989). Identification of an encephalitogenic determinant of myelin proteolipid protein for SJL mice. J Immunol *142*, 1523-1527.

103. Vogel, C.F., Khan, E.M., Leung, P.S., Gershwin, M.E., Chang, W.L., Wu, D., Haarmann-Stemmann, T., Hoffmann, A., and Denison, M.S. (2014). Cross-talk between aryl hydrocarbon receptor and the inflammatory response: a role for nuclear factor-κB. J Biol Chem *289*, 1866-1875.

104. Wallace, C., Rotival, M., Cooper, J.D., Rice, C.M., Yang, J.H., McNeill, M., Smyth, D.J., Niblett, D., Cambien, F., Cardiogenics, C.*, et al.* (2012). Statistical colocalization of

monocyte gene expression and genetic risk variants for type 1 diabetes. Hum Mol Genet *21*, 2815-2824.

105.   Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res *38*, e164.

106.   Wang, Z., Sadovnick, A.D., Traboulsee, A.L., Ross, J.P., Bernales, C.Q., Encarnacion, M., Yee, I.M., de Lemos, M., Greenwood, T., Lee, J.D*., et al.* (2016). Nuclear Receptor NR1H3 in Familial Multiple Sclerosis. Neuron *92*, 555.

107.   Ware, C.F., and Sedý, J.R. (2011). TNF Superfamily Networks: bidirectional and interference pathways of the herpesvirus entry mediator (TNFSF14). Curr Opin Immunol *23*, 627-631.

108.   Watanabe, N., Gavrieli, M., Sedy, J.R., Yang, J., Fallarino, F., Loftin, S.K., Hurchla, M.A., Zimmerman, N., Sim, J., Zang, X*., et al.* (2003). BTLA is a lymphocyte inhibitory receptor with similarities to CTLA-4 and PD-1. Nat Immunol *4*, 670-679.

109.   Westra, H.J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E*., et al.* (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet *45*, 1238-1243.

110.   Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A *109*, 1193-1198.

111.   Šedý, J., Bekiaris, V., and Ware, C.F. (2014). Tumor necrosis factor superfamily in innate immunity and inflammation. Cold Spring Harb Perspect Biol *7*, a016279.