UPO
UNIVERSITÀ DEL PIEMONTE ORIENTALE

UNIVERSITA' DEGLI STUDI DEL PIEMONTE ORIENTALE

Department of Health Sciences

MIRIAM ZUCCALA’

PhD in Science and Medical Biotechnology

Identification and functional validation of genetic variants associated with multiple sclerosis susceptibility in the Italian population: high density fine mapping and functional analysis of TNFSF14 gene

Annual Report 2015

Tutor: **Prof.ssa Sandra D’Alfonso**

September 2015

Background

Multiple sclerosis (MS) is a multifactorial autoimmune disease. The role of the Human Leukocyte Antigen region (HLA) in determining this predisposition is well established, an more recently, other 103 non-HLA loci have been identified . The genetics laboratory has participated in two major international studies that have allowed the identification of most of these loci and now it is conducting more detailed studies to better understand the typical susceptibility variants in the Italian population by performing functional analysis on genes with SNPs primarily associated.

Study design

This region of chromosome 19, which covers 17451 bp, including the entire *TNFSF14* gene and its 5 'and 3' flanking regions has been studied in order to identify which is the primarily MS associated variant between the two variants emerged from these studies (rs1077667 and rs2291668) and to functionally characterize them. In addition, it was also performed a fine mapping analysis of this region doing an analysis of short tandem repeats, polymorphic sequences with potential functional relevance, not analysed in chip of the genome-wide studies because of technical complexity related to the nature of these sequences. We used TaqMan technique for genotyping of two different polymorphisms (rs1077667 and rs2291668) in TNFSF14 gene and Sanger sequencing for genotyping of a microsatellite in the same gene. For the analysis of gene expression we used qRT-PCR.

Results

The association analysis in the region of chromosome 19 containing TNFSF14 has led us to identify the primarily associated variant (rs1077667). Expression studies have detected that patients have lower levels of both transmembrane isoform (full length) and of soluble isoform lacking the transmembrane domain (ΔTM) generated by alternative splicing. Furthermore stratification analysis for genotypes of the two polymorphisms (rs1077667 and rs2291668) showed that both patients and controls that are homozygous for the MS susceptibility allele , have reduced levels of expression. In silico analysis predicted that the presence of the risk allele allows the binding of transcription factor AhR, while the other allele this binding was not predicted.. The data of our study are still very preliminary to define which is the precise role of TNFSF14 in the predisposition to MS. For this reason our efforts will focus in understanding the functional role of intronic variant primarily associated and to look for other possible factors that can affect TNFSF14 gene expression.

1.INTRODUCTION

1.1 Multiple sclerosis

Multiple sclerosis (MS) is a chronic inflammatory demyelinating disease of the central nervous system (Compston 2008) that leads to changes of nerve conduction due to damage resident cells, primarily oligodendrocytes and neurons. CD4+ T cells are of primary importance in the immune cascades leading to tissue damage, but also CD8+T cells, NK cells, B cells and antibodies contribute to tissue damage. In addition, the innate immune response and mainly microglial cells participate in the events leading to lesions. The incidence and prevalence of MS differ depending on the region of the world with most affected patients distant from the equator. In Europe the incidence is about 6/100.000/year and the prevalence 1/1000 (Robert Weissert, 2013).The prevalence of MS in the Italian population shows different rates depending on the regions, in particular, in the Central and the South of Italythere are 53 cases per 100,000 inhabitants, while in the North there are 81 cases per 100,000 inhabitants (Totaro et al 2003). The prevalence in Sardinia is higher than that observed in mainland, to about 150 cases per 100,000 inhabitants (Granieri et al.2000; Pugliatti et al.2001).

1.2 Background

Multiple sclerosis is a multifactorial disease, in fact bothenvironmental and genetic factors contribute to theetiology of the disease.In fact, 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 (Mumford et al., 1994; Willer et al., 2003; Hansen et al., 2005). The genetic susceptibility is mainly due tohuman HLA II region (Human Leukocyte Antigen), and in particular with HLA-DRB1*15.01allele, with an increasing of the risk of three times (Lincoln et al., 2005; Oksenberg et al., 2004).

Recently, international studies analyzing large datasets at the genome-wide level identified 103 loci involved in the susceptibility of the disease in addition to the HLA region.

In particular, two studies published in 2011 (IMSGC, Nature, 2011) and 2013 (IMSGC, Nature Genetics 2013) , mainly contributed to these results: aGenome Wide Association Study (GWAS) performed in 2011,involving 9772 patients and 17376 controls, collected by 23 research groups working in 15 different countries andanalyzing approximately 450,000 SNPs) and the Immunochipproject in 2013 involving 12 different autoimmune diseases. In particular, for MS, approximately 14.498 subjects and 24.091 healthy controls have been analysed for196,524 SNPsin 184 genes.

- In August 2011 the International Multiple Sclerosis Genetics Consortium(IMSGC) have performed a Genome Wide Association Study (GWAS) involving 9772 patients and 17376 controls (collected by 23 research groups working in 15 different countries) and analyzing approximately 450,000 SNPs. This study has confirmed 23 of the 26 known multiple-sclerosis-associated loci and has identified 29 novel susceptibility loci(p-value <5x10-8) and further 5 new regions with strong evidence for association (p-value <5x10-7) (IMSGC, Nature, 2011). Gene Ontology analysis have shown that in the 30% of association regions, the nearest gene to the lead SNP is an immune system gene. These are genes involved in the lymphocyte function, in particular 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)andonly twogeneswitha role inaxonalneurodegeneration(GALC, KIF21B).Each of thesegenescontributesonly minimallyto the total riskof development of the disease(odds ratio, OR~1.2) andthe most part of the heritabilityof MS(about 80%) remainsunexplained. This means thatprobablymany otherlow-frequencyallelic variantsandrare variants(MAF<5%), contribute significantlyto the etiologyof MS.There are probablymore than100-200genes, each of whichcouldcontributeminimallyto the risk ofdisease(OR=1.1-1.3).

- 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 emergedin genome-wide association studiespreviouslyconducted, 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, remautoide arthritis, systemic lupus erythematosus, type I diabetes and ulcerative colitis), including 55 genes associated with MS (IMSGC, 2013).The analysis were performed on

14,498 MS patients and 24,091 healthy controls belonging to 12 countries.The results were replicated in 14,802 patients and 26,703 controls. They have identified 48 new regions associated with multiple sclerosis 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) and they have confirmed the role of these in the immune response control.

- More recently, a fine mapping of the HLA region (IMSGC, Nature Genetics 2015) analysingthe same dataset of the Immunochip project, identified a total of 12 HLA genetic variants within this region roboustly associated with MS in addition to the well-established HLA-DRB1*15.01 allele. Interestingly, these HLA variants include either risk and protective factors, and evidenced for the first time interactions among pair of HLA class II alleles (HLA-DQA1*01:01–HLADRB1*15:01 and HLA-DQB1*03:01–HLA-DQB1*03:02). Conversely, no evidence for interactions between classical HLA alleles and the 103 non-HLA risk-associated variants were identified.

1.3 Summary: in state of art of the project

Our Genetics in Novara has participated (with a total of 1,726 patients and 2,257 controls) at these threeinternational studies(GWAS –IMSGC, 2011 IC-IMSGC, 2013, HLA-IC, 2015)thanks to the cooperation of various centers, PROGEMUSand PROGRESS consortium and, coordinated respectively from the University of Eastern Piedmont and Maggiore Hospital of Novara andSan Raffaele Hospital in Milan .We have used data derived from these studies to perform a GWAS analysis only on Italian population and the results have shown that the strongest non-HLA signal for Italian population was an intronic variant (rs1077667) in the Tumor Necrosis Factor (ligand) superfamily member 14 (*TNFSF14*) gene (p-value=5.9×10-8), located in a known MS associated region. In order to define the primarily associated variant in this region, our group hassequenced the whole region and performed a preliminary association analysis on sequencing data followed by the replication of the data through individual genotyping of a selection of the detected variants. In particular, weperformed a target resequencing (using Next Generation Sequencing , NGS, techniques) on the whole genomic region (17.500 bp, flanked by recombination hot spots) on 588 multiple sclerosis patients (MS) and 408 healthy controls (HC) from the Italian population, pooled in groups of 12 samples.Validation of 129 variants (randomly located in the genome) by individual genotyping demonstrated a high correlation with allele frequency (AF) estimated in the NGS data of the pools (R^2=0.98). After quality controls (QC), there were 113 variants: 63 private variations and 50 with an AF>1%; 6 variants were in the coding region and 15 showed a significant ($p<0.05$) association with MS. Following this,we performed a fine-mapping of the gene usinga custom array (Open Array Technology, Life Technology)on an independent, individually typed sample set (867 MS and 878 HC) for 59 *TNFSF14* variations, covering all common (AF>1%), all significantly associated and coding variants derived from the NGS experiment in the pools. After QC, we observed a significant association for 7 variants, confirming 5 of the associations observed in pools (n=10 SNPs tested in open array). The strongestassociations were reported for the known MS risk variant in intron 1 (rs1077667) (p=6.2e-5) and for a synonymous variant (rs2291668).(p=4.6e-4).

1.3 TNFSF14

Tumor necrosis factor ligand superfamily member14, also known as LIGHT, is a type II transmembrane glycoprotein expressed by activated T lymphocytes, natural killer and immature dendritic cells.

LIGHT 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 aglycosylation 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-g, and LTβR (lymphotoxinβ 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 et al., Cytokine and growth factor Reviews, 2003).

There are three physical forms of LIGHT (figure 1) 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.(Steve W. Granger et al., the Journal of Immunology, 2001).

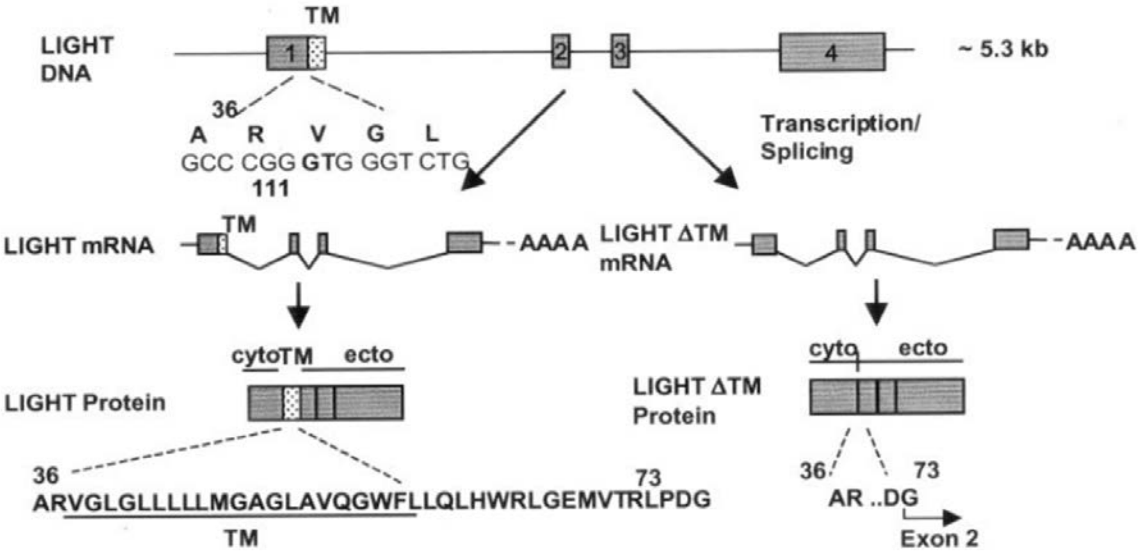


Figure 1: Representation of the isoforms produced by alternative splicing (Granger et al., J Immunol, 2001)

Aim of the project

The general aim of my project is to study genetic variants associated to susceptibility to multiple sclerosis.In detail,during this year my attention was focused on the region of chromosome 19, which covers 17451 bp, including the entire *TNFSF14* gene and its 5 'and 3' flanking regions. This region was chosen because it is strongly associated with susceptibility to MS both in the international studies (GWAS-IMSGC, rs1077667, p = 2.10x10-12), and in the Italian study, where we observed the most strongly associated region (p = 5.8x10-8 rs1077667).This region has been studied in order to identify which is the primarily MS associated variant between the two variants emerged from these studies (rs1077667 and rs2291668) and to functionallycharacterize them. In addition, I have contributed to the fine mapping analysis of this region performing an analysis of short tandem repeats, polymorphic sequences with potential functional relevance, not analysed in chip of the genome-wide studies because of technical complexity related to the nature of these sequences.

2Materials and methods

2.1 Samples

The study was approved by the ethics committee of the local hospital.Written informed consent for genetic analysis for research purposes was obtained for all MS patients and healthy controls. MS patients had been diagnosed according to McDonald criteriaand were of continental Italian ancestry. Controls used in the study do not have family history of autoimmune diseases and have similar geographic provenience of patients. A peripheral blood sample was obtained by each participant to the study. A DNA sample was purified for all individuals. For 83 patients and 79 controls (for whom fresh biological material was available) we also obtained peripheral mononuclear blood cells (PBMCs) and serum.

2.2 Acid nucleic extraction

The DNA was extracted from peripheral whole blood with standard salting out method or by using the QIAamp®DNA Blood Mini kit (QIAGEN, GmbH, Hilden , Germany). Blood samles were treated with EDTA in order to prevent its coagulation. Total RNA was extracted from pellets of PBMC using the RNeasy Plus Mini Kit (QIAGEN GmbH, Hilden, Germany), according to the protocol in the supplying company.

DNA and RNA extracts were quantified by spectrophotometric reading at Nanodrop (Fisher Scientific SAS, IllkirchCedex, France).

2.3 Peripheral blood mononuclear cell isolation

The peripheral mononuclear blood cells (PBMCs) were isolated by density gradient centrifugation using Lympholyte-H (Cedarline, Burlington, NC, USA). Blood with EDTA (about 8 ml) was layered on 4 ml of Lympholyte and centrifuged at 1800 rpm for 30 minutes without brake. At the end we collected the mononuclear cells layered at the interface Lympholyte-plasma and washed them with phosphate buffered saline (PBS, phosphate buffered saline). Finally theywere covered with RNAlater, a reagent for the preservation and RNA stabilization.

2.3 TaqMan genotyping

Genotyping of the polymorphisms rs1077667 and rs2291668 of TNFSF14 was conducted using the TaqMan technology with two commercially available Real Time-PCR assays provided by Life

Technologies Each assay comprises two primers (sense and antisense) for the amplification of the fragment in which the polymorphism is present, and two probes (labeled with the fluorophore FAM and VIC respectively) for allelic discrimination (VIC-labeled probe which recognizes allele 1 and FAM-labeled probe for allele 2) as indicated by the SDS software tool. Data from each experiment were analyzed withTaqMan Genotyping Software,which allowsto group and display data from different experiments in a single study.The reaction was performed by adding 1 ul of DNA (20 ng) to the reaction mixture (Table1). The thermal protocol shown in table 2 was run on the instrument StepOnePlus ™ Real-Time PCR System (LifeTechnologies).

Reagents	Volume(uL)
Master Mix (2X)	5
Probe (40X)	0,25
Nuclease free water	3,75
total	9

Table1: Reaction mix for TaqMan genotyping

Temperature	Time	Cycle
95° C	10 min	
95° C	15 sec	40
60° C	1 min	

Table 2: Thermal protocol for TaqMan genotyping

2.4 PCR and sequencing for microsatellite analysis

The primers used to genotype the microsatellite (CA_n): Chr19:6,669,748-6,669,793 (hg 17/human), were designed using the bioinformaticprogramme Primer 3 (http://primer3.ut.ee/) to design primers and Blast algorithm to check their specificity. PCR reaction was performed on the thermocycler Eppendorf Mastercycler by adding 50ng of genomic DNA to the reaction mixture described in Table 3and using the thermal protocol written in Table 4 and analyzed through capillary electrophoresis on the automatic sequencer ABI Prism 3100 XL genetic Analyzer (Applied Biosystem).

To optimize costs and time we designed three primers: one forward primer (labeled with FAM) and two reverse primers with the same annealing sequence but differing in base leght: a first one with a 21-bp not-annealing tail at the 3’ end and other one without this tail. The PCR reaction performed with the reverse primer with the tail generates a amplicon 21 bases longer than the product of a PCR reaction performed with the reverse primer without tail (see below). In this way in the post-PCR analysis we can load two different samples in the sequencer machine and analyze them at the same time. Below there are the sequences of the primers and details about the PCR product. Product sizes are relative to the reference allele of the microsatellite marker (45 repetitions, according to the web site STR catalog Viewer I)

Amplicon:chr19:6,669,623-6,669,814(hg 17/human)

Product size:192 (without tail); 213 (with tail)

Forward primer (labeled with FAM) :CCCCTATGAAGAGACTGCCC

Reverse primer (with tail) : gcggtcccaaaagggtcagttCCCTGTGTCATTGTGGGTC

Reverse primer (without tail) : CCCTGTGTCATTGTGGGTC

Reagents	Starting concentration	Final concentration	Volume (uL)
Buffer go TaqPromega	5X	1X	2
dNTPs	2,5mM	200uM	0,8
$MgCl_2$	25mM	1,5mM	0,6
Oligo forward	10uM	0,5uM	0,5
Oligo reverse	10uM	0,5uM	0,5
Taqpolimerase	5U/uL	0,036U/uL	0,05
H_2O	-	-	3,55
DNA	25ng/uL	50ng each one	2
total	-	-	10

Table 3: PCR mix for microsatellite CA_n of TNFSF14

Temperature	time	cycles
96° C	5min	1
96° C	40sec	35
57° C	40sec	
72° C	30sec	
72° C	5min	1
10° C	infinity	-

Table 4: thermal protocol

The PCR products were displayed on 1,5% agarose gel, PCR products were diluted to the optimal concentration depending on the intensity of the band prior to loading them on the sequencer machine. We prepared the samples in an optical plate and for each well we add 10 ul of formamide, 0,3 ul of molecular weight marker (Gene Scan 500 ROX) and 1 ul of PCR product. The results were analyzed by the sequencer software Gene Mapper v4.1.

2.5 Reverse transcription and quantitative Real Time PCR

For this analysis we used 83 patients and 79 controls from two different cohorts:

-cohort 1: 107 RNA samples (44 patients and 63 controls) purified from PBMCs

-cohort 2: 55 RNA samples (39 patients and 16 controls) purified from whole blood.

The expression of the two splicing isoforms of *TNFSF14* was determined by quantitative Real Time PCR with SYBR Green method using the GoTaq 2-step RT-qPCR system (Promega A6010). Its components allow the synthesis of cDNA using GoScript Reverse Transcription System and thesubsequent quantification by GoTaq qPCR Master Mix.

1) *Reverse transcription PCR*: For the reverse transcription we used two different kind of primers: Random Examer Primers and oligo dT primers. We performed the reaction with 4-8ul of RNA depending on its concentration (final reaction volume 10μl). The samples were then placed in the thermocycler Mastercycler (Eppendorf) for the primers annealing at 70 ° C for 5 minutes and at 4° C for 5 minutes. Afterwards a reaction mixture containing the other reaction reagents (table 6) was added to the samples, which were placed in the thermocycler for the second reaction step (table 7). The cDNA can be stored at -20 ° C and used for subsequent PCR reactions for the analysis of gene expression.

Reagents	Concentration	Volume (uL)
Oligo dT	0,025ug/uL	1
Random Primer	0,025ug/uL	1
H_2O	-	6
RNA	-	4
total	-	10

Table5: first mixture for reverse transcription

Reagents	Concentration	Volume (uL)
GoScript 5x Reaction Buffer	1X	4
$MgCl_2$, 25mM	2.5 mM	2
PCR nucleotide Mix, 10mM	0.5 mM	1
RecombinantRNasinRibonucleaseInhibitor	-	0,5
GoScript Reverse Transcriptase	-	1
Nuclease-Free Water	-	1,5
Total	-	10

Table 6: second mixture for reverse transcription

Temperature	time
25° C	4 min
42° C	1 hour
70° C	15 min
4° C	hold

Table 7: thermal protocol for reverse transcription

2) *Quantitative Real-Time PCR:* We designed with Primer 3 (http://primer3.ut.ee/) the PCR primers specific for the two TNFSF14 splicing isoforms (long and short). The qRT-PCR reaction was conducted in the instrument C1000 Thermal Cycler CFX96 Real Time System (Bio-Rad) using the thermal protocol reported in the table 10. 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. For 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. Below we report the sequences of the primers and details about real time PCR reaction with thermal protocol.

Primers	Sequence
TNFSF14 long forward	GGTGGGTCTGGGTCTCTT
TNFSF14 long reverse	AGACCTTCGCTCTTGTATCAGC
TNFSF14 short forward	AGTGTGGCCCGGGACGGA
TNFSF14 short reverse	GCTGGAGTTGGCCCCTGTGA
bactin forward	CGCCGCCAGCTCACCATG
bactin reverse	CACGATGGAGGGGAAGACGG

Table 8: primers for real time PCR

Reagents	Concentration	Volume (uL)
GoTaq qPCR Master Mix 2x	1X	10
Forward primer	10 uM	1
Reverse primer	10 uM	1
H2O	-	7
cDNA		1

total	-	20

Table 9: real time PCR mix

Temperature	Time	Cycle
95°C	2	1
95°C	15’	40
60°C	1’	
60°-90°C	-	1

Table 10: thermal protocol of real time PCR

2.6 Bionformatic Analysis

The bioinformatic analysis of the whole genomic region containing the TNFSF14 gene (CpG islands, cromatin state, transcription levels and histone mark) was conducted with the UCSC genome browser (https://genome.ucsc.edu/).

We searched for short tandem repeats within the region using STR catalog Viewer I.

To evaluate the functional role of the two SNPs rs1077777 (intron 1) and rs2291668 (synonymous variant in exon 1) we evaluated if the different alleles influence the binding of transcriptional factors (Transfac program) and if rs2291668 generates or removes consensus sequences for splicing regulatory proteins with the software ESE finder (rulai.cshl.edu/tools/ESE/).

2.7 Statistic Analysis

Genotype association was conducted with PLINK software (Purcell) and MedCalc Software.

Conditional analysis was performed trough a logistic regression (covariated for sex and conditioned for one SNP at a tyme) with PLINK software.

Linkage Disequilibrium between the multi-allelic microsatellite (CA$_n$) alleles and bi-allelic SNPs was calculated with the software MIDAS v1.0 (MultiallelicInterallelic Disequilibrium Analysis Software), which allows the analysis af multi-allelic markers.

Differences in the expression values between patients and controls and between carriers of the different genotypes were evaluated with a non-parametric Mann Whitney test using Med Calc software (© 1993-2014 MedCalc Software bvba) and box plots were generated. For the analysis stratified by genotype we merged heterozygous individuals and homozygous for the rarest allele in the same category, due to the rarity of the latter genotype.

In order to merge together the different cohorts and to take into account all the various variables and confundents we performed two linear regression analysis: 1) We tested for association between expression levels and genotype assuming an additive model and using sex, cohort, and case-control status as covariates. 2) We tested for association between expression levels and case-control status using genotype, sex and cohort as covariates. This analysis was performed with PLINK software.

2 Results

2.1 Preliminary results

The strongest non-HLA signal for the Italian population was an intronic variant (rs1077667) in the Tumor Necrosis Factor (ligand) superfamily member 14 (*TNFSF14*) gene (p-value=5.9×10-8), located in a known MS associated region. The association of this region was again confirmed in NGS: in particular the analysis on pools detected 172 variants, with 13 variants showing a significantly different allele frequencies (AF) between patients and controls, including the variant rs1077667. It was also identified a new synonymousvariant (rs2291668) in linkage disequilibrium with rs1077667 (r2 = 0.808), not present in the genotyping platforms.This SNP is located in exon 1, near the site of alternative splicing which leads to the transcript isoform ΔTM (short isoform) of TNFSF14.In the light of these results, the region containing the TNFSF14gene has been subjected to a follow up through a fine mapping: 58 of the 172 variants were genotyped through a custom array (Open Array Technology, Life Technologies) in 901 Italian patients and 883 controls. After Quality Control, the results confirmed a significant association for 7 of them (out of the 10 SNPs with a significant result in NGS), including rs1077667 (p = 3.24x10-5) and the new synonymous variant (rs2291668) in linkage disequilibrium with it (r2 = 0.75).

2.2 Genotyping analysis

My contribute to the project starts at this stage in order to increase the number of samples analysed with the custom array and to understand which is the primarily associated polymorphism variant in this region. The total number of analysed samples is 3357 (1680 healthy controls and 1677 patients) andthese samples were genotyped for two polymorphisms rs1077667 and rs2291668 by TaqMan method (materials and methods) in order to perform a conditioning analysisbetween the two variants.

First of all, I have conducted a covariateanalysis for sex, that confirmed that both polymorphisms are significantly associated with multiple sclerosis in this dataset (rs1077667: Odds Ratio = 0.6336, p-value 1.114x10-10; rs2291668: Odds Ratio = 0.6641, p-value 5.659x10-8).

Then, the analysis of association of the two SNPs conditioned one on the other,showed that only the rs1077667 (intron 1) maintains statistical significance if conditioned for rs2291668 (exon 1) (Odds Ratio 0.6034 p = 4,228x10-4) while rs2291668 becomes not statistically significant when conditioned for rs1077667 (Table 1). This analysis suggest that the intronic variant rs1077667 is primarily associated with MS while the association with the exonic variant is only the consequence of its linkage disequilibrium with the intronic variant.

SNP		Covariate for sex		Conditioning for rs2291668	
Chr	SNP	Odds ratio	pvalue	Odds ratio	pvalue
19	rs1077667	0.6336	1.114e-0.10	0.6034	0.0004228

SNP		Covariate for sex		Conditioning for rs1077667	
Chr	SNP	Odds ratio	pvalue	Odds ratio	pvalue
19	rs2291668	0.6641	5.659e-008	1.062	0.695

Table 1Analysis of association of rs1077667 and rs2291668 covariate for sex and conditioning for the polymorphism.

In order to continue the fine mapping analysis in this region I have looked for other variants near to the two most associated SNPs and in particular I focused my attention on short tandem repeats.

Short tandem repeats (STR) are among the most polymorphic loci in the human genome. These loci are highly prone to mutations due to their susceptibility to slippage events during DNA replication. To date, STR variations have been linked to at least 40 monogenic disorders including a range of neurological conditions such as Huntington's disease, amyotrophic lateral sclerosis, and certain types of ataxia. Multiple studies have suggested that STR variations contribute to splicing in humans. Population genetic studies have utilized STRs in a wide range of methods to find signatures of selection and to elucidate mutation patterns in nearby SNPs. (Thomas Willems et al., 2014). STRs are not included among the polymorphism present in the chip utilized for the genome wide association studies. Using the bioinformatics software (STR catalog Viewer) I looked for possible STR in this gene and I discovered in the region identified as active promoter, a short tandem repeat (CA) (chr19:6669748-6669793, hg19) (heterozygosity 0,744) in the same intron of rs1077667. Considering its location, this STR has been chosen to investigate the presence of linkage disequilibrium between this repeat and the two polymorphisms (rs1077667 and rs2291668). For this reason I have performed the analysis by genotyping a first series of 334 individuals (161 patients and 173 controls). In order to maximize the informativity of the linkage disequilibrium analysis,our analysis was conducted by selecting individuals based on their genotypes at the two polymorphisms (rs2291668 and rs1077667) : in detail, 57.5% of individuals have a genotype AA/GG, 35.3% genotype AG/CT and 8.4% a genotype AA/TT, and another 8% have a mixed genotype.

rs2291668 vs STR

Size	Dprime	r-squared
G_186	0.87057	0.05689
G_188	0.88636	0.25355
G_190	-0.46534	0.1583
G_192	-1.0	0.02471
G_196	-1.0	0.12464
G_198	-1.0	0.00611
A_186	-0.87057	0.05689
A_188	-0.88636	0.25355
A_190	0.46534	0.1583
A_192	1.0	0.02471
A_196	1.0	0.12464
A_198	1.0	0.00611

rs1077667 vs STR

Size	Dprime	r-squared
C_186	0.86849	0.05941
C_188	0.96592	0.31594
C_190	-0.55841	0.23917
C_192	-1.0	0.02355
C_196	-1.0	0.11879
C_198	-1.0	0.00582
T_186	-0.86849	0.05941
T_188	-0.96592	0.31594
T_190	0.55841	0.23917
T_192	1.0	0.02355
T_196	1.0	0.11879
T_198	1.0	0.00582

Table 2 analysis of linkage disequilibrium between the alleles at chr19:6669748-6669793, hg19 STR and the two rs1077667 and rs2291668 SNPs

As we can see in the tables, the results of the analysis to assess the linkage disequilibrium (performed by MIDAS) show a certain association between the longer alleles of the microsatellite (>190) (corresponding to allele >43) and the A allele for the rs2291668 and the T allele for the rs1077667 . Moreover, analyzing the allelic frequency of these STR alleles (table 3), we can observe that they show a higher frequency in the controls compared to patients in the three layers of individuals classified on the basis of the combination of rs1077667 and rs2291668 genotypes .

Genotype GG/CC	**MS**		**CTR**	
allele	n	freq	n	freq
186	34	0,205	40	0,253
188	91	0,548	84	0,532
190	40	0,241	32	0,203
>190	1	**0,006**	2	**0,012**
totale	166	1	158	1

Genotype AG/CT	**MS**		**CTR**	
allele	n	freq	n	freq
186	13	0,112	9	0,075
188	31	0,267	36	0,300
190	58	0,500	58	0,483
>190	14	**0,121**	17	**0,142**
totale	116	1	120	1

Genotype AA/TT	**MS**		**CTR**	
allele	n	freq	n	freq
186	0	0	0	0

188	0	0	0	0
190	15	0,833	27	0,711
>190	3	**0,167**	11	**0,289**
totale	18	1	39	1

Table 3Stratified analysis according to rs1077667 /rs2291668 SNPs genotypes

We performed a statistical analysis of association with multiple sclerosis of the two polymorphisms rs1077667 and rs2291668 and the intronicmicrosatellite in an unbiased cohort (i.e. without a selection for the genotypes of the two polymorphisms) analyzing 2664 samples (1214 patients and 1450 controls). The distribution of the microsatellites alleles was significantly different between cases and controls (Yates' p-value: 0.016). In particular, we observed that the allele 196 shows a significantly higher frequency in the controls (0,033) compared to patients (0,017) (p-value=0,0001504,odds ratio=0,5). As expected, a significant association is also present for the rs1077667 (p-value=0,0000041, odds ratio=0,7).

allele STR	**patients**	
allele	**n**	**frequency**
182	0	0
184	4	0,002
186	378	0,156
188	1181	0,486
190	779	0,321
192	39	0,016
194	1	0,0004
196	**41**	**0,017**
198	4	0,002
200	1	0,0004
TOTAL	2428	1

allele STR	**controls**	
allele	**n**	**frequency**
182	1	0,0003
184	1	0,0003
186	414	0,143
188	1406	0,485
190	934	0,322
192	41	0,014
194	5	0,002
196	**97**	**0,033**
198	0	0
200	1	0,0003
TOTAL	2900	1

rs1077667	**patients**	
allele	**n**	**frequency**
T	330	0,14
C	2098	0,86
TOTAL	2428	1
rs1077667	**controls**	
allele	**n**	**frequency**
T	531	0,18
C	2369	0,82
TOTAL	2900	1

Table 4 distribution of the microsatellites allele frequencies, of rs1077667 and rs2291668 in patients and controls.

Then we performed a conditioning analysis in order to understand which is the variant primarily associated with the disease. In this analysis, we have considered the 196 microsatellite allele showing a significant MS association, as well as the presence of all the alleles longer than 190 and. This analysis shows that the intronic variant rs1077667 is primarily associated with MS while the association with exon variant rs2291668 and the alleles of the microsatelliteare only the consequence of its linkage disequilibrium with the intron (table 5). In this part of the study we confirmed that the rs1077667 variant intron is directly responsible for the association with MS for this region.

brute		**p-val**	**OR**
rs1077667	T	1,08E-06	0.6748
microsatellite	>190	0,01711	0.7086
rs2291668	A	0,001138	0.7345

microsat	196	0,0005852	0.5106

Conditioned forrs1077667		**p-val**	**OR**
msat	>190	0.4221	0.8838
rs2291668	A	0.1826	1.29
msat	196	0.05907	0.6748
Conditioned for rs2291668		**p-val**	**OR**
msat	>190	0.058	0.7153
rs1077667	A	0.000673	0.5382
msat	196	0.01942	0.5707
Conditioned formicrosat>190		**p-val**	**OR**
rs1077667	T	1,43E-05	0.6903
rs2291668	A	0.01858	0.7876
Conditionedformicrosat 196		**p-val**	**OR**
rs1077667	T	9,90E-05	0.7153
rs2291668	A	0.03561	0.8068

Table 5: conditioning analysis between the microsatllite alleles and the two SNPs.

2.3 Bioinformatics analysis in TNFSF14 gene

At the same time I have performed a bioinformatic analysis in the TNFSF14 gene in order to look for possible mechanisms in regulation of gene expression. To this purpose, I analysed the region for methylation profileusing the browser UCSC. In fact in literature is known that the methylation in CpG islands, especially if it is present in promoter region, can regulate gene expression (Anna Portela&ManelEsteller 2010) The results are shown in the graph below.

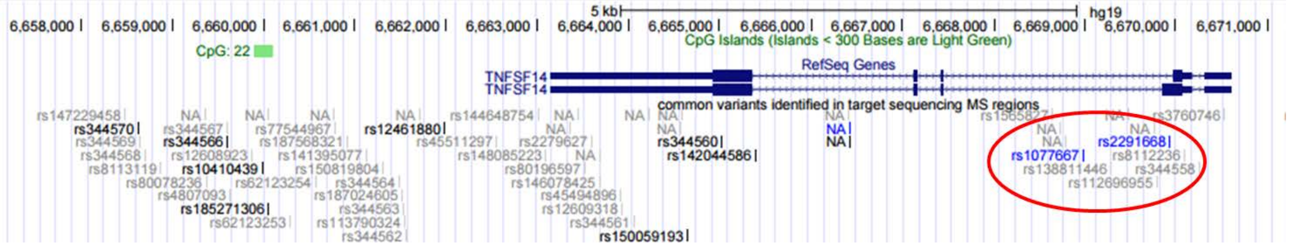


Figure 1: Methylation analysis in TNFSF14 gene by UCSC genome browser.

As we can see, the first CpG island is localized about 3 kb downstream the gene and for this reason its implication in the gene regulation is unlikely.

If we analyse the region of TNFSF14 gene which includes the two polymorphism rs1077667 and rs2291668, we observe that they are within a region identified as active promoter by UCSC genome browser thanks to computationally integrating ChIP-seq data using a Hidden Markov Model

(HMM) (Ernst and Kellis, 2010). This region covers 2600 bp and shows high levels of enrichment of the H3K4Me3 histone mark as determined by a ChIP-seq assay. The H3K4Me3 histone mark is the tri-methylation of lysine 4 of the H3 histone protein, and it is associated with promoters that are active or poised to be activated.

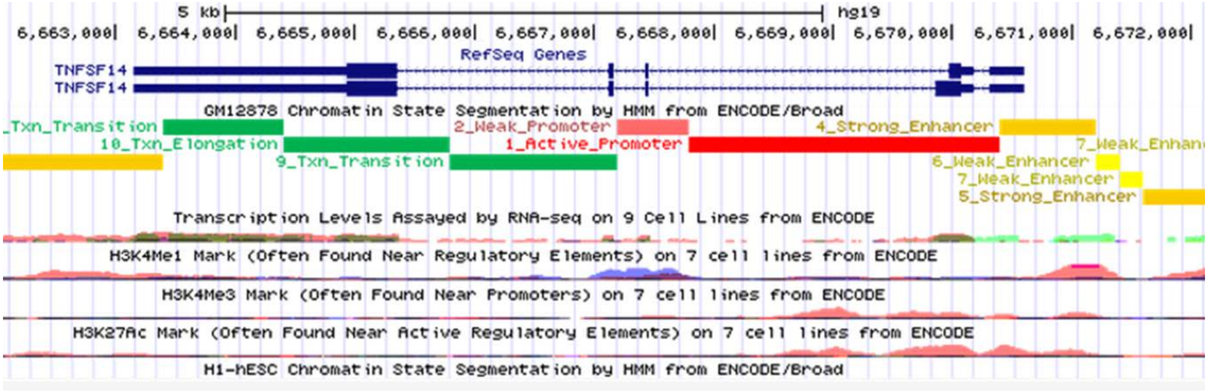

5 kb
hg19
6,663,000| 6,664,000| 6,665,000| 6,666,000| 6,667,000| 6,668,000| 6,669,000| 6,670,000| 6,671,000| 6,672,000|
RefSeq Genes
TNFSF14
TNFSF14
GM12878 Chromatin State Segmentation by HMM from ENCODE/Broad
_Txn_Transition
2_Weak_Promoter
4_Strong_Enhancer
10_Txn_Elongation
1_Active_Promoter
7_Weak_Enhan
9_Txn_Transition
6_Weak_Enhancer
7_Weak_Enhancer
5_Strong_Enhancer
Transcription Levels Assayed by RNA-seq on 9 Cell Lines from ENCODE
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
H1-hESC Chromatin State Segmentation by HMM from ENCODE/Broad

Figure 2: Bioinformatics analysis in the TNFSF14 gene by UCSC genome browser.

In light of these results the second step of my analysis focuses on these two polymorphisms. In detail, I have performed in silico analysis for the synonymous variant (rs2291668) and I have discovered that it introduces an ESE sequence (exonic splicing enhancer). In fact, the presence of C allele creates the consensus sequence of two different splicing regulatory proteins: SRSF1 and SRSF2 (Serine / Arginine-Rich Splicing Factor 1 and 2) (ESE finder3.0). Using bioinformatic databases like TRANSFAC (www.biobaseinternational.com) and MatInspector (www.genomatix.de) that allow to predict the binding sites of transcription factors, I have found that the C allele of the intronic SNP (rs1077667) allows the binding of transcription factor AHR, while the presence of the A does not allow this. In fact, the third nucleotide in the consensus sequence CACGC, corresponds to the rs1077667 polymorphic site, and it is the most important for binding. Conversely, the presence of the A allele, does not affect the binding of other transcription factors such as p53, p300, egr-3 and DEC1 (figure3). In conclusion in light of these bioinformatics analyses, the next step will be to perform a functional validation of these two variants in TNFSF14.

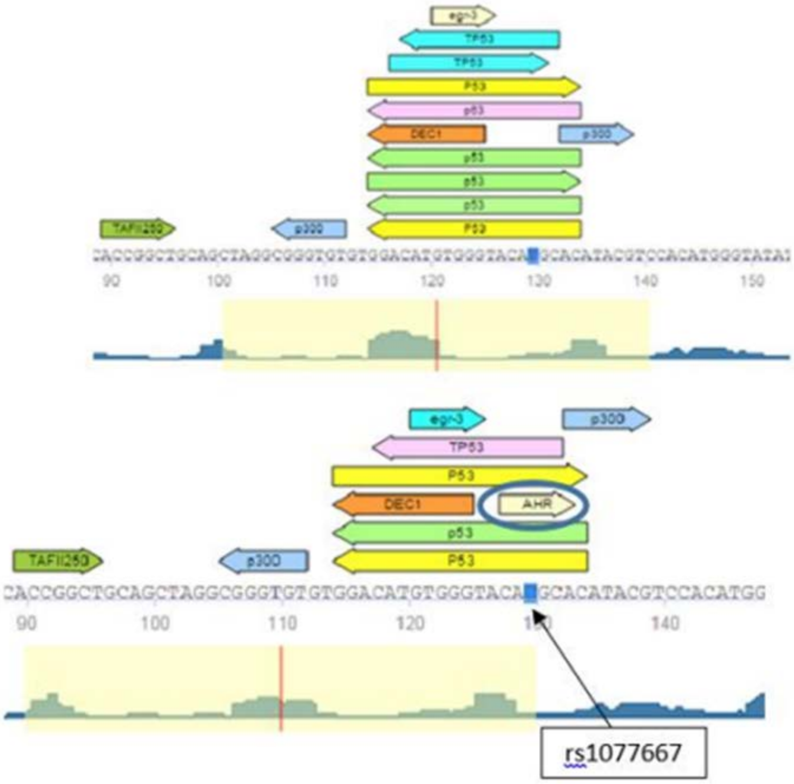


Figure 3: Schematic diagram of the binding sites of transcription factors predicted by TRANSFAC within a gene fragment of TNFSF14. The C allele of the polymorphism (in the lower part of the figure) creates a consensus sequence which allows the binding of AhR.

2.4 Expression analysis in TNFSF14

As reported in the literature (introduction, paragraph 1.3), we evaluated the existence oftwo different mRNA transcripts of TNFSF14 in PBMC (peripheral blood mononuclear cell): the full length isoform of 2894bp and the short isoform also called ΔTM, about 100 bp smaller, as a result of an alternative splicing, devoid of the transmembrane domain. We conducted a reverse transcription using primers outside the site of alternative splicing (Materials and methods), and then we amplified cDNA from 5 healthy controls, 3 with heterozygous genotype for rs1077667 and 2 homozygous for the susceptibility allele to multiple sclerosis (GG). The full length isoform was observed at levels considerably higher than the ΔTM (Figure 4). To verify if the visible bands on gels are the two isoforms of TNFSF14 mRNA splicing, we cut both the band of 779bp (corresponding to full-length isoform) and that of 652 bp (corresponding to ΔTM isoform) and then those bands were purified and sequenced. The results of sequences confirmed that the bands are the two forms of alternative splicing in the literature.

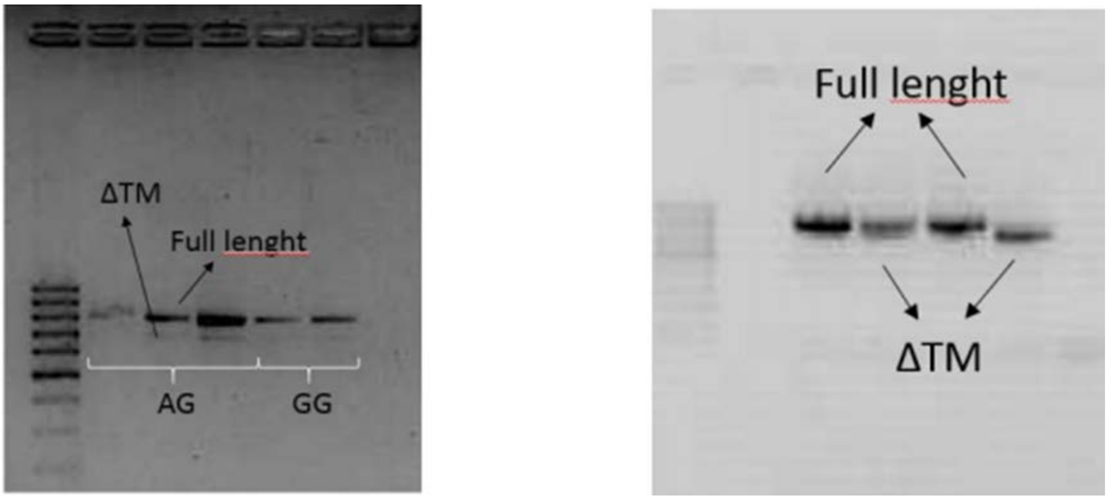


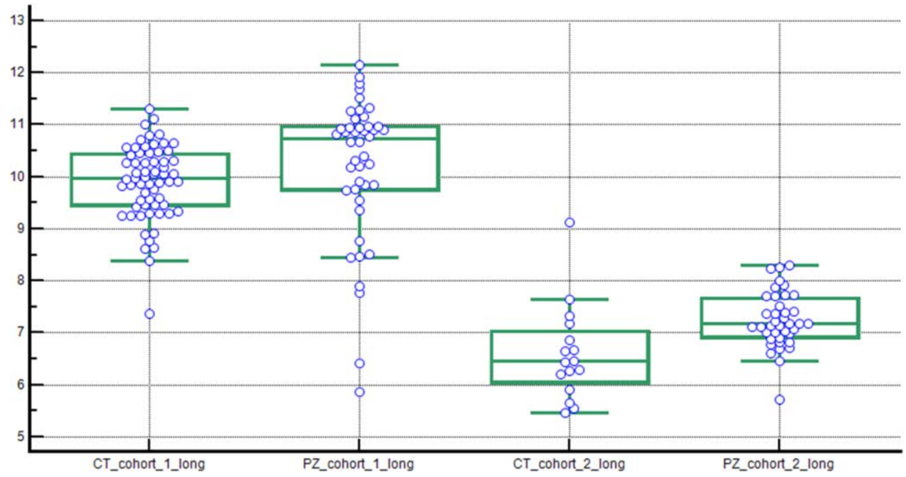
Figure 4 Electrophoresis on agarose gel 1.5% of RT-PCR of TNFSF14 cDNA . Left side: RT-PCR products of cDNA of TNFSF14. Right side: RT-PCR using the two TNFSF14 bands (extracted from the gel on the left side) as a template ".

During my project I have evaluated the TNFSF14 expression in RNA samples from two different cohorts .

The first cohort includes 64 healthy controls and 45 patients without drug treatment and their RNA was obtained from frozen pellets of PBMC (peripheral blood mononuclear cell), while for the second cohort, consisting of 16 controls and 39 patients also without drug treatment, the RNA was obtained by extraction from whole blood.

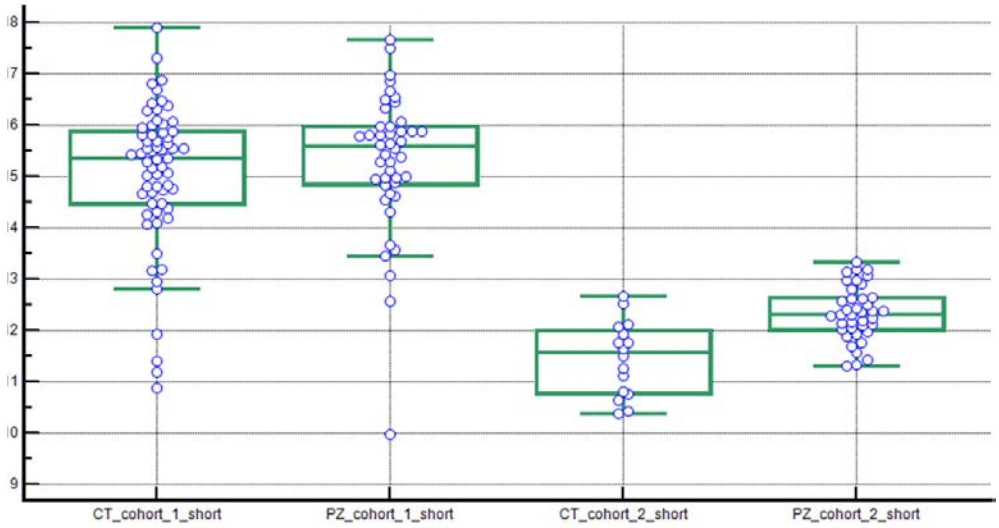
I quantified the expression of both isoforms of TNFSF14 (indicated as long and short isoform), because, since the two polymorphic variants of our interest are located in proximity of the site of alternative splicing, we wanted to verify whether they could affect the different splicing of TNFSF14. The RNA expression was measured by a relative quantification (using dCt method), and the patients showed a lower TNFSF14 expression in both cohorts and for both isoforms, in comparison with controls.

In particular, this difference was statistically significant both in the cohort 1 and 2 (respectively, $p = 0.0032$ and $p = 0.0009$, Mann-Whitney test with median values) for the long isoform (Figure 5), while for the short isoform the difference between patients and controls is statistically significant in cohort 2 ($p < 0.0001$) (figure 6) but the same trend was not statistically significant in Cohort 1 (Figure 6). For both patients and controls, and for both isoforms, the TNFSF14 expression is higher in the second cohort compared to the first. These data can suggest that the expression in this gene could be higher in cell populations of the whole blood (cohort 2) compared to PBMC (cohort 1).



median	9.99	10.77	6.45	7.18
mean	9.91	10.25	6.60	7.26

Figure 5: Expression of long TNFSF14 isoform in controls (CT) and patients (PZ) of the two cohorts. In X-axis there is the group, in Y-axis the ΔCT, each dot represents an individual, the green line are the median and the standard deviation. Cohort1 p = 0.0032; Cohort2 p = 0.0009 (Mann-Whitney test)



median	15.34	15.55	11.56	12.39
mean	15.06	15.42	11.46	12.47

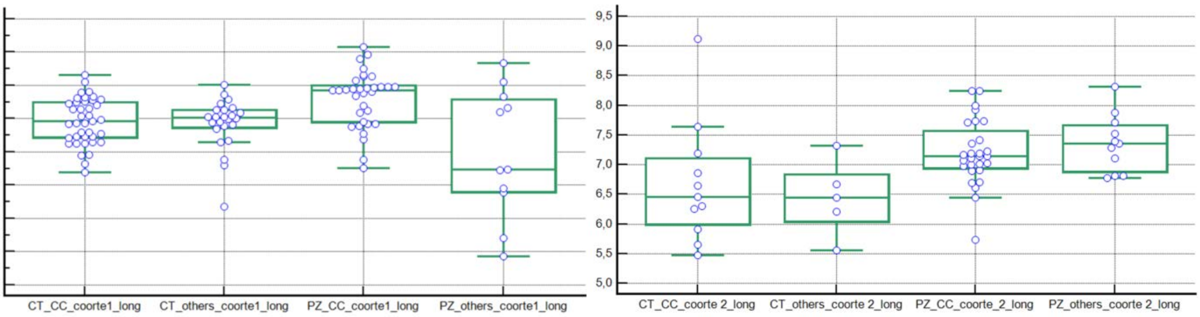
Figure 6: Expression of TNFSF14 short isoform in controls (CT) and patients (PZ) of the two cohorts. In X-axis there is the group, in Y-axis the ΔCT, each dot represents an individual, the green line are the median and the standard deviation (DS). Cohort 1 p = 0,1269; Cohort 2 p <0,0001, (Mann-Whitney test)

In order to further analyze the association data between expression levels and case-control status we performed a logistic regression on the merged sample set (161 individuals, including patients and controls from both cohorts). We took into account of the cohort heterogeneity by covariating for the cohort. The analysis was performed with SAS software. Higher ΔCT levels (thus lower expression levels) confer a significantly increased risk of multiple sclerosis with an odds ratio of 1.61 (p=0.015) for the long isoform and of 1.39 (p=0.031) for the short isoform.

isoform	Odds ratio	P value
long	1.61	0,015
short	1.39	0,031

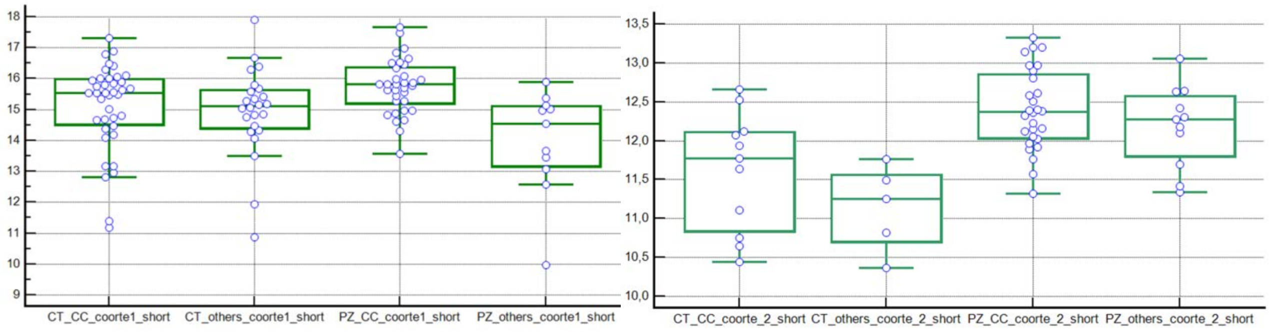
Table 6: logistic regression analysis on the whole sample set (n=161) between case-control status and expression levels.

In the second step of this analysis, I have used these expression data to perform a stratified analysis based on genotype of the two polymorphisms (rs1077667 and rs2291668). The stratification conducted revealed that the expression is lower in individuals homozygous for the risk allele (CC) of rs1077667 and for allele risk (GG) of rs2291668, for both patients or controls, for both cohorts and for both isoforms, except for the long isoform in patients of cohort 2 (right side in figure 7 and 9). This difference reached statistical significance in patients in the cohort 1 for the long isoform (p = 0.011, Mann-Whitney test, Figure 7) and for the short isoform (p = 0.0016, Mann-Whitney test, Figure 8) stratified for the intronic SNP. We observed similar results for patients in the cohort 1 stratified for the exonic SNP (rs2291668) (short isoform: p = 0.0006; long isoform: p = 0.016, Mann-Whitney test, Figures 9,10).



median	9.96	10.00	10.86	8.47	6.46	6.44	7.14	7.36
mean	9.91	9.85	10.60	8.99	6.68	6.44	7.20	7.36

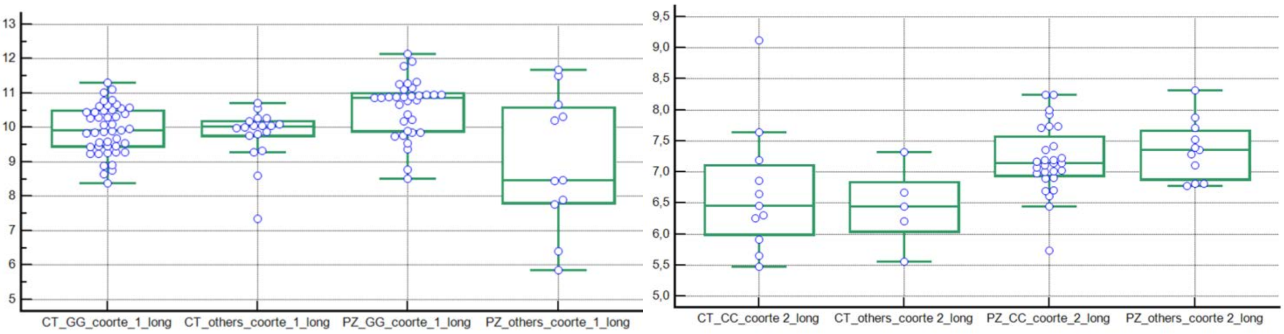
Figure 7: Expression data of the long isoform in controls (CT) and patients (PZ) stratified by rs1077667genotypes for Cohort 1 (Left side) and Cohort 2 (right side). A comparison between individuals homozygous for C allele and individuals with the other genotypes (CT and TT) was performed. . X-axis shows the group, Y-axis the ΔCT, each dot represents an individual, the green lines are the median and standard deviation. Cohort 1: Patients p = 0.0011; Controls p = 0,9 (Mann-Whitney test); Cohort 2: Patients p =0,4; Controls p = 0,8 (Mann-Whitney test)



median	15,55	15,11	15,82	14,54	11,77	11,25	12,37	12,28

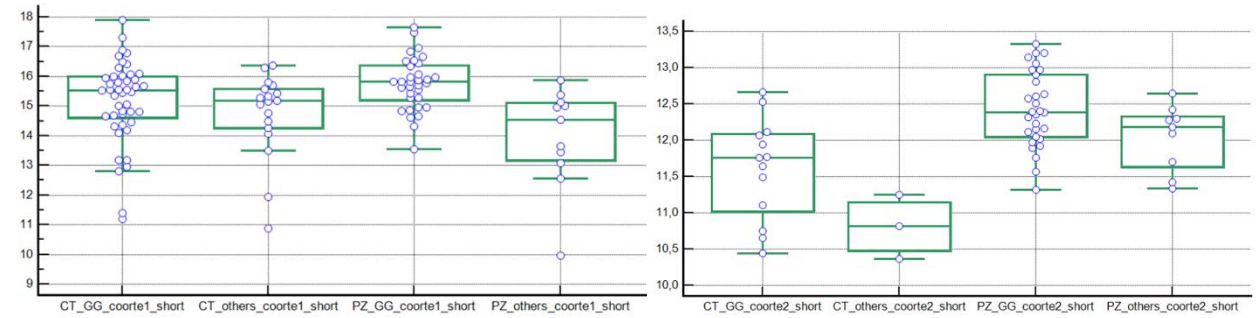
mean	15,17	14,73	15,75	13,97	10,65	10,82	12,4	11,42

Figure 8 : Expression data of the short isoform in controls (CT) and patients (PZ) stratified by rs1077667 genotypes for Cohort 1 (left side) and Cohort 2 (right side). A comparison between individuals homozygous for C allele and individuals with the other genotypes (CT and TT) was performed. . X-axis shows the group, Y-axis the ΔCT, each dot represents an individual, the green lines are the median and standard deviation. Cohort 1: Patients p = 0.0007; Controls p = 0,35 (Mann-Whitney test); Cohort 2: Patients p = 0.22; Controls p = 0,42 (Mann-Whitney test)



median	9,92	10,02	10,86	8,47	6.46	6.44	7,15	7,29
mean	9,94	9,79	10,59	9,02	6.68	6.44	7,23	7,28

Figure 8: Expression data of the long isoform in controls (CT) and patients (PZ) stratified by rs2291668 genotypes for Cohort 1 (left side) and Cohort 2 (right side). A comparison between individuals homozygous for G allele and individuals with the other genotypes (GA and AA) was performed. X-axis shows the group, Y-axis the ΔCT, each dot represents an individual, the green lines are the median and standard deviation. Cohort 1 Patients p = 0,0165; Controls p = 0,66 (Mann-Whitney test) Cohort 2. Patients p = 0,90; Controls p = 0,36 (Mann-Whitney test)



median	15,53	15,17	15,82	14,54	11,77	11,25	12,38	12,18
mean	15,17	14,73	15,75	13,97	10,65	10,82	12,43	12,04

Figure 10: Expression data of the short isoform in controls (CT) and patients (PZ) stratified by rs2291668 genotypes for Cohort 1 (left side) and Cohort 2 (right side). A comparison between individuals homozygous for G allele and individuals with the other genotypes (GA and AA) was performed. X-axis shows the group, Y-axis the ΔCT, each dot represents an individual, the green lines are the median and standard deviation. Cohort 1 Patients p = 0,006; Controls p = 0,20 (Mann-Whitney test) .Cohort 2. Patients p = 0,11 Controls p = 0,95 (Mann-Whitney test)

In order to merge together the different cohorts stratified for genotype and to take into account all the various variables and confounders we performed two linear regression analysis on the whole sample set (n=161, including both patients and controls from both cohorts):

1) we tested for association between expression levels and genotype assuming an additive model and using sex, cohort, and case-control status as covariates (table 7);

2) we tested for association between expression levels and case-control status using genotype, sex and cohort as covariates(table 8). This analysis was performed with PLINK software.

Regarding the association with genotype (table 7), we observed that at the increase in the number (0, 1 or 2) of minor (protective) alleles the Δct value significantly decreased (beta minor than 1) for both SNPs and with both isoforms. This means that at the increase in the number of susceptibility (common) alleles the expression levels significantly decrease.

Regarding the association with the disease status (table 8), we observed that Δct values are significantly higher in MS patients (thus expression levels are significantly lower) for both isoforms and when covariating either for rs1077667 or for rs2291668. The significance levels are similar to those obtained in the logistic analysis performed without covariating for sex and genotype.

SNP	allele	isoform	BETA	STAT	P
rs1077667	T	long	-0.3872	-2.885	**0.004464**
	T	short	-0.5347	-3.135	**0.002057**
rs2291668	A	long	-0.4097	-2.837	**0.005165**
	A	short	-0.7273	-4.042	**8.29*10-5**

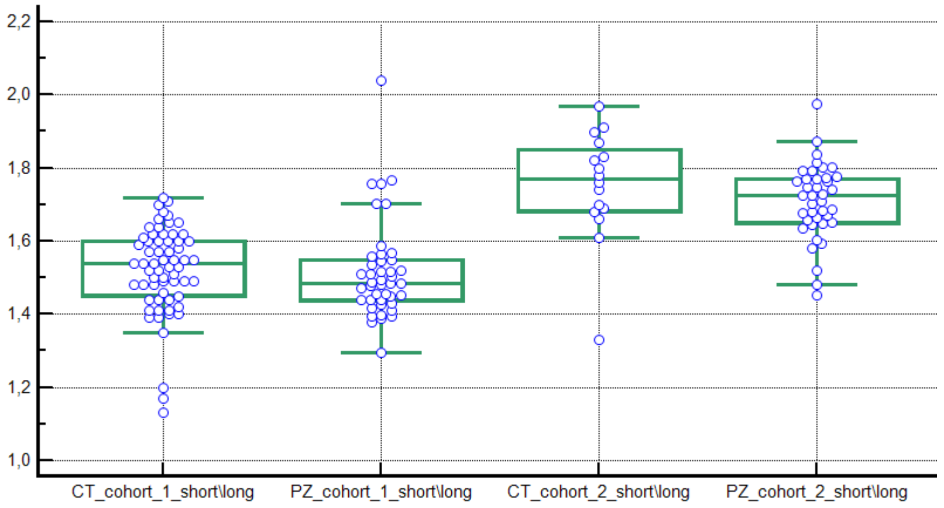
Table 7: linear regression analysis on the whole sample set (n=161) between Δct values and genotype covariated for sex, disease status and cohort.

isoform	conditioned for SNP	BETA	STAT	P
long	rs1077667	0.3906	2.573	**0.011**
	rs2291668	0.413	2.724	**0.007189**
short	rs1077667	0.4391	2.276	**0.02421**
	rs2291668	0.4679	2.477	**0.01432**

Table 8: linear regression analysison the whole sample set (n=161) between Δct values and disease status covariated for sex, genotype and cohort.

In addition, to compare the level of expression of the two isoforms, I evaluated the level of expression of the ratio of the two isoforms to test the hypothesis that the two polymorphisms in the study may influence the alternative splicing.

However, there were not significant differences in the ratio of the short isoform and the long isoform in controls and patients in cohort 1 (mean ΔCT controls = 1.52, mean ΔCT patients = 1.51) as well as in Cohort 2 (mean ΔCT controls = 1.75, mean ΔCT patients = 1.71) (Figure 11).



median	1,54	1,49	1,77	1,72
mean	1,52	1,51	1,75	1,71

Figure 9: the expression ratio between short and long isoforms in controls (CT) and patients (PZ) in the two cohorts. In X-axis there is each group, in Y-axis there are the dct value of this ratio, each dot represents an individual, the green lines are the median and standard deviation

In the light of these results I drew a scatter plot in order to calculate the coefficient of correlation between the short and long isoform (Figure 12).

We observed that the two variables are positively associated with high correlation coefficient

(r = 0.89, p <0.0001), in fact an individual with high levels of long isoform expression also expresses high levels of short isoform, as well as a low expression of one isoform is associated with a low expression of the other one.

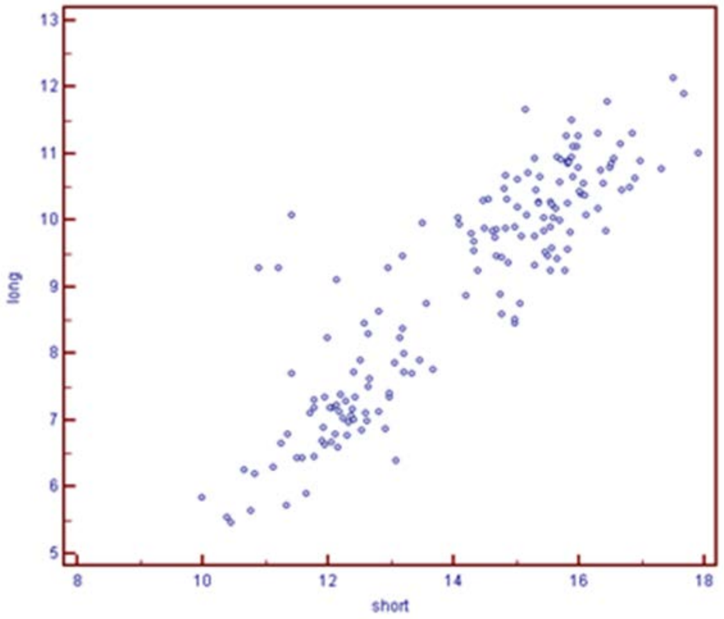
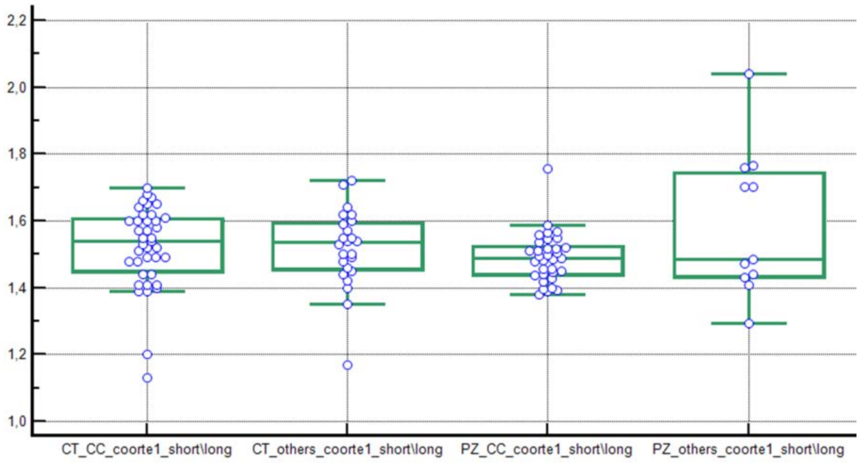


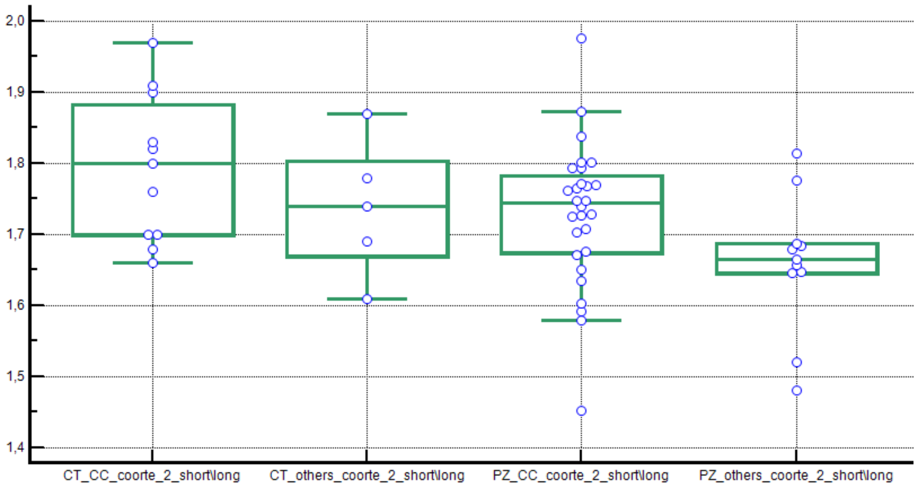
Figure 10: Scatter plot showing the correlation between the two isoforms derived from TNFSF14 alternative splicing. In the X-axis there are values of ΔCT for short isoform, in the Y-axis there are those for long isoform, each dot corresponds to an individual. Coefficient of correlation: r = 0.89, p <0.0001

Finally, I evaluated the relationship between the two isoforms stratifying the samples on the basis ofgenotypes of the two polymorphism (homozygous for the allele for susceptibility vs the others). None of the differences appears to be significant, although we observed that the individuals in the second cohort (both patients and controls) which are homozygous for the risk allele, have higher levels of the ratio between the two isoforms (Figures 13-16).



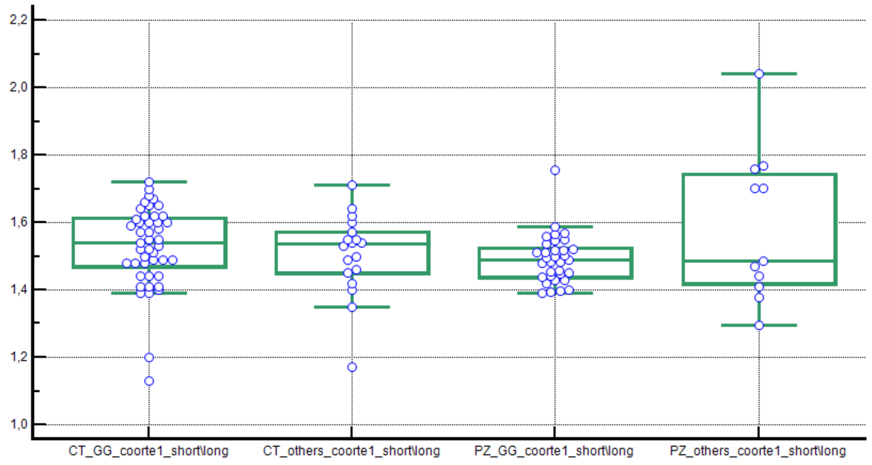
median	1,54	1,54	1,50	1,59
mean	1,52	1,52	1,49	1,59

Figure 11: ratio between short isoform and long isoform in controls and patients of Cohort 1 stratified by genotype for rs1077667. In X-axis there is the group, in Y-axis the ΔCT, each dot represents an individual, the green lines are the mean and standard variation. Controls: p = 0,75;patients :p = 0,4 (Mann Whitney test)



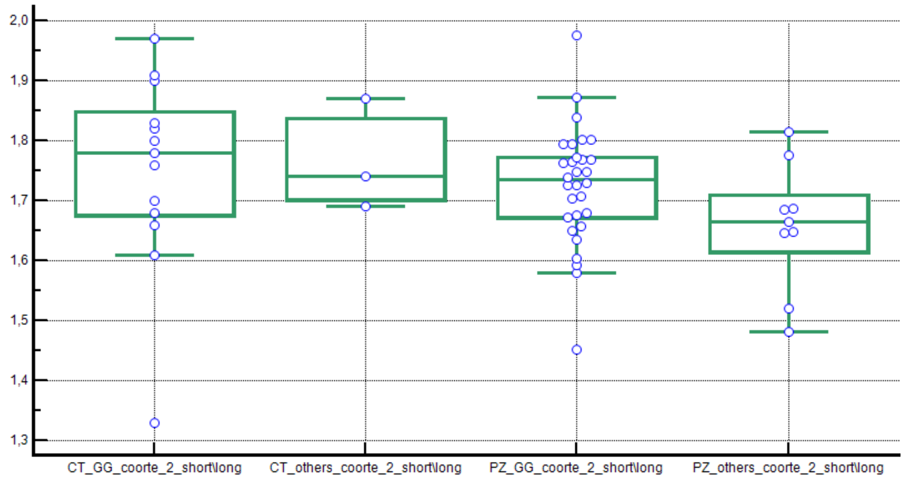
median	1,80	1,75	1,76	1,68
mean	1,76	1,74	1,73	1,66

Figure 12: ratio between short isoform and long isoform in controls and patients of Cohort 2 stratified by genotype for rs1077667. In X-axis there is the group, in Y-axis the ΔCT, each dot represents an individual, the green lines are the mean and standard variation. Controls: p = 0,33;patients :p = 0,06 (Mann Whitney test)



median	1,57	1,57	1,50	1,50
mean	1,53	1,51	1,49	1,59

Figure 13: ratio between short isoform and long isoform in controls and patients of Cohort 1 stratified by genotype for rs2291668. In X-axis there is the group, in Y-axis the ΔCT, each dot represents an individual, the green lines are the mean and standard variation. Controls: p = 0,48;patients: p = 0,53 (Mann Whitney test)



median	1,78	1,74	1,74	1,68
mean	1,76	1,74	1,72	1,66

Figure 14: ratio between short isoform and long isoform in controls and patients of Cohort 2 stratified by genotype for rs2291668. In X-axis there is the group, in Y-axis the ΔCT, each dot represents an individual, the green lines are the mean and standard variation. Controls: p = 1;patients: p = 0,12 (Mann Whitney test)

4 Discussion and conclusion

Multiple sclerosis is a multifactorial autoimmune disease with an aetiology still poorely known. For this reason, it is important to study genetic basis of the disease in order to know better its pathogenesis and to design targeted therapies. In these last few years genome wide association studies have led to a better understanding of the genetic basis of disease with the discovery of 103 loci involved in the susceptibility of MS in addition to the already known HLA region. Our genetics laboratory took part in three international genome wide studies (GWAS, 2011 and Immunochip 2013 and 2015) and conducted the studies on the Italian population in order to identify MS susceptibility variants specific for our population. In particular, our studies are focused in the identification of genetic variants of functionally responsible for the until now known associations . In detail, we focused our attention on the region of chromosome 19, which showed a very high association (p = 5.8x10-8) in the Italian population. Through a massive sequencing of the region and a fine mapping we was able toidentify, as primarily associated variant, a SNP (Single Nucleotide Polymorphism) (rs1077667) located in 1 intron of the gene TNFSF14, already observed in the International GWAS in 2011. It was also identified a new synonymous variant (rs2291668) in linkage disequilibrium with the intronic variant (r2 = 0.808), located in exon 1, near to the splicing site and not present in the genotyping platforms. To understand which is the primarily associated variant in this region, we genotyped 3357 samples (1680 healthy controls and 1677 patients) for the two polymorphism and we found that the association of exonic variant is not statistically significant after conditioning analysis with the intronic polymorphism. At the same time, we genotyped 2664 samples (1214 patients and 1450 controls) for a short tandem repeat (CA) which is present in the same intron as the rs1077667. In fact,in literature it is reported that STR variations contribute to splicing in humans (Thomas Willems et al., 2014). Furthermore, this STR together with the two SNPs mapin a region identified as active promoter (UCSC genome browser) and for this reason it could play a role in the gene regulation. Our analysis demonstrated that the intronic variant rs1077667 is primarily associated with MS while the association with exonic variant rs2291668 and the alleles of the microsatellite are only the consequence of the linkage disequilibrium with the intronic variant. The intronic SNP is that primarily associated among all 172 polymorphisms identified in this region, which include both those described in databases and those emerged from the sequencing of our region performed in a large group of patients and controls. Secondly, we conducted an analysis of gene expression in TNFSF14 with the purpose of starting to define a possible link between genetic association and gene function. Our results show that in general the patients express a lower level of this gene, while the controls show higher expression levels, although not always this difference is statistically significant. Stratifying our series for the genotype of rs1077667 and rs2291668 we observe a trend of expression that is replicated by all groups, although not all comparisons are statistically significant. In particular, we observe that people with risk genotype (GG) produce lower levels of TNFSF14 than the other and that, between patients and controls with this genotype, patients always are the minor producers. We did not detect significant differences in expression between the two isoforms: an individual who produces a low amount of TNFSF14 produces both a low amount of both short and long isoform. In the future, itwill be necessary to increase the number of samples for gene expression analysis in order to increase the statistical power. In addition, another point will be to determine the expression of this gene in different cell populations from peripheral whole blood. In fact our data of gene expression performed in the second cohort with RNA extracted from peripheral whole blood have shown different level of gene expression compared to samples with RNA extracted from peripheral blood lymphocytes. In addition, in silico data have shown that this intronic variant can modify the binding of the AhR transcription factor. In fact the risk allele (G) allows the binding while the other allele does not allow it. AhR was initially discovered and well characterized as a transcription factor responsible for the activation of genes encoding different enzymes to the metabolism of xenobiotics (Vogel et al.2014). Recent studies also indicate that activation of AhR plays different roles in cellular functions, including the regulation of the immune system (Singh et al., 2007), for example, it is involved in the differentiation of T-lymphocytes, in particular in regulatory T cells and T helper 17 (Kimura et al., 2008). Recently it has been also shown that 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), so it is interesting that the AhR transcription factor is involved in the differentiation of these lymphocytes. Inflammation is one of the responsible mechanisms of multiple sclerosis , at least in the early stages (Frischer et al., 2009). It was seen that molecules such as NR4A2, with a key role in protecting neurons from neurotoxicity induced by inflammation through the signaling of NF-kB pathway, have a reduced gene expression in peripheral blood of patients (Navone et al., 2014). In the light of these results, we can suppose that also TNFSF14 can play a similar role. This glycoprotein in particular contexts, is able to induce, through the link with LTβR, cell death due to the recruitment of TRAF 3 and the activation of caspases (Granger et al., 2003). Also Shui et al., in 2011, have noted that although the interaction between TNSFSF14 and its receptor HVEM has an costimulatory effect when it interacts with HVEM BTLA has opposite functions which led to inhibition in the activation of T lymphocytes. The data of our study are still very preliminary to define which is the precise role of TNFSF14 in the predisposition to MS. They seem to suggest that a reduced expression of TNFSF14, caused by genetic variants in a regulatory region in the first intron potentially involved in the recognition of the transcription factor AhR, can predispose to the disease. The genetic risk allele is in fact associated with a reduced expression of TNFSF14, and in general, patients have an expression of TNFSF14 significantly reduced compared with healthy controls. This model will require further studies to be confirmed.

Bibliography

Compston Alastair, Alasdair Coles, Multiple sclerosis. Department of Clinical Neurosciences, University of Cambridge Clinical School, Addenbrooke's Hospital, Cambridge, UK. Lancet 2008; 372: 1502–17

Durelli, L., Conti, L., Clerico, M., Boselli, D., Contessa, G., Ripellino, P., Ferrero, B., Eid, P. and Novelli, F., *T-helper 17 cells expand in multiple sclerosis and are inhibited by interferon-ß. Ann Neurol*., 2009; 65: 499–509. doi: 10.1002/ana.21652 Frischeret al., 2009

Ernst J, Kellis M., Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol. 2010 Aug;28(8):817-25

Granger et Rickert, *LIGHT-HVEM signaling and the regulation of T-cell-mediated immunity*, Cytokine Growth Factor Reviews, 2003

Granger SW, Butrovich KD, Houshmand P, Edwards WR, Ware CF*, Genomic characterization of LIGHT reveals linkage to an immune response locus on chromosome 19p13.3 and distinct isoforms generated by alternate splicing or proteolysis*, The Journal of Immunology, 2001;167(9):5122-8.
Granieri E., Exogeneous factors in the aetiology of multiple sclerosis. J Neurovirol 2000; 6 Suppl 2: S141-6

Hansen, T., A. Skytthe, Stenager E., Peterson HC., Brannum-Hansen H., Kyvik KO., *Concordance for multiple sclerosis in Danish twins: an update of a nationwide study*. MultScler 2005; 11(5): 504-10.

IMSGC. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat Genet. 2013 Nov;45(11):1353-60.

Kimura A. Naka T, Nohara K, Fujii-Kuriyama Y, Kishimoto T., *Aryl hydrocarbon receptor regulates Stat1 activation and participates in the development of Th17 cells*, Proc Natl AcadSci U S A. 2008 Jul 15;105(28):9721-6

Lincoln, M. R., Montpetit A., Cader MZ., Saarela J., Dyment DA, Tislar M., et al. *A predominant role for the HLA class II region in the association of the MHC region with multiple sclerosis*. Nat Genet 2005; 37(10): 1108-12

McDonald WI1, Compston A, Edan G, Goodkin D, Hartung HP, Lublin FD, McFarland HF, Paty DW, Polman CH, Reingold SC, Sandberg-Wollheim M, Sibley W, Thompson A, van den Noort S, Weinshenker BY, WolinskyJS.Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis.*Ann Neurol.* 2001;50:121-7.

Mumford, C. J., N. W. Wood, et al. *The British Isles survey of multiple sclerosis in twins*. Neurology 1994; 44(1): 11-5.

Navone ND, Perga S, Martire S, Berchialla P, Malucchi S, Bertolotto A., *Monocytes and CD4+ T cells contribution to the under-expression of NR4A2 and TNFAIP3 genes in patients with multiple sclerosis*., J Neuroimmunol. 2014 Jul 15;272(1-2):99-102

Oksenberg, J. R., L. F. Barcellos, et al. *Mapping multiple sclerosis susceptibility to the HLADR locus in African Americans*. Am J Hum Genet 2004; 74(1): 160-7

Pugliatti M., S. Sotgiu, Solinas G., Castiglia P., Pirastru MI., Murgia B., Mannu L., Sanna G., Rosati G., Multiple sclerosisepidemiology in Sardinia: evidence for a trueincreasingrisk. ActaNeurolScand 2001; 103(1): 20-6

Purcell S, Neale B, Todd-Brown K, et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet*. 2007; 81: 559–575.

Shui JW, Steinberg MW, Ware CF, Kronenberg M. Regulating the mucosal immune system: the contrasting roles of LIGHT, HVEM, and their various partners SeminImmunopathol. 2009 Jul;31(2):207-21.

Singh NP, Hegde VL, Hofseth LJ, Nagarkatti M, Nagarkatti P, *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*, MolPharmacol. 2007 Dec;72(6):1508-21

The International Multiple Sclerosis Genetic Consortium & the Wellcome Trust Case Control *Consortium Genetic risk and primary role for cell-mediated immune mechanisms in multiple sclerosis*. Nature 2011; 476:214-218

The International Multiple Sclerosis Genetics Consortium; International IBD Genetics Consortium (IIBDGC); Class II HLA interactions modulate genetic risk for multiple sclerosis. International BD Genetics Consortium IIBDGC. Nat Genet. 2015 Sep 7. doi: 10.1038/ng.3395.

Totaro R., C. Marini, Cialfi A., Giunta M., Carolei A., Prevalence of multiple sclerosis in the L'Aquila district, central Italy. J NeurolNeurosurg Psychiatry 2000; 68(3): 349-52

Vogel, C. Khan EM, Leung PS, Gershwin ME, Chang WL, Wu D, Haarmann-Stemmann T, Hoffmann A, Denison MS., *Cross-talk between aryl hydrocarbon receptor and the inflammatory response: a role for nuclear factor-κB,*, J BiolChem 2014; 289 (3) 1866-75

Weissert Robert, The Immune Pathogenesis of Multiple Sclerosis, J NeuroimmunePharmacol (2013) 8:857–866

Willems T, Gymrek M, Highnam G; 1000 Genomes Project Consortium, Mittelman D, Erlich Y.et al., The landscape of human STR variation. Genome Res 2014; 24(11):1894-904

Willer, C. J., D. A. Dyment, et al. *Twin concordance and sibling recurrence rates in multiple sclerosis*. Proc Natl AcadSci U S A, 2003: 100(22): 12877-82

Seminari e congressi seguiti:

Lesson of Prof Prat: " Stem cell in the regeneration and repair of the tissues and organs"

IL LUPUS ERITEMATOSO SISTEMICO la ricerca IRCAD nel contesto internazionale- 22 Novembre 2014 Sala del Consiglio della Provincia di Novara Piazza Matteotti 1, Novara

“Nuove sfide ed opportunità dell'epidemiologia molecolare per lo studio dei tumori"” (Prof. Laura Bagliettodell'Inserm - Centre for Research in Epidemiology and PopulationHealth,)

“Humoral responses to HCV infection and clinical outcomes” (Arvind Patel, PhD Programme Leader MRC Centre for Virus Research, University of Glasgow)

“Regulation of hepatocytes differentiation during the transitions betweenepithelial and mesenchymal states”. (DrToninoAlonzi)

“Targeting the liver to cure myocarditis: a lesson from a model ofSTAT3-dependent auto-immune myocarditis” (Prof Valeria Poli)

Lesson of Prof. Antonio Sica: "Myeloid cells as therapeutic target in cancer"

Seminar in the framework of "Regenerative Medicine programme:"From the legend of Prometheus to regenerative medicine": Main Lecture Hall Prof. Dr Yong-Sang Song

Lesson of professor Antonio Sica: “Myeloid cells as therapeutic target in cancer ”

Partecipazione al congresso MOLECULAR MECHANISMS OF NEURODEGENERATION (Milan, May 28th-30th, 2015)
(presentazione dati come poster dal titolo: Genome-wide analysis of DNA tandem repeats in ALS: development of a NGS method
Lucia Corrado[#*1], Roberta Bordoni[#2], Loredana M. Genovese[#3], Eleonora Mangano[2], Filippo Geraci[3], Marco Severgnini[2],Romina D’Aurizio[3], Clarissa Locci[1], Miriam Zuccalà[1], Letizia Mazzini[4], Alfredo Brusco[5], Giovanni Manzini[3, 6], Marco Pellegrini[#3], Gianluca De Bellis[#2], Sandra D’Alfonso[#1])

seminario di Nefrologia tenuto dal Dott. Vincenzo Cantaluppi dal titolo: "Le cellule staminali nel danno renale acuto e nel trapianto di rene"

Cell based models for studying molecular mechanism of Facioscapulohumeral Muscular Dystrophy (FSHD)
Speaker: Prof. Darko Boshnakovski, PhD,University "Goce Delcev" Stip, Faculty of Medical Sciences,
Stip, R. Macedonia

“Recent Developments in (cutaneous) Human Polyomavirus Research ”(MarietFelktamp scheduled on
Assoc Prof of Medical VirologyDept of Medical Microbiology,Leiden University Medical Center)