UNIVERSITY OF EASTERN PIEDMONT

DEPARTMENT OF HEALTH SCIENCES

Phd Program In Medical Sciences And Biotechnology

XXXVII Cycle

MASSIVELY PARALLEL REPORTER ASSAY (MPRA) COMBINED WITH BAYESIAN FINE MAPPING APPROACHES PRIORITIZE VARIANTS ASSOCIATED WITH MULTIPLE SCLEROSIS (MS) FROM GENOME WIDE ASSOCIATION STUDIES (GWAS) WITH A POTENTIAL FUNCTIONAL ROLE AND INTERACTION WITH ENVIRONMENTAL FACTORS.

PhD Candidate: **Endri Visha**

TUTOR: **PROF.SSA SANDRA D'ALFONSO**

Academic Year: 2023/2024

SSD: MED/03

DICHIARAZIONE E AUTORIZZAZIONE AL RILEVAMENTO ANTIPLAGIO

(Art. 47 e Art. 38 del D.P.R. 28 dicembre 2000, n. 445) e s.m.i. esente da bollo ai sensi dell'art. 37 D.P.R. 445/2000 e s.m.i.

Il/La sottoscritto/a Endri Visha nato/a a Tirana Albania prov (AL) il 1/9/96, regolarmente iscritto/a per l'a.a. 2021 al 2024 anno del Corso di Dottorato di ricerca Scienze e Biotecnologie Mediche, XXXVII Ciclo

consapevole del fatto che l'Università ha adottato un servizio *web-based* per rilevare il plagio, attraverso un sistema software chiamato "Turnit.in",

consapevole delle sanzioni penali previste in caso di dichiarazioni non veritiere e di falsità negli atti e della conseguente decadenza dei benefici di cui agli artt. 75 e 76 del D.P.R. 445/2000 e s.m.i.,

DICHIARA

- di aver sottoposto la propria tesi di dottorato alla scansione di Turnit.in.

Ritiene che, dalla verifica, la Tesi è ragionevolmente risultata un documento originale, che cita correttamente la letteratura.
È infine consapevole che la propria tesi di dottorato può essere verificata dal proprio Supervisore e/o dal Coordinatore del Corso di dottorato per confermarne l'originalità.

Novara 10/06/2025 Endri Visha

(luogo e data) (firma)

INDEX

SUMMARY

Multiple Sclerosis (MS) is a complex autoimmune disease affecting the central nervous system (CNS) which leads to a chronic inflammatory condition that causes demyelination and neuronal damage, along with neurological impairment and disability. The underlying etiology of MS is still unknown but is thought to be related to an interplay of genetic susceptibility and environmental factors. Several factors have been investigated as the causes of the disease. Meta-analyses suggest that, among non-genetic factors, the strongest evidence of association is related to Epstein-Barr virus biomarker positivity, infectious mononucleosis, and smoking (Belbasis L. et al. 2015). The presence of a genetic component in the pathogenesis of Multiple Sclerosis is supported by the observation of familial clustering and the higher prevalence of MS in certain ethnic populations. Studies involving twins from different populations consistently demonstrate that the risk of MS is higher in monozygotic twins (25-30% concordance) of individuals with MS compared to dizygotic twins (2-5%). (Sadovnick et al., 2004)

There have been important advancements in the recent years in the field of MS shifting interest to other genetic factors associated with the disease apart from regions of the MHC. The HLA (Human Leukocyte Antigens) class II and I genes are particularly relevant modifiers of the disease risk. The development of Genome-Wide Association Studies (GWAS) allowed the simultaneous identification of hundreds of thousands of SNPs, spaced across the entire genome for the association with a particular trait in case-control datasets composed of genetically unrelated individuals, giving new insights into genetic variants that contribute to diseases. International studies analyzing large datasets at the genome-wide level have identified 200 loci involved in the susceptibility of MS in addition to the HLA region. These discoveries were mainly due to the contribution of three international studies in 2011 by the International Multiple Sclerosis Genetics Consortium (IMSGC), 2013 (IMSGC, 2013), and 2019 (IMSGC, 2019) of which our laboratory has been partaker. The 2019 GWAS study (IMSGC, 2019) increased the number of statistically independent associations with MS susceptibility to 233. They identified 200 risk loci in the autosomal non-major histocompatibility complex (non-MHC) associated with susceptibility to the disease. GWAS studies carry some intrinsic limits represented by Linkage Disequilibrium (LD) which don't allow the pinpointing of the real causative variants in highly complex regions. Pinpointing causal variants in MS-associated regions that contain drug target genes (genes encoding for proteins targeted by an already approved drug or by a molecule currently tested in clinical trials), could potentially lead to the repurposing of already known drugs, whose target pathogenic mechanism may overlap with MS ones. The combination of fine mapping with functional techniques such as Massively Parallel Reporter Assay (MPRA) would resolve this problem since the latter is a high-throughput screening method able to test thousands of sequences for their putative transcription regulation role within a single assay. This allows for identifying variants that modulate gene expression among Multiple Sclerosis drug target-associated loci derived from GWAS.

Evidence from epidemiological studies conducted recently that report a 32-fold increased risk of developing MS in individuals who converted to Epstein Barr (EBV) seropositive compared to individuals who remained seronegative highlight the importance and involvement of Epstein Barr Virus to Multiple Sclerosis. MPRA could serve as a vessel to measure differences in reporter gene expression influenced by SNPs associated with MS when exposed to environmental factors associated with MS such as Epstein Barr Virus. Notably, Harley et al., 2018 have proven the preferential binding of the Epstein-Barr Nuclear Antigen 2 (EBNA2) to MS risk loci in the presence of different alleles proving its involvement in gene expression. Similarly, Keane et al., 2021 were able to demonstrate by using an allele-specific Chip assay that EBNA2 binds to MS risk loci in an allele-dependent manner in LCLs, by using the EBNA2 inhibitor EBNA2-TAT. Published work from our collaborators Mechelli et al., 2015 have studied five major alleles of the EBV type 1 strain, the most frequent strain in the Caucasian population, which were identified based on the nucleotide variation within the most variable region of EBNA2. Specifically, they showed that the MS risk significantly correlates with an excess of the 1.2 allele of the EBNA2 gene (odds ratio (OR) =5.13; 95% confidence interval (CI) 1.84-14.32; p=0.016). This data fuels another work hypothesis, to test the changes in the effect of functional SNPs associated with MS in regions, that are druggable, when exposed to disease-related environmental factors, such as in this case the Epstein-Barr EBNA2 variant 1.2. Using the basis of the MPRA technique which can test the putative transcriptional regulatory role of a large number of variants delivering a quantitative value of transcription regulation changes in the presence of different alleles for each variant, we can measure statistically significant differences in the presence of the Epstein-Barr EBNA2 variant 1.2.

Initially, we applied a Bayesian Fine Mapping with the final aim to identify functional variants in complex MS-associated regions. We started from a large cohort of 5,903 individuals from the continental Italian Population, including 5,259 MS patients and 1,644 healthy controls, with 6,339,414 imputed SNPs covering the entire genome. We wanted to identify potential drug target genes within the regions associated with MS, so we cross-mapped replicated MS regions (statistically significant with a p-value <0.05, 2Mb-wide) with genes listed in the Drug-Gene interaction database v4.2. This analysis yielded 36 regions surrounding the replicated SNPs, which contained 238 druggable genes. Then we applied Paintor and Caviar BF, which both utilize a Bayesian framework, which assigns a posterior probability of causality, known also as posterior inclusion probability (PIP). For a more accurate functionally informed fine mapping, we included a set of annotation scores, to take into account the biological relevance of each SNP within the region. For this purpose, we incorporated GWAVA, CADD, FINSURF, and Regulome Db annotation scores into the fine mapping analysis. Once we identified the SNPs in the credible sets, we had to determine which genes they potentially influenced. For this, we conducted SNP-to-gene mapping using the Open Targets Genetics

Database. We obtained 19 regions that contain SNPs with compelling evidence of causality with statistical significance (Paintor PIP>0.7). Among these 19 regions, 18 displayed as a possible causal SNP a different one than the Lead SNP that is the SNP showing the most statistically significant association in the original IMSGC article. Among these, 10 regions have causal SNPs that target drug-target genes. *IFNGR1, TUBB4A, TEC, TGFBR3, IDE, and CD40* are the genes that are targets of drugs approved for other diseases but also of the SNPs that are shown as "causal". Additionally, in four regions the "causal" SNPs target potential drug target genes (*EOMES, FCRL3, CTSH,* and *ADCY3*) according to Open Targets Genetics. Simultaneously with the fine mapping analysis, we performed the MPRA experiment involving 5 out of the 36 MS-associated regions which contained druggable regions, specifically the regions around the GWAS most associated lead SNP were the regions of CD40, PRDX5, TEC-TKX, IFNGR2, CHRNA9. We selected these 5 regions due to their architectural complexity due to the large number of SNPs in high linkage disequilibrium with the lead SNP of the region. This analysis was performed independently from the statistical analysis, however after the fine mapping analysis we confirmed two regions that were also within the credible set predicted by fine mapping, were also selected for the MPRA analysis, respectively the CD40 region and the TEC region. The total number of SNPs tested across these 5 regions was 83, and the filtering criteria for a SNP to be considered in high LD with the Lead SNP was $r2 \geq 0.77$. MPRA enabled the simultaneous functional testing of numerous potential regulatory elements, using the basis of the conventional luciferase assay in a high throughput manner. The foundation of this technique consists of creating a library in which each tested variant is represented by an oligo sequence containing either the reference or alternative nucleotide in a region of 145 bases. The novelty instead lies in the fact that each sequence is represented by a 10 bp nucleotide sequence called a TAG that serves as a barcode to distinguish each oligo sequence. Each SNP to be tested is portrayed by a 145 base pair sequence containing either an Alternative or Reference allele and by a sequence with the same length that contains a 21 base pair segment surrounding the variant nucleotide, which serves as a null hypothesis sequence later on called Scramble. Each of either Reference, Alternative, or Scramble is reflected 10 times in both DNA strands, resulting in 60 representations for each SNP to be tested, leading to a total of 4980 probes each with a unique bar code. To test each of the variants, we utilized two distinct vectors, namely pMPRAdonor1 and pMPRAdonor2. Both constructs were created to contain the library of SNPs to be tested however; pMPRAdonor1 contained only the LUC gene Open Reading Frame (ORF) and was utilized the test the variant's effect as promoters. pMPRAdonor2 contains the LUC gene coupled with a minimal promoter, which will investigate the enhancer effect of each tested variant. Once ready, it was transfected to 2 different cell lines, respectively HEK293T for transfection testing and JURKAT cell lines as a disease-relevant cell line which was performed 2 times for experimental evaluation. Cells were transfected in quadruplicates for statistical

significance for both PMPRAdonor1 and pMPRAdonor2, they were left to express the construct for 48 hours before cell harvesting. We performed NGS using the Illumina DNA Prep and Nextera XT index kits on the cDNA obtained from the transfected cells and input DNA was used to compare the change in reads for each barcode post-transfection. Data was analyzed using the mpralm tool, which performs a ratio from DNA to RNA counts for each Reference and Alternative allele of each probe, illustrating this ratio by a LogFC associated with a P-value, which gives it significance. The significance filter for the p-value was at an FDR<0.01. We were able to identify SNPs that exert a significantly different impact on the expression of the Luc gene between the two allelic variants (Alternative vs. Reference sequence). However, we focused solely on the probes that demonstrated a significant deviation from the null sequence (scramble) when compared to both the Alternative and the Reference sequence. For added reassurance in our results, we decided to perform the experiment on the Jurkat cell line in duplicate at different times and then measure the outcomes. Following this security step only the probes that showed the same direction of effect and surpassing the filters applied to the mpralm outcome between both experiments were selected for further analysis and evaluation. We compared the Log FC as given by mpralm of both the experiments conducted on the Jurkat cell line and observed a high correlation (R2=0.8) between both experiments, suggesting a high reproducibility in our methodological approach. This analysis proved that when we replicated the experiment, the Log FC of the tested variants had a high correlation. Then to discard any probes that did not follow the same trend in both experiments we performed a simple linear regression test. The simple linear regression with a 95% confidence interval showed that 4 probes that were tested were not in the confidence interval, as they exhibited a significant difference in the log FC between the two Jurkat experiments. Respectively, the four probes, which show discrepancies, were not taken into account for further evaluation. We chose to assess only the variants that showed the same direction of effect in both experiments and with a statistically significant difference from the null hypothesis, which left us with 8 variants for follow-up. Along with this, further research using online databases such as RegulomeDB, the ENCODE project, Screen registry V3, and UCSC revealed that some of the variants that show the same direction of effect in both cell lines fall in regions of particular interest such as promoter like signatures, H3K36me3, proximal or distal enhancer signatures, etc. Then we performed transcription factor (TF) binding analysis using MotifBreakR, which predicts preferential binding of TF in the presence of different alleles. We performed MotifBreakR in combination with MEME suite and TomTom to predict at least one transcription factor that preferentially binds in the presence or absence of an allele of interest for 7 out of 8 MPRA significant variants. We then associated transcription factors to the risk allele of each SNP, to obtain TF that could explain the effect of the risk allele as measured in vitro to the reporter gene. This would possibly mimic what happens biologically even though further evaluation is most certainly required. After

having assessed which were the causative variants through MPRA, and then estimated the effect of the risk allele on the reporter gene, we linked these data by possible TF that bind in the presence of the risk allele. Then by using the Open Target Genetics, we were able to find the target gene of each MPRA significant SNP and associate the possible negative effect of the SNP to a gene, with a drug employing the opposing effect.

The same procedure when using the MPRA techniques is followed as a path to test the effect of known environmental factors associated with MS such as EBNA2 variant 1.2. For this, we performed a sequential transfection on the Jurkat 2 cell line, firstly with a construct that contains the EBNA2 1.2 variant, and after 24 hours with the MPRA construct to observe the effect in the expression of the sequences as measured by the reporter gene. The outcome of each SNP is measured by comparing tag counts for each probe after transcription, using the MPRAlm tool to compare RNA reads to input DNA. The resulting counts are utilized to quantify the activity of a given putative regulatory sequence. Then to explain why the effect of the SNP on the reporter gene had changed we applied MotifBreakR, which predicted several Transcription Factors (TF) that preferentially bind in the presence of either allele. We then selected TF that were known to be affected in the presence of viral agents. To assess the target gene of these SNPs we applied the Open Target Genetics tool, which can predict the target gene of the tested SNPs by showing the eQTL effect. Variants that were discarded from the replicability experiment on the Jurkat cell line mentioned before were not taken into account in this experiment either. When we applied our experimental pipeline, we performed a sequential transfection and each experiment was performed in quadruplicates to enhance statistical significance. Firstly we transfected the Jurkat cells with the pCDNA 1.2 plasmid which contains around 1602 bp of the EBNA2 variant coupled with the GFP sequence. In parallel as a positive control we transfected the same number of Jurkat cells with a basic GFP construct which is innocuous to the cells. Through mpralm, we obtained the Log Fc and P values corresponding to each tested SNP. Our hypothesis is based on this crucial step where we want to compare the LogFC measured by the mpralm tool for each variant tested between the control experiment and the experiment containing EBNA2. Therefore, we measured the effect on the expression of the reporter gene in the presence of the EBNA2 variant 1.2 for each reference or alternative allele, and compared it to the effect on the expression of the reporter gene in the presence of an innocuous GFP plasmid as a substitution for EBNA variant 1.2. To have statistical confidence in our results we compared the LogFC of each tested probe with a simple linear regression test between the experiment with the exposure to the pathogen and the control. We selected a 95% confidence interval, which revealed 8 probes with significant changes in the effect measured by the reporter gene when exposed to the EBNA variant. Out of the 8 variants, only 2 variants fulfilled the predetermined MPRA filters, which are crucial since they facilitate the finding of sequences tested that have a prominent and

robust effect linked to a significant p-value. We then focused on transcription factors that bind to the risk allele with a strong effect and are modulated by the presence of the EBNA2 protein. This evaluation revealed important information about the probable pathway to which these TF alter the effect of the MS-associated variants as measured by the reporter gene, explaining some of the noted changes in the presence of the EBNA2 variant 1.2 on the cell line. Altogether, these data support the efficiency of our methodology in examining functional variants associated with MS among non-functional ones. We were able to find the target genes of these variants, and a probable pathway by which they exert their function by predicting TF that binds to them when the risk allele is present. This allowed us to find possible drugs that wield an opposite effect on the gene to that of the MS-associated variant. Furthermore, we verified that using the MPRA as an estimator of the influence of environmental factors associated with MS; in this case, the EBNA2 variant 1.2 reveals promising results. Our project opens the possibility to apply the same approach in other regions characterized by high Linkage Disequilibrium associated with MS or other diseases, to pinpoint functional variants amongst others, and to perform an insilico prediction of the altered pathways in the presence of the SNP.

RIASSUNTO

La Sclerosi Multipla (SM) è una complessa malattia autoimmune che colpisce il sistema nervoso centrale (SNC) e porta a una condizione infiammatoria cronica che causa demielinizzazione e danni neuronali, oltre a deficit neurologici e disabilità. L'eziologia sottostante della SM è ancora sconosciuta, ma si ritiene sia correlata a un'interazione tra suscettibilità genetica e fattori ambientali. Diversi fattori sono stati indagati come possibili cause della malattia. Le meta-analisi suggeriscono che, tra i fattori non genetici, le prove più forti di associazione riguardano la positività ai biomarcatori del virus Epstein-Barr, la mononucleosi infettiva e il fumo (Belbasis L. et al., 2015). La presenza di una componente genetica nella patogenesi della Sclerosi Multipla è supportata dall'osservazione di raggruppamenti familiari e dalla maggiore prevalenza di SM in alcune popolazioni etniche. Studi su gemelli di diverse popolazioni dimostrano costantemente che il rischio di SM è maggiore nei gemelli monozigoti (25-30% di concordanza) di individui con SM rispetto ai gemelli dizigoti (2-5%) (Sadovnick et al., 2004).Negli ultimi anni sono stati fatti importanti progressi nel campo della SM, spostando l'interesse su altri fattori genetici associati alla malattia oltre alle regioni del complesso maggiore di istocompatibilità (MHC). I geni della classe II e I degli antigeni leucocitari umani (HLA, Human Leukocyte Antigens) sono particolarmente rilevanti come modificatori del rischio di malattia. Lo sviluppo degli studi di associazione su tutto il genoma (GWAS, Genome-Wide Association Studies) ha permesso l'identificazione simultanea di centinaia di migliaia di SNP (polimorfismi a singolo nucleotide) distribuiti su tutto il genoma per l'associazione con un particolare tratto in dataset caso-controllo composti da individui geneticamente non correlati, fornendo nuove intuizioni sulle varianti genetiche che contribuiscono alle malattie. Studi internazionali che analizzano ampi dataset a livello genomico hanno identificato 200 loci coinvolti nella suscettibilità alla SM oltre alla regione HLA. Queste scoperte sono state principalmente il risultato del contributo di tre studi internazionali condotti nel 2011 dal Consorzio Internazionale di Genetica della Sclerosi Multipla (IMSGC), nel 2013 (IMSGC, 2013) e nel 2019 (IMSGC, 2019), a cui il nostro laboratorio ha partecipato. Lo studio GWAS del 2019 (IMSGC, 2019) ha aumentato il numero di associazioni statisticamente indipendenti con la suscettibilità alla SM a 233. Sono stati identificati 200 loci di rischio nella regione non-MHC autosomica associati alla suscettibilità alla malattia. Gli studi GWAS presentano alcuni limiti intrinseci rappresentati dal disequilibrio di linkage (LD), che non consente di individuare con precisione le varianti causative reali in regioni altamente complesse. Identificare varianti causali nelle regioni associate alla SM che contengono geni target per farmaci (geni che codificano proteine bersaglio di un farmaco già approvato o di una molecola attualmente testata in studi clinici) potrebbe potenzialmente portare al riutilizzo di farmaci già noti, il cui meccanismo patogenetico target potrebbe sovrapporsi a quello della SM. La combinazione di mappatura fine con tecniche funzionali come il Massively Parallel Reporter Assay (MPRA) potrebbe risolvere questo problema, poiché quest'ultima è un metodo di screening ad alto rendimento in grado di testare migliaia di sequenze per il loro

potenziale ruolo nella regolazione della trascrizione in un singolo esperimento. Questo consente di identificare varianti che modulano l'espressione genica tra i loci associati ai target farmacologici della Sclerosi Multipla derivati dagli studi GWAS. Le evidenze provenienti da studi epidemiologici condotti di recente, che riportano un rischio 32 volte maggiore di sviluppare la sclerosi multipla (SM) negli individui che hanno convertito la sieropositività per Epstein-Barr Virus (EBV) rispetto a quelli rimasti sieronegativi, evidenziano l'importanza e il coinvolgimento del virus Epstein-Barr nella sclerosi multipla. La tecnica MPRA potrebbe servire come strumento per misurare le differenze nell'espressione del gene reporter influenzate dagli SNP associati alla SM, quando esposti a fattori ambientali associati alla SM come il virus Epstein-Barr. In particolare, Harley et al., 2018, hanno dimostrato il legame preferenziale dell'antigene nucleare Epstein-Barr 2 (EBNA2) ai loci di rischio per la SM in presenza di diversi alleli, provando il suo coinvolgimento nell'espressione genica. Analogamente, Keane et al., 2021, hanno dimostrato, utilizzando un saggio Chip allele-specifico, che EBNA2 si lega ai loci di rischio per la SM in maniera dipendente dall'allele nelle LCL, utilizzando l'inibitore di EBNA2, EBNA2-TAT. Studi pubblicati dai nostri collaboratori, Mechelli et al., 2015, hanno analizzato cinque principali alleli del ceppo di tipo 1 del virus EBV, il ceppo più frequente nella popolazione caucasica, identificati in base alla variazione nucleotidica nella regione più variabile di EBNA2. Hanno mostrato specificamente che il rischio di SM correla significativamente con un eccesso dell'allele 1.2 del gene EBNA2 (rapporto di probabilità (OR) = 5.13; intervallo di confidenza al 95% (CI) 1.84-14.32; p = 0.016). Questi dati alimentano un'altra ipotesi di lavoro, ossia testare le variazioni nell'effetto degli SNP funzionali associati alla SM in regioni potenzialmente trattabili con farmaci, quando esposti a fattori ambientali correlati alla malattia, come in questo caso la variante 1.2 di EBNA2 del virus Epstein-Barr. Utilizzando la tecnica MPRA, che può testare il ruolo regolatorio trascrizionale putativo di un grande numero di varianti, fornendo un valore quantitativo delle variazioni nella regolazione della trascrizione in presenza di diversi alleli per ciascuna variante, possiamo misurare differenze statisticamente significative in presenza della variante 1.2 di EBNA2 del virus Epstein-Barr. Inizialmente, abbiamo applicato un fine mapping bayesiano con l'obiettivo finale di identificare varianti funzionali in regioni complesse associate alla SM. Abbiamo iniziato con una vasta coorte di 5.903 individui della popolazione italiana continentale, comprendente 5.259 pazienti con SM e 1.644 controlli sani, con 6.339.414 SNP imputati che coprivano l'intero genoma. Volevamo identificare potenziali geni bersaglio di farmaci all'interno delle regioni associate alla SM, quindi abbiamo mappato le regioni replicate della SM (statisticamente significative con un valore p <0.05, 2Mb di ampiezza) con i geni elencati nel database Drug-Gene Interaction v4.2. Questa analisi ha identificato 36 regioni circostanti gli SNP replicati, che contenevano 238 geni trattabili con farmaci.

Successivamente, abbiamo applicato Paintor e Caviar BF, entrambi basati su un framework bayesiano, che assegna una probabilità posteriore di causalità, nota anche come probabilità di inclusione posteriore (PIP). Per un fine mapping funzionalmente più accurato, abbiamo incluso un set di punteggi di annotazione per considerare la rilevanza biologica di ciascun SNP all'interno della regione. A tal fine, abbiamo incorporato i punteggi di annotazione di GWAVA, CADD, FINSURF e Regulome Db nell'analisi di fine mapping. Una volta identificati gli SNP nei set credibili, abbiamo dovuto determinare quali geni potessero influenzare. Per questo, abbiamo condotto una mappatura SNP-gene utilizzando Open Targets Genetics. Abbiamo ottenuto 19 regioni contenenti SNP con prove convincenti di causalità e significatività statistica (Paintor PIP > 0,7). Tra queste 19 regioni, in 18 è stato identificato come possibile SNP causale uno diverso dallo SNP principale, ovvero quello che mostra l'associazione statisticamente più significativa nell'articolo originale di IMSGC. Tra queste, 10 regioni presentano SNP causali che mirano a geni target per farmaci. IFNGR1, TUBB4A, TEC, TGFBR3, IDE e CD40 sono geni target di farmaci approvati per altre malattie ma anche di SNP identificati come "causali". Inoltre, in quattro regioni, gli SNP "causali" mirano a potenziali geni target di farmaci (EOMES, FCRL3, CTSH e ADCY3) secondo Open Targets Genetics. Contemporaneamente all'analisi di fine mappatura, abbiamo condotto l'esperimento MPRA coinvolgendo 5 delle 36 regioni associate alla sclerosi multipla (SM) che contenevano regioni farmacologicamente rilevanti, in particolare quelle intorno agli SNP principali più associati nei GWAS. Le regioni selezionate sono state CD40, PRDX5, TEC-TKX, IFNGR2 e CHRNA9. Queste 5 regioni sono state scelte per la loro complessità architettonica, dovuta all'elevato numero di SNP in forte linkage disequilibrium con lo SNP principale della regione. Questa analisi è stata condotta indipendentemente dall'analisi statistica; tuttavia, dopo l'analisi di fine mappatura, abbiamo confermato che due regioni (CD40 e TEC) selezionate per l'analisi MPRA rientravano anche nell'insieme credibile previsto dall'analisi di fine mappatura. Il numero totale di SNP testati in queste 5 regioni è stato di 83, con il criterio di filtro che considerava uno SNP in forte LD con il Lead SNP se $r^2 \geq 0,77$. L'MPRA ha permesso il test funzionale simultaneo di numerosi potenziali elementi regolatori, utilizzando la base del test convenzionale con luciferasi in modalità ad alta produttività. La tecnica si basa sulla creazione di una libreria in cui ogni variante testata è rappresentata da una sequenza oligonucleotidica contenente il nucleotide di riferimento o alternativo in una regione di 145 basi. La novità consiste nel fatto che ogni sequenza è rappresentata da una sequenza nucleotidica di 10 bp chiamata TAG, che funge da codice a barre per distinguere ogni oligo. Ogni SNP testato è rappresentato da una sequenza di 145 basi contenente un allele alternativo o di riferimento e da una sequenza della stessa lunghezza contenente un segmento di 21 bp attorno al nucleotide variante, che funge da sequenza ipotetica nulla (in seguito chiamata Scramble). Ogni sequenza, di riferimento, alternativa o Scramble, è rappresentata 10 volte su entrambi i filamenti di DNA, portando a 60 rappresentazioni per ciascun SNP testato e a un totale di

4980 sonde, ciascuna con un codice a barre univoco. Per testare ogni variante, sono stati utilizzati due vettori distinti: pMPRAdonor1 e pMPRAdonor2. Entrambi i costrutti contenevano la libreria di SNP da testare; tuttavia, pMPRAdonor1 conteneva solo l'ORF del gene LUC ed è stato utilizzato per testare l'effetto della variante come promotore. pMPRAdonor2 conteneva il gene LUC accoppiato a un promotore minimale, utile per indagare l'effetto enhancer di ciascuna variante testata. Una volta pronti, i costrutti sono stati trasfettati in due diverse linee cellulari, rispettivamente HEK293T per testare la trasfezione e JURKAT come linea cellulare rilevante per la malattia. L'esperimento è stato eseguito due volte per valutazioni sperimentali, con trasfezioni effettuate in quadruplicato per garantire la significatività statistica sia per PMPRAdonor1 e pMPRAdonor2: sono stati lasciati esprimere il costrutto per 48 ore prima della raccolta delle cellule. Abbiamo eseguito il sequenziamento NGS utilizzando i kit Illumina DNA Prep e Nextera XT index sul cDNA ottenuto dalle cellule transfettate, mentre il DNA di input è stato utilizzato per confrontare la variazione nel numero di letture per ogni codice a barre post-transfezione. I dati sono stati analizzati utilizzando lo strumento mpralm, che calcola un rapporto tra i conteggi di DNA e RNA per ciascun allele di Riferimento e Alternativo di ogni sonda, illustrando questo rapporto tramite un LogFC associato a un valore P, che conferisce significatività. Il filtro per la significatività del valore P è stato impostato a un FDR<0,01. Siamo stati in grado di identificare SNP che esercitano un impatto significativamente diverso sull'espressione del gene Luc tra le due varianti alleliche (sequenza Alternativa vs. Riferimento). Tuttavia, ci siamo concentrati esclusivamente sulle sonde che hanno dimostrato una deviazione significativa dalla sequenza nulla (scramble) rispetto sia alla sequenza Alternativa che a quella di Riferimento. Per garantire maggiore affidabilità nei nostri risultati, abbiamo deciso di eseguire l'esperimento sulla linea cellulare Jurkat in duplice copia in momenti diversi, per poi misurare i risultati. Dopo questo passaggio di verifica, solo le sonde che mostravano la stessa direzione dell'effetto e superavano i filtri applicati all'output di mpralm tra i due esperimenti sono state selezionate per ulteriori analisi e valutazioni. Abbiamo confrontato il Log FC fornito da mpralm dei due esperimenti condotti sulla linea cellulare Jurkat e abbiamo osservato una correlazione elevata (R^2=0,8) tra i due esperimenti, suggerendo un'alta riproducibilità nel nostro approccio metodologico. Questa analisi ha dimostrato che, replicando l'esperimento, il Log FC delle varianti testate presentava un'elevata correlazione. Successivamente, per escludere eventuali sonde che non seguissero la stessa tendenza in entrambi gli esperimenti, abbiamo eseguito un semplice test di regressione lineare. La regressione lineare semplice con un intervallo di confidenza del 95% ha mostrato che 4 sonde testate non rientravano nell'intervallo di confidenza, poiché presentavano una differenza significativa nel Log FC tra i due esperimenti Jurkat. Di conseguenza, le quattro sonde che mostravano discrepanze non sono state prese in considerazione per ulteriori valutazioni.

Abbiamo scelto di valutare solo le varianti che mostravano la stessa direzione dell'effetto in entrambi gli esperimenti e una differenza statisticamente significativa rispetto all'ipotesi nulla, il che ci ha lasciato con 8 varianti per ulteriori analisi. Inoltre, ulteriori ricerche utilizzando database online come RegulomeDB, il progetto ENCODE, Screen registry V3 e UCSC hanno rivelato che alcune delle varianti che mostrano la stessa direzione dell'effetto in entrambe le linee cellulari si trovano in regioni di particolare interesse, come firme simili a promotori, H3K36me3, firme di enhancer prossimali o distali, ecc. Successivamente, abbiamo eseguito un'analisi del legame dei fattori di trascrizione (TF) utilizzando MotifBreakR, che predice il legame preferenziale dei TF in presenza di diversi alleli. Abbiamo utilizzato MotifBreakR in combinazione con MEME suite e TomTom per prevedere almeno un fattore di trascrizione che si lega preferenzialmente in presenza o assenza di un allele di interesse per 7 delle 8 varianti significative MPRA. Abbiamo quindi associato i fattori di trascrizione all'allele di rischio di ciascun SNP, per ottenere TF che potessero spiegare l'effetto dell'allele di rischio misurato in vitro sul gene reporter. Questo potrebbe imitare ciò che avviene biologicamente, anche se è certamente necessaria una valutazione ulteriore. Dopo aver valutato quali fossero le varianti causative attraverso MPRA e aver stimato l'effetto dell'allele di rischio sul gene reporter, abbiamo collegato questi dati ai possibili fattori di trascrizione (TF) che si legano in presenza dell'allele di rischio. Successivamente, utilizzando Open Target Genetics, siamo riusciti a individuare il gene bersaglio di ciascun SNP significativo rilevato con MPRA e a collegare il possibile effetto negativo dello SNP a un gene, associandolo a un farmaco che esercita un effetto opposto. La stessa procedura utilizzata con le tecniche MPRA è stata seguita per testare l'effetto di fattori ambientali noti associati alla sclerosi multipla (SM), come la variante EBNA2 1.2. Per questo, abbiamo eseguito una trasfezione sequenziale sulla linea cellulare Jurkat 2, inizialmente con un costrutto contenente la variante EBNA2 1.2 e, dopo 24 ore, con il costrutto MPRA per osservare l'effetto sull'espressione delle sequenze, misurato tramite il gene reporter. L'effetto di ogni SNP è stato misurato confrontando i conteggi dei tag per ciascuna sonda dopo la trascrizione, utilizzando lo strumento MPRAlm per confrontare le letture RNA con il DNA di input. I conteggi risultanti sono stati utilizzati per quantificare l'attività di una determinata sequenza regolatoria putativa. Per spiegare perché l'effetto dello SNP sul gene reporter fosse cambiato, abbiamo applicato MotifBreakR, che ha predetto diversi fattori di trascrizione (TF) che si legano preferenzialmente in presenza di uno dei due alleli. Abbiamo quindi selezionato i TF noti per essere influenzati in presenza di agenti virali. Per identificare il gene bersaglio di questi SNP, abbiamo applicato lo strumento Open Target Genetics, che può predire il gene bersaglio degli SNP testati mostrando l'effetto eQTL. Le varianti escluse dall'esperimento di replicabilità sulla linea cellulare Jurkat menzionato in precedenza non sono state prese in considerazione in questo esperimento.

Nella nostra pipeline sperimentale, abbiamo eseguito una trasfezione sequenziale, con ogni esperimento effettuato in quadruplicato per migliorare la significatività statistica. Inizialmente, abbiamo trasfettato le cellule Jurkat con il plasmide pCDNA 1.2 contenente circa 1602 bp della variante EBNA2, accoppiato con la sequenza GFP. In parallelo, come controllo positivo, abbiamo trasfettato lo stesso numero di cellule Jurkat con un costrutto GFP di base innocuo per le cellule. Attraverso MPRAlm, abbiamo ottenuto i valori di Log Fc e P corrispondenti a ciascun SNP testato. La nostra ipotesi si basa su questo passaggio cruciale, in cui vogliamo confrontare il Log FC di ciascuna variante testata individualmente per verificare se ci siano cambiamenti significativi nella direzione dell'effetto. Pertanto, abbiamo misurato l'effetto sull'espressione del gene reporter in presenza della variante EBNA2 1.2 per ciascun allele di riferimento o alternativo, confrontandolo con l'effetto sull'espressione del gene reporter in presenza di un plasmide GFP innocuo come sostituto della variante EBNA2 1.2. Per avere fiducia statistica nei risultati, abbiamo confrontato il LogFC di ciascuna sonda testata con un semplice test di regressione lineare tra l'esperimento con esposizione al patogeno e il controllo. Abbiamo selezionato un intervallo di confidenza del 95%, che ha rivelato 8 sonde con cambiamenti significativi nell'effetto misurato dal gene reporter quando esposte alla variante EBNA. Di queste 8 varianti, solo 2 hanno soddisfatto i filtri MPRA predefiniti, cruciali per individuare le sequenze testate che mostrano un effetto prominente e robusto legato a uno SNP con un **v**alore p significativo**.** Ci siamo quindi concentrati sui fattori di trascrizione che si legano all'allele di rischio con un forte effetto e che sono modulati dalla presenza dell'antigene EBNA2. Questa valutazione ha rivelato informazioni importanti sul probabile percorso attraverso il quale questi TF alterano l'effetto delle varianti come misurato dal gene reporter, spiegando i cambiamenti in presenza della variante EBNA2 1.2. Complessivamente, questi dati supportano l'efficienza della nostra metodologia nell'esaminare varianti funzionali associate alla SM tra quelle non funzionali. Siamo stati in grado di trovare i geni target di queste varianti e un probabile percorso attraverso il quale esercitano la loro funzione prevedendo i TF che si legano a esse quando l'allele di rischio è presente. Questo ci ha permesso di trovare farmaci possibili che esercitano un effetto opposto sul gene rispetto a quello della variante. Inoltre, abbiamo verificato che l'utilizzo dell'MPRA come stimatore dell'influenza dei fattori ambientali associati alla SM, in questo caso la variante EBNA2 1.2, ha dato risultati promettenti. Il nostro progetto apre la possibilità di applicare lo stesso approccio in altre regioni caratterizzate da un alto Disequilibrio di Legame associato alla SM o ad altre malattie, per identificare varianti funzionali tra le altre e per eseguire una previsione in silico dei percorsi alterati in presenza dello SNP.

INTRODUCTION

Multiple Sclerosis and Epidemiology

Multiple Sclerosis (MS) is a complex chronic autoimmune disorder that affects the central nervous system (CNS), including the brain, spinal cord, and optic nerves. (Houtchens and Khoury 2013). It is the main neurological disorder affecting young adults and middle-aged individuals in developed countries with an incidence 2-4 higher in women than in men. (Walton et al. 2020) The latest data from the Atlas of Multiple Sclerosis shows that now there are an estimated 2.9 million people that are living with MS around the world. There is an overall increase in MS prevalence, which reflects both improved diagnosis and a broader demographic range being impacted, including more diverse racial and ethnic groups in the U.S. and Europe.

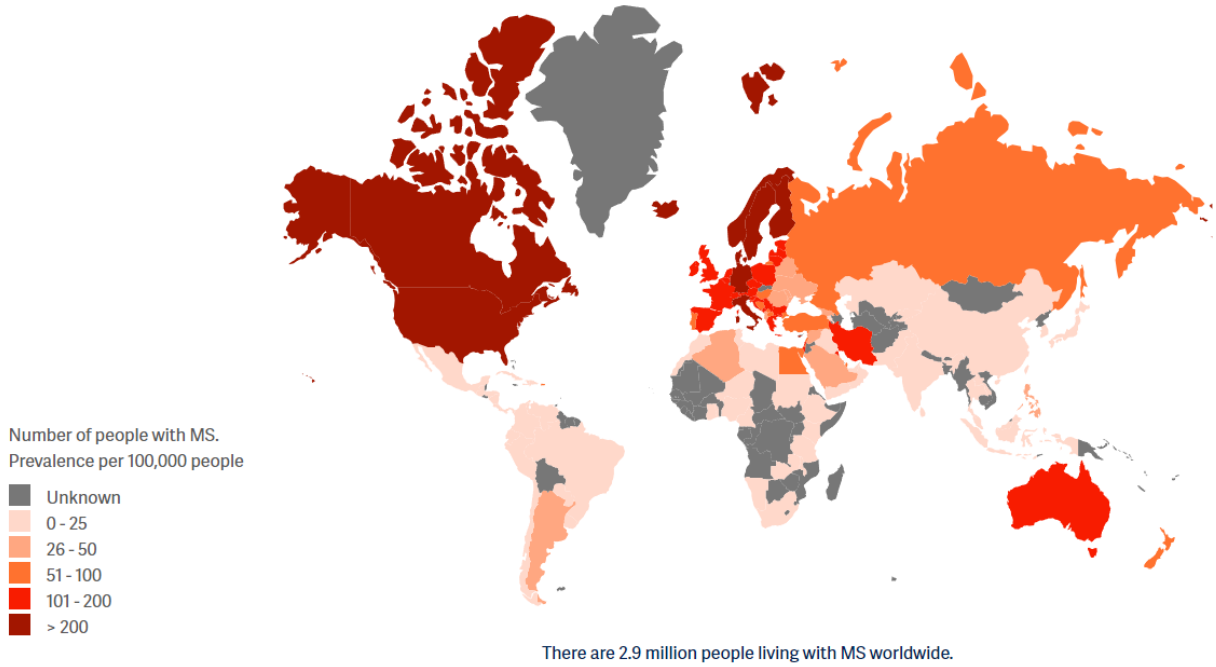


Fig.1 *Representation of epidemiological distribution of MS patients*

Although there is variation between countries, there was no overall change to the global incidence of MS, which still stands at 2.1 cases per 100,000 people per year. This is equivalent to someone in the world being diagnosed with MS every 5 minutes.

Pathology and Subtypes of MS

Clinically, MS manifests itself as a neurological deficit that frequently exhibits a relapsing-remitting pattern and can resolve completely or leave residual deficits. The deficits can involve a part of the central nervous system (CNS) alone or in combination. The most common manifestations involve somatosensory,

pyramidal-motor, and visual manifestations, the latter due to inflammatory demyelination in the afferent visual pathways (optic neuritis) or the efferent visual pathways (ocular motility disorders such as internuclear ophthalmoplegia). The MS lesions are focal areas of demyelination associated with variable inflammation and axonal loss that predominantly affect the white matter of the brain, spinal cord, and optic nerves (Sobel and Moore et al., 2008). MS lesions can be further classified histologically as active, chronic, and remyelinated. Active lesions are common in relapsing-remitting MS and are characterized by myelin degradation and inflammation caused by macrophage infiltration, reactive astrocytes, and perivascular and parenchymal inflammation. Chronic or inactive lesions are more often seen in patients with progressive disease. They are associated with more extensive demyelination, often with marked axonal depletion and relative absence of active inflammation. "Shadow lesions" are lesions that show an incomplete remyelination and are observed in patients with relapsing and progressive disease. However, the observed pathologic heterogeneity may not be exclusive to a subset of MS patients and it's probably related to the stage of disease in a patient (Barnet and Prineas et al., 2004). There are four key pathological features of MS: (a) inflammation, of complex pathogenesis, which is generally believed to be the main trigger of the events that lead to CNS tissue damage in the majority of cases; (b) demyelination, the hallmark of MS, where the myelin sheath or the oligodendrocyte cell body is destroyed by the inflammatory process; (c) axonal loss or damage; and (d) gliosis (astrocytic reaction to CNS damage).

In terms of the clinical course, there are several MS subtypes: relapsing-remitting MS (RRMS), with relapses (flare-ups) of disease separated by periods without clinical progression; secondary progressive, SPMS, which represents the phase of the disease where a gradual neurological deterioration follows a period of RR disease; primary progressive, PPMS, affecting approximately 15% of people with MS where the neurological deterioration is present from the onset, most frequently without superimposed relapses. The rare variant where a few acute escalations are imposed on the gradual PPMS-like course is called progressive-relapsing MS (PRMS) (Lublin and Reingold et al., 1996)

Etiopathogenesis of the Disease

The etiology of MS remains unknown; however, it is assumed to be caused by immune dysregulation triggered by genetic and environmental factors (Ascherio and Munger et al., 2007). Environmental, genetic,

and epigenetic factors are causal in MS and potentially interact with modifiable risk factors (Tomas Olsson et al., 2016). Although MS is not an inherited disease, there is a strong genetic component to its etiology as evidenced by the clustering of MS cases within families. The risk among first-degree relatives of MS patients is 10-15 times higher than the general population (absolute risk 2-5%); the concordance rate in monozygotic twins is about one-third (Weinshenker 1996; Kantraci 2008). Linkage analysis studies have revealed several gene loci as risk factors, with the major histocompatibility complex (MHC) HLA DR15/DQ6 allele being the strongest one (Barcellos et al., 2003; Sawcer et al., 2011). More recently, alleles of interleukin-2 receptor alpha gene (IL2RA) and interleukin-7 receptor alpha gene (IL7RA) have also been identified as inheritable risk factors (Hafler et al., 2007).

The pathogenesis of MS involves an immune response against CNS antigens mediated through activated CD40+ myelin-reactive T cells with a possible contribution by B cells. The majority of our understanding of the immunopathogenesis of MS is derived from the study of experimental autoimmune encephalomyelitis (EAE), an animal model of CNS inflammatory demyelination that can be induced by peripheral immunization with myelin protein components. EAE shares many of the histologic features of MS including active demyelination, oligodendrocyte, and axonal loss, all of which are presumably mediated by myelin-specific T cells (Yong et al., 2004; Gold et al., 2006). The immunopathogenesis of MS is best described as a loss of self-tolerance towards myelin and other CNS antigens, which leads to persistent peripheral activation of autoreactive T cells (Hafler et al., 2007). In individuals, which might be more, genetically susceptible, this loss of self-tolerance may be triggered by an environmental antigen, presumably an infectious agent. This infection might cause bystander activation of T cells or result in the release of autoantigens due to the cellular damage, which in turn causes activation of T cells by cross-reactivity between an endogenous protein and the pathogenic exogenous protein (viral or bacterial antigen) in a process known as molecular mimicry (Fujinami and Oldstone et al., 1985; Wucherpfennig and Strominger et al., 1995)

Once the myelin-reactive T cells are activated in the periphery, they can migrate across the blood-brain barrier (BBB). The transmigration process involves interaction between very late antigen-4(VLA-4) present on T lymphocytes and the vascular adhesion molecule-1 (VCAM-1) expressed on capillary endothelial cells, which is facilitated by expression and upregulation of various adhesion molecules, chemokines, and matrix metalloproteinases (MMPs) (Gold and Wolinsky 2011). After entering the CNS, the autoreactive T cells can be reactivated upon encountering the autoantigenic peptides within the brain parenchyma in the context of MHC class II molecules expressed by local antigen-presenting cells. This leads to an inflammatory cascade, which releases cytokines and chemokines, along with the recruitment of additional inflammatory cells including T cells, monocytes, and B cells, and persistent activation of microglia and

macrophages resulting in myelin damage (Hemmer et al., 2002, Frohman et al., 2006). This series of events which causes local inflammation and demyelination results in the exposure of sequestered myelin autoantigens providing an additional target of self-reactive T cells, a phenomenon called "epitope spreading" (Miller et al., 1997). Activation of resident CNS glial cells (such as microglia) results in persistent inflammation even in the absence of further infiltration of exogenous inflammatory cells. Evidence-based on animal studies suggests that CD4+ T-helper 1 (TH1) cells which release proinflammatory cytokines such as interferon-gamma, interleukin-2 (IL-2), and tumor necrosis factor- α *(TNF-* α*)* are the key players in mediating inflammation in MS with some role for the novel CD4+ T-helper-17 (TH17) cell subset which secretes IL-17 (O'Connor et al., 2001; Selter and Hemmer et al., 2013). Although demyelination is the hallmark of MS pathology, early axonal injury and axonal loss also occur and may drive disability progression. The mechanism of both myelin and axonal injury are not completely understood but are likely to include both direct injury to myelin and oligodendrocytes and axons by CD4+ and CD8+ T lymphocytes and complement.

Therapeutic options for Multiple Sclerosis

Immunomodulatory agents that help alter the course of the disease have been used widely helping with symptomatic management focusing on relieving specific symptoms such as fatigue, splasticity, bladder dysfunction, and pain. Corticosteroids (methylprednisolone) and adrenocorticotropic hormone (ACTH) have anti-inflammatory and immunomodulatory effects and are used to treat acute relapse to fasten recovery (Berkovich et al 2013). Immunomodulatory therapies (IMT) have made major improvements in the treatment of MS in the last two decades. Since the introduction of the first immunomodulation medication, interferon beta-1 in 1993, several other medications with a different mechanism of action, mode, and frequency of administration have become available. Currently, there are 12 medications approved for the treatment of MS, including six injectable, three infusion-based, and three oral medications. The mechanism of action of IMT used for the treatment of MS is mainly about the suppression of the immune response mediated by autoreactive lymphocytes. The majority of these drugs are effective in relapsing-remitting MS where inflammatory demyelination is the primary process (Winstock-Guttman et al., 1995; Rudick et al., 1997). The aim of using such therapies is to reduce the frequency of relapses and number of MRI lesions furthermore slowing the progression of disease. The beta-interferons (IFN-β*)* have multiple actions including stabilizing the BBB, and by this limiting the entry of T cells into the CNS, modulating T- and B-cell function, and altering the expression of cytokines. Glatiramer Acetate (GA) or Copolymer 1 is a synthetic complex that mimics myelin basic protein (MBP), one of the many autoantigens targeted by the

T cells. Due to this similarity, GA blocks the formation of myelin reactive T cells and induces GA-specific regulatory T-cell expression and Th2 anti-inflammatory cytokine production (Rudic et al., 1997). The clinical efficacy of GA in terms of reducing relapse rate and MRI lesions is similar to *(IFN*-β). Natalizumab is a humanized monoclonal antibody that prevents the transmission of lymphocytes across the BBB (Ransohoff 2007). It has shown a major superiority as opposed to other drugs in relapse rate reduction and disability progression. However, it presents some concerns, as it is associated with progressive multifocal leukoencephalopathy, by the reactivation of the JC virus (Yousry et al., 2006). Due to this risk, Natalizumab has been limited to being used as a second-line drug in patients with breakthrough disease or intolerable side effects with first-line therapies. Fingolimod is a sphingosine-1-phosphate receptor (S1P1) modulator and is the first drug approved for the treatment of MS. This drug prevents the migration of activated T cells from lymph nodes thereby limiting their entry into the CNS. A potential side effect of Fingolimod includes first-dose bradycardia, and reported cases of PML (Cohen et al ., 2010). Another emerging therapeutic approach is the use of CD19 chimeric antigen receptor (CAR)-T cell therapies in progressive MS, which has resulted as an acceptable approach showing no clinical signs of neurotoxicity (Fischbach et al., 2024).

Genetic Factors contribcuting to MS

The presence of a genetic component in the pathogenesis of Multiple Sclerosis is supported by the observation of familial clustering and the higher prevalence of MS in certain ethnic populations, particularly in those of northern European descent, compared to others such as African and Asian groups, regardless of their geographic location (Ebers and Sadovnick et al., 1994; Oksenberg et al., 1996).

Studies involving twins from diverse populations consistently demonstrate that the risk of MS is higher in monozygotic twins (25-30% concordance) of individuals with MS compared to dizygotic twins (2-5%). These findings provide evidence supporting a complex etiology for MS, involving multiple genetic factors of moderate effect as well as environmental influences. (Sadovnick et al., 2004)

Our knowledge of MS genetics has evolved dramatically in the recent year shifting the interest to other genetic factors, which are associated with the disease apart from Regions of the Major Histocompatibility Complex (HLA, Human Leukocyte Antigens) on the short arm of the chromosome 6 that extends for about 4 Mb. The HLA class II and I genes are particularly relevant modifiers of the disease risk: variants of the class II genes encode, products that present antigens to the CD4+ T lymphocytes, and class I products present antigens to the CD8+ lymphocytes. Instead, the class II variant HLA-DRB1-15:01 is highly associated with the disease with an increased risk of MS (odds ratio (OR) ~3). In contrast, the class I variant HLA-A*02 is related to the protection from the disease (OR ~0.6). The absence of HLA-A*02 and the presence of DRB1*15:01 have a combined OR of ~5. (Sawcer et al., 2011; Brynedal et al., 2007; Beecham

et al 2013) The importance of genetics in multiple sclerosis is also confirmed by data showing an increased risk for siblings of affected individuals to develop the disease, which is around 20 times higher for them than for an individual from the general population (IMSGC, 2005).

The development of Genome-Wide Association Studies (GWAS) allowed the simultaneous identification of hundreds of thousands of SNPs, spaced across the entire genome for the association with a particular trait in case-control datasets composed of genetically unrelated individuals, giving new insights into genetic variants that contribute to diseases (Manolio T.A et al., 2010). GWAS relies on the concept of linkage disequilibrium (LD), which refers to the non-random association of alleles at different loci within a population. It is created by evolutionary forces such as mutation, drift, and selection, and is broken down by recombination (Hartl et al., 1997). Generally, physically close loci together exhibit closer LD than loci that are farther apart on a chromosome. The larger the (effective) population size, the weaker the LD for a given distance. Although GWASs are unbiased concerning prior biological knowledge (or prior beliefs) and concerning genome location, they are unbiased in terms of what is detectable. GWASs rely on LD between genotyped SNPs and ungenotyped causal variants. The strength of statistical association between alleles at two loci in the genome strongly depends on their allele frequencies, such that a rare variant, (frequency <0.01) will be low in LD (as measured by r2) with a nearby common variant, even if they map at the same recombinant interval (Wray N.R et al., 2005). But, SNPs that are on the SNP chip have been selected to be common, therefore GWASs are by design powered to detect association with causal variants that are relatively common in the population. Most recent GWASs have reported a large number of associations with genes outside of the HLA region. In the last few years, international studies analyzing large datasets at the genome-wide level, have identified 200 loci involved in the susceptibility of MS and the HLA region. These discoveries were mainly due to the contribution of three international studies in 2011 by the International Multiple Sclerosis Genetics Consortium (IMSGC), 2013 (IMSGC, 2013), and 2019 (IMSGC, 2019). The first MS GWAS was reported in 2007 by IMSGC, which analyzed a substantial portion of common genetic variations to identify factors associated with MS. This study focused on common variants in family trios and utilized advanced genotyping techniques, analyzing 334,923 SNPs in 931 family trios (one affected child and both parents) capturing 2.2 million SNPs in individuals of European ancestry. The screening confirmed with genome-wide significance the association of the previously identified locus containing the Interleukin-7 Receptor α *(IL7Rα)* gene and detected the novel non-HLA disease-risk locus, defined by the presence of the Interleukin-2 Receptor *α* (IL2R*α*) gene. One of the most important GWAS studies, published in 2011, was conducted by "The International Multiple Sclerosis Genetics Consortium & the Wellcome Trust Case Control Consortium 2" (IMSGC-WTCCC2) which identified 57 different genes associated with Multiple Sclerosis, employing nearly 10.000 MS cases and

20.000 healthy controls of European Ancestry and analyzing approximately 450.000 SNPs. This study identified genes encoding for the cytokine pathway, the costimulatory pathway, and other molecules involved in the immune system functions. Assuming the presence of genetic susceptibility factors shared by autoimmune diseases, in 2013 IMSGC started the Immunochip project, which was performed on 14.498 MS patients and 24.703 HC. This study identified 48 new loci associated with MS and confirmed 49 already known regions bringing the number of risk variants that have been pointed out outside the MHC locus to 110 (International Multiple Sclerosis Genetics Consortium 2013). Chip-based technologies allowed us to further characterize the MS association signals, improving the definition of the association also in the HLA region. The IMSGC 2015 project which included also the Italian population, identified 9 HLA alleles and 2 allelic combinations involved in the association with MS. In particular, this study confirmed the HLA-DRB1*1501, HLA-DRB1*0301, and the HLA-DRB1*0303 as risk alleles and the HLA-A*0201 allele as a protective one. In addition, this study identified new risk alleles (HLA-DRB1*0801, HLA DQB1*0302) and new protective alleles (HLA-B*4402, HLA-B*3801, HLA-B*5501). The 2019 GWAS study (IMSGC, 2019) increased the number of statistically independent associations with MS susceptibility to 233. They identified 200 risk loci in the autosomal non-major histocompatibility complex (non-MHC) associated with susceptibility to the disease. Additionally, 30 HLA markers, excluding HLA-DRB1*15, and one locus on chromosome X were discovered to be linked to MS. The genome-wide and suggestive effects jointly could explain about 48% of the estimated heritability. They used an ensemble of methods to prioritize 551 putative susceptibility genes that implicate multiple innate and adaptive pathways distributed across the cellular components of the immune system. Although these efforts in identifying these associated variants and loci and although several signals are near genes involved in immunologic processes, the effector mechanism for most associations remains unknown. However, even though most of these genetic variants are non-coding they may influence indirect pathways of gene expression and splicing in immune cells, affecting both innate and adaptive immunity, which are implicated in the pathogenesis of MS (Leuven et al., 2022).

In 2022, another GWAS focused on identifying genetic variants, that influence Multiple Sclerosis patients. This study identified two variants, one located in the DYSF-ZNF638 gene and another in the DNM3-PIGC gene, which consistently affects various aspects of the disease, such as severity, suggesting potential targets for future drug development. (IMSGC 2023)

Environmental Factors contributing to MS

Genetic predisposition alone explains only one part of MS etiopathogenesis, environmental factors play an important role too. Various environmental factors are associated with the risk of MS, including early obesity, vitamin D deficiency, Infections (Epstein-Barr virus (EBV)), gut microbiome imbalance, and smoking.

Smoking

Smoking was initially suggested as a risk factor for MS by a pooled analysis of several small studies that showed an OR ~1.5 (Hawkes et al., 2007; Handel et al., 2011) which was later confirmed in a large case-control study (Hedstrom et al., 2009). It has been established that the relationship between MS and smoking is dose-dependent: cumulative smoking is associated with an increase in the risk of developing the disease. Elevated levels of cotinine in the serum or plasma (>10ng/ml), which reflect smoking, from patients before developing MS were associated with a similar pattern in risk increase. Passive exposure to smoking has also been associated with increased risk for MS, suggesting that even minor lung irritation is important. There has been evidence that smoking, not only is associated with an increased risk of developing MS but also with the risk of developing neutralizing antibodies against drugs used in treating MS such as natalizumab. The main mechanism by which smoking increases the risk of developing MS is by provoking lung inflammation and promoting proinflammatory pathways. If CNS auto-antigen cells are present in the lung, such cells might be activated to attack the CNS; supported by the experimental autoimmune encephalomyelitis (EAE) rodent model of MS. Smoking displays also a considerable interaction with MS-associated HLA risk alleles. In the Scandinavian population, having the class II HLA-DRB1*15:01 MS risk allele confers an OR of ~3, and lack of HLA-A*02 confers an OR of ~1.8, resulting in a combined OR of ~ 5 among non-smokers; however, in smokers, the combined OR is much higher, at ~14 (Hedstrom et al 2011). Smoking is also able to influence CD4+ T cells development since it can cause post-translational modifications of peptides like citrullination, which can lead to a bypass of central thymic tolerance, favoring autoimmunity (Klareskog et al., 2009). The interaction of smoking with MS HLA risk alleles can shed light on disease mechanisms in MS.

Sun exposure and vitamin D levels

Epidemiological observations of a latitude-dependent variation in MS incidence and prevalence, although founded by the increased prevalence of the MS-predisposing HLA DRB1*15:01 allele enriched in the Northern gradients, have provoked a large number of studies investigating sun exposure and vitamin D in the risk of MS. It has been demonstrated that increasing vitamin D levels before the age of 20 years, is associated with a decreased risk of MS later in life (Munger et al., 2006). Recently, high vitamin D levels were shown to correlate with decreased axonal damage, as assessed by cerebrospinal fluid (CSF) neurofilament light chain levels in people with MS (Sandberg et al., 2015). The role of vitamin D in MS is also supported by findings from genetic studies, where it has been shown that polymorphisms close to a central vitamin D metabolism enzyme gene *CYP27B1* are associated with an increased risk of MS(Beecham et al., 2013). Interestingly, based on in vitro studies, vitamin D was proposed as the first example of a gene-environment interaction involving the strongest genetic risk factor for MS, HLA DRB1*15:01 but this finding has not been reproduced in case-control studies. Because of this evidence on the protective effect of vitamin D, this vitamin has also been added to conventional therapy in man studies but it is still to be determined if it can have a therapeutic effect once a patient is diagnosed with MS.

Early Obesity

Growing evidence from studies in the past years strongly supports the role of early obesity in the risk of MS. The association is stronger when a BMI >27 is present, although being modestly overweight is also associated with an increased risk of MS. According to observations, adolescence seems to be the critical period in which weight affects the risk of MS in adulthood, as a high BMI at 10 years was not associated with future disease risk (Wesnes et al., 2015). However, Mendelian randomization studies show that genetic determinants for high BMI are associated with increased MS risk. Although the association between BMI and MS is well established, whether high BMI influences disease course is unknown. As with smoking and EBV, BMI interacts with HLA genetic variants: individuals with a high BMI who carry the DRB1*15:01 and do not have the protective HLA-A*02 have a ~14-fold increased risk of MS. This supports the hypothesis that risk factors affect common biological pathways involving inflammation and adaptive immunity, as well as support for a causal role for obesity. The exact mechanism by which obesity may attribute to MS is still unknown, but it is thought to be an overlap of three possible pathways. Firstly, obesity is characterized by a low-grade inflammation in which increased levels of proinflammatory mediators are produced in fat tissue. Second, obesity is associated with increased levels of leptin, a mediator connected to proinflammatory processes. And lastly, obesity leads to decreased bioavailability of vitamin D, which further promotes inflammatory processes. Any of these mechanisms can enhance the activation of adaptive,

autoreactive immune cells, which can trigger bouts of neuroinflammatory activity. The relevance of obesity about a putative immune attack on the CNS is also strongly supported by the observed interaction between EBC and BMI: where a history of either EBV infection or BMI increases the risk of MS two-fold, but a combination of both factors increases the risk of MS 14-fold (Hedstrom et al., 2015) independent of the contribution of the HLA-DRB1*15:01 class II risk allele.

Microbiota

Experiments involving transgenic mice that express T-cell receptors for a neuroantigen, demonstrate that T cells can become activated and cause MS-like neuroinflammation. Such transgenic mice that are free from bacteria, however, did not develop neuroinflammation, suggesting the role of bacteria in the gut in triggering the activation of adaptive immune cells to attack the CNS (Berer K et al., 2011). The gut microbiome consists of different communities of bacteria, fungi, viruses, and other microorganisms that coexist in symbiosis inside the human body. The role of microbiota is well established as it has functions in nutrient metabolism, immune modulation, and protection against pathogens when dysregulated can cause dysbiosis. This has led to research focused on the so-called gut-brain axis, a bidirectional neuro-hormonal communication system between the intestinal microbiota and the CNS. In particular, some components of the intestinal microbiota may be pro-inflammatory, promoting the migration of immune cells into the CNS, and contributing as a key parameter to the development of autoimmune disorders such as MS (Maglione et al., 2021). Furthermore, spontaneous EAE developed differently in such mice when kept at different animal facilities, suggesting that different strains of bacteria differ in the way they cause neuroinflammation. These data altogether support the hypothesis that the type and/or distribution of gut microbiota could affect the risk and course of MS.

Epstein - Barr virus (EBV): The role of EBNA1/EBNA2 antigens

There is no specific pathogen regarded as the consistent initiator of the disease, this is mainly due to the complexity and variation within individuals suffering from MS. However, the accumulation of epidemiological, serological, and virological data has increasingly supported the involvement of Epstein-Barr Virus in the genesis of MS. The Epstein-Barr virus is a type of gammaherpesvirus that results in long-term infection in over 90% of the population worldwide (Keane et al., 2021). Recent extensive studies suggest that Epstein-Barr infection is likely a necessary factor for the development of the disease (Soldan and Lieberman, 2023). The largest and the most accurate epidemiology study of the relationship between EBV and MS was performed by A. Ascherio and his group (Bjornevik et al., 2022), which performed a follow-up on more than ten million US Army personnel for over 20 years and showed a 32-fold increased risk of MS diagnosis in individuals who converted to EBV seropositivity compared with those who remained seronegative. Indeed; this is the largest and most comprehensive study strongly suggesting that EBV infection is required for subsequent development of MS. Even though more than 200 genetic variations associated with the risk of MS have been identified, and 47 of these are linked to functions of the Epstein-Barr virus, the interplay between these risk-associated genetic variations and EBV that might influence the susceptibility to MS is still not well established. However, there is evidence showing that the Epstein-Barr virus can engage with genetic factors linked to MS such as the HLA-DRB15*1501 region increasing the probability of carriers to 14 folds higher to develop the disease as opposed to the carriers not infected with EBV. (Keane et al., 2021). Another mechanism by which EBV is associated with MS is molecular mimicry, which proposes a mechanistic pathway where immune responses elicited by EBV mistakenly target epitopes on CNS protein that share similar amino acid sequences, potentially contributing to the pathogenesis of MS. Chronic latent and persistent infection with EBV provides an ongoing source of viral antigens that provoke immune response. During the hyperproliferative phase, EBV enters a type III latency stage characterized by the expression of several latency-associated genes, including EBNA1, EBNA2, EBNA3A, EBNA3B, EBNA3C, and various non-coding RNAs (Soldan and Lieberman, 2023). A most striking observation in a nested case-control study was that primarily all EBNA1-negative individuals had serologically converted to being EBNA1 antibody-positive before MS onset (Levin et al., 2010).

EBNA1

The primary function of the Epstein-Barr Nuclear Antigen 1 (EBNA1) lies in facilitating the replication of the EBV genome within the host cell. In MS patients, the genetic risk for elevated EBNA1 concentration is positively correlated with the development of MS (Zhoy, Y et al., 2016). In a pattern similar to that seen with smoking, infectious mononucleosis, and increased EBNA1 antibody concentrations interact with HLA MS risk variants, and infectious mononucleosis interacts with HLA DRB1*15:01 to increase the risk of MS (Nielsen T.R. et al., 2009). Recent studies have highlighted the therapeutic potential of EBNA1 inhibitors in disrupting EBV latency and inhibiting tumor proliferation in experimental models. This underscores the significance of EBNA1 as a target for developing strategies to combat EBV-associated malignancies (Damania et al., 2022).

EBNA2

Epstein-Barr Nuclear Antigen 2 (EBNA2) is essential for maintaining the latency III growth phase of EBV and acts by regulating both viral and cellular genes. There has been proven an association between several autoimmune disorders including MS, which have an over-representation of EBNA2 binding sites at disease-risk loci in EBV-infected B cells (Harley et al., 2018). They were able to identify many autoimmune-associated risk variants that are associated with altered EBNA2 binding in the human genome. There has been evidence that 6 out of the 200 MS risk-associated loci were co-located with EBNA2 chromatin immunoprecipitation sequencing (ChiP-seq) binding peaks, where EBNA2 expression levels also correlated with the expression of these risk genes (Afrasiabi et al ., 2019). In a study conducted by Keane et al., 2021 they were able to demonstrate by using an allele-specific ChiP assay that EBNA2 binds to MS risk loci in an allele-dependent manner. They were able to confirm that selective binding of EBNA2 to MS risk loci, contributes to changes in gene expression by using the previously described EBNA2 inhibitor EBNA2-TAT. However, their study was limited by only testing MS risk loci associated with MS risk gene expression in LCLs. This work aligns with other data reviewed by Bar-Or et al 2020, supporting the hypothesis of targeting EBNA2 for therapeutic benefit in MS.

Linkage Disequilibrium

Linkage disequilibrium (LD) is the non-random association of alleles of different loci and it is a sensitive indicator of the population genetic forces that structure the genome (Lewontin et al 1960). Linkage disequilibrium (LD) poses significant challenges in identifying causal genetic variants, as it often links non-causal variants with causal ones, obscuring precise localization. Fine-mapping methods address this by modeling LD structures and prioritizing potential causal variants, though they often struggle with heterogeneity and assumptions about single causative loci. Another limiting factor is the need for computational trade-offs to balance power and false discovery control (Sesia et al.,2020). LD

presents a complex issue in pinpointing a truly causative variant as most variants are associated with traits but they do not cause those traits. This problem is addressed by fine-mapping methods, which attempt to prioritize putative causal variants for functional follow-up studies.

Fine mapping

Genome-wide association studies have been widely used to identify the genomic regions on chromosomes that harbor genetic determinants of complex traits (WTCCC et al 2007). However, SNPs from GWAS studies typically do not cause the trait, and being that patterns of LD among SNPs can be complex, it can be challenging to determine the real causal variants. Fine mapping could be the solution to this problem since it can determine the genetic variant (or variants) responsible for complex traits, given evidence of an association of a genomic region with a trait and assuming at least one causal variant exists. Usually in fine mapping studies, each region is visually explored for its LD structure for genes known to be mapped to the region, which is individually studied by a fine mapping strategy. The goal of fine mapping is to determine which variants are most likely to contribute to the disease and the strength of evidence. This is highly valuable since it can be used for follow-up studies, such as laboratory functional studies. Although GWASs can provide statistical evidence that a region is likely to harbor a causal variant, additional statistical methods are needed to discriminate likely functional variants from variants that are merely correlated with the functional variants (Schaid et al 2018). Fine mapping is also crucial for gaining insight into the underlying mechanism from a GWAS and translating the findings into potential therapeutic approaches. Several factors impact the effectiveness of fine-mapping, which includes the local linkage disequilibrium (LD) structure, the number of causal SNPs in a specific region, their effect sizes on the trait of interest, the sample size available for analysis, the density of SNPs, and the feasibility of accurately measuring the causal variants. (Schaid et al. 2018)

There are many strategies to conduct fine mapping, some of those are the Heuristic LD approach, penalized regression, and Bayesian fine mapping.

Functional annotations also have a crucial role in fine mapping by providing valuable biological information about the variants, including epigenetic marks, conservation scores, and other relevant scores. This prior knowledge enables the identification of causal variants even before analyzing specific GWAS data, a strategy known as functionally informed fine-mapping. There are various approaches to conducting functionally informed fine-mapping, and one effective method is to incorporate these data into the Bayesian method. (Wang and Huang 2022)

Bayesian Fine Mapping

Bayesian statistics is an approach to data analysis based on Bayes's theorem, where available knowledge about parameters in a statistical model is updated with the information in observed data. The background knowledge is expressed as a prior distribution and combined with observational data in the form of a likelihood function to determine the posterior distribution Ren van de Schoot et al., 2021). These methods for fine mapping have been specialized to focus on the SNPs that have the largest chance of being causal. This provides an alternative approach to assessing associations that alleviates the limitations of p-values at the cost of some additional modeling assumptions.

Bayesian fine mapping along with many models of fine mapping is based on the PIP (Posterior Inclusion Probability) for a SNP, as causal in any of the models. The PIP is computed by the sum of the posteriors over all models that include that SNP as causal. However, this model can be used to determine the credible set, so the minimum set of SNPs that contain all causal SNPs with a probability α. When assuming only one causal SNP, α is the sum of PIPs for SNPs in a set. This means that an α credible set is equivalent to ranking SNPs from largest to smallest PIPs and taking the cumulative sum of PIP until it is later at least α (Maller et al., 2012). There are numerous advantages to Bayesian for fine mapping: Firstly, unlike P values, posterior probabilities for a SNP can be directly compared. Secondly, they tend to select fewer SNPs as potentially causative compared to selecting SNPs based on their correlation with the lead SNP. Then, stimulation studies have shown Bayesian methods to perform better than both conditional stepwise regression and penalized regression models. Nevertheless, most importantly since Bayesian models are based on the joint effects of the SNPs, they control for SNPs with large effects, improving the power of the model to detect SNPs with lesser effects.

Functionally informed Bayesian Fine Mapping: Paintor, Caviar BF

PAINTOR (Probabilistic Annotation INTegratOR), is a framework to combine external functional annotations (sets of variants that localize within certain genomic features, e.g enhancers, repressors) with genetic association data (the strength of association between genetic variants and the phenotype) to improve the prioritization of causal variants n fine-mapping studies (Kichaev et al., 2014). This framework combines two lines of evidence to estimate variant-specific probabilities for causality: functional relevance and genotype-phenotype association. These probabilities are then used to prioritize the variants for functional validation studies to determine biological causality. More in detail, this model incorporates the external functional annotation data through an Empirical Bayes prior with parameters inferred from targeted fine-mapping data.

Indeed, by prioritizing variants using PAINTOR posterior probabilities, higher accuracy is observed compared to existing methodologies. PAINTOR can identify a greater number of causal variants in comparison to other methods. This improves the performance, because of PAINTOR's capability to model multiple causal variants while incorporating functional priors. (Kichaev et al. 2014)

CAVIAR (CAusal Variants Identification in Associated Regions) is a statistical framework that quantifies the probability of each variant to be causal while allowing an arbitrary number of causal variants. This is accomplished by jointly modeling the observed association statistics at all variants in the risk locus; posterior probabilities for sets of variants to be causal are then estimated using the conditional distribution of all association statistics in the locus conditional on the set of causal variants (Hormozdiari et al 2014). The approach Caviar takes as input is the association statistics for all of the SNPs at the locus together with the correlation structure between the variants obtained from a reference data set such as HapMap (Gibbs et al., 2003) or 1000 Genomes project (Abecasis et al., 2010). This information allows the method to predict a subset of the variants that have the property that all the causal SNPs are contained in the set with the probability *p*. This is the "p causal set" which is used in the follow-up studies by validating only the SNPs that are present in the set. This approach can be applied to a vast set of genetic variants, including structural variants.

CaviarBF and Paintor

Both CAVIAR and PAINTOR use the marginal test statistics directly and are likelihood-based based and they demonstrate a better performance over other fine mapping methods. They require only the marginal test statistics and the correlation coefficients among SNPs, instead of the original genotype data, which makes it easier to share data among different groups. The relationship and usage of CAVIAR/PAINTOR lead to a unified Bayesian framework for both fine mapping and association testing using marginal test statistics. This proposed method is called CAVIAR BF (Caviar Bayesian Framework).

Fine-Mapping efforts in Multiple Sclerosis

To pinpoint variants that are truly consequential to MS pathology; the regions surrounding the associated variants must be narrowed. By conducting a comparative analysis of follow-up studies of GWAS, it's possible to identify consistently implicated loci associated with autoimmune diseases, such as MS. These loci may prioritize a single gene based on multiple sources of functional and statistical evidence. Integrating such approaches into drug-discovery pipelines can help in focusing time and resources on the most promising targets for further functional investigation. Statistical analysis of fine-mapping data is employed to determine the number of independent association signals in a given region, the variant(s) most strongly associated with the disease (that is, the 'lead' or 'index' SNP), and set of highly correlated (linkage disequilibrium r2>0.9) SNP variants, all of which are equally likely to be causal. Studies conducted by the International Multiple Sclerosis Genetics Consortium (IMSGC 2013) with a high-resolution Bayesian fine-mapping, were able to identify five regions where one variant accounted for more than 50% of the posterior probability of association in the regions of *TNFSF14, IL2RA, TNFRSF1A, IL12A,* and *STAT4*. This study enhanced the catalog of multiple sclerosis risk variants and illustrates the value of fine-mapping in the resolution of GWAS signals. However, the regions where fine mapping is essentially required are characterized by high linkage disequilibrium making difficult the identification of causal variants at risk loci. Similarly, in 2021 our group conducted a fine mapping analysis performed in a large cohort of Italian multiple sclerosis patients and healthy controls. This study focused on a specific region with the strongest non-human leukocyte antigen (HLA) associated with MS in the Italian population that maps to the *TNFSF14* gene encoding LIGHT, a glycoprotein involved in dendritic cell maturation. Initially, the International Multiple Sclerosis Genetics Consortium confirmed the involvement of *TNFSF14* in the pathogenesis of MS in 2019. TNFSF14 gene encodes the glycoprotein LIGHT of which the intronic SNP rs1077667 was identified as the primary MS-associated variant in the region. Further analysis revealed that the MS risk allele is associated with reduced *TNFSF14* mRNA levels in blood cells, and MS patients exhibit lower *TNFSF14* gene expression compared to healthy controls. Additionally, individuals homozygous for the MS risk allele showed an increased percentage of LIGHT-positive immune cells. These findings suggest a potential role of rs1077667 in MS susceptibility and provide insights into the modulation of immune responses in MS. (Zuccalà et al. 2021)

In 2023, a fine mapping approach was used to integrate disease-specific genetic and epigenetic data with the human protein interactome and transcription factor knowledge enriching for drug target genes. This approach allows for the comprehensive analysis of genetic and epigenetic factors in the context of protein interactions and transcriptional regulation, providing valuable insights into disease mechanisms and

potential therapeutic targets. This study discovered synergistic signals of expression quantitative trait loci (eQTL) for variants in the CD40 gene's promoter region. These findings indicate that these variants have an influence on both epigenetic mechanisms and gene expression about multiple sclerosis. (Manuel et al. 2023).

Functional validation through high throughput techniques – MPRA

Massively Parallel Reporter Assay (MPRA) is an in vitro technique that uses the basis of the conventional luciferase assay in a massive way, which facilitates the systematic dissection of transcriptional regulatory elements. In MPRA, microarray-synthesized DNA regulatory elements and unique sequence tags are cloned into plasmids to generate a library of reporter constructs (Melnikov et al., 2012). These constructs are transfected into cells and tag expression is assayed by high-throughput sequencing, this approach gives resulting data that define accurate maps of functional transcription factor binding sites. As mentioned before the association between genetic variations and traits is often in non-coding regions with strong linkage disequilibrium (LD) which makes it difficult to identify a single causative variant among multiple correlated variants, even though statistical and functional fine-mapping approaches have been developed to identify credible sets of variants containing the causal variant (Schaid et al., 2018). However, these approaches cannot distinguish between proximal or highly linked variants and lack systematic prior information on the number of causal variants underlying association signals. To identify causal variants while controlling for LD we can apply MPRAs. MPRAs measure the effect of synthetic DNA libraries on the expression of a reporter gene, typically luciferase or GFP, containing a 3' UTR barcode. Such assays have screened potential regulatory elements in diverse cellular contexts, and have applications in saturation mutagenesis or tiling along regulatory regions of interest (Ernst J et al., 2016). Instead of relying on fluorescence measurements, researchers can utilize RNA-seq to identify which regulatory elements are active and assess the number of transcripts produced from their respective reporter genes. Each transcript is distinguishable by its unique barcode, and each SNP is represented by a Reference (Ref), Alternative (Alt), and Scramble (Scr) which allows for comprehensive analysis. The MPRA technique has been utilized to examine DNA variants positioned at various genomic sites, including promoter regions and enhancer/silencer sequences. Studies conducted recently have been able and have proved MPRA functionality in identifying susceptibility genes/variants from multiple GWAS loci. A recent study conducted by Choi et al., 2020 demonstrates how they were able to identify 39 candidate

functional variants starting from 832 high-LD variants, in 14 loci that displayed allelic transcriptional activity, a subset of which corroborates four colocalizing melanocyte cis-eQTL genes (Choi et al., 2020).

In another study where MPRA was applied to 32,373 variants from 3,642 cis-expression quantitative trait loci and control regions, they identified 842 variants showing differential expression between alleles, including 53 well-annotated variants associated with diseases and traits (Tewhey et al 2016).

MPRA as a tool to study the interaction between environmental factors and common polymorphisms associated with complex diseases.

Epidemiological studies conducted by A. Ascherio and his group have demonstrated that there is a 32-fold increased risk of developing MS in individuals who converted to Epstein Barr Virus (EBV) seropositivity as opposed to those who remained seronegative. This and many more studies have highlighted the importance and involvement of Epstein Barr Virus to Multiple Sclerosis. Our hypothesis is to use the MPRA construct as a tool to measure the effect of Epstein Barr virus in common SNPs associated with multiple sclerosis. Notably, Harley et al., 2018 have proven the preferential binding of the Epstein-Barr Nuclear Antigen 2 (EBNA2) to MS risk loci in the presence of different alleles proving its involvement in gene expression. Similarly, Keane et al., 2021 were able to demonstrate by using an allele-specific Chip assay that EBNA2 binds to MS risk loci in an allele-dependent manner in LCLs, by using the EBNA2 inhibitor EBNA2-TAT. Published work from our collaborators Mechelli et al., 2015 have studied five major alleles of the EBV type 1 strain, the most frequent strain in the Caucasian population, which were identified based on the nucleotide variation within the most variable region of EBNA2. Specifically, they showed that the MS risk significantly correlates with an excess of the 1.2 allele of the EBNA2 gene (odds ratio (OR) =5.13; 95% confidence interval (CI) 1.84-14.32; p=0.016).

This data provides promising grounds to lay our work hypothesis, to test the changes in the effect of functional SNPs associated with MS in regions, that are druggable when exposed to disease-related environmental factors, such as in this case the Epstein-Barr EBNA2 variant 1.2. Using the basis of the MPRA technique that can test the putative transcriptional regulatory role of a large number of variants delivering a quantitative value of transcription regulation changes in the presence of different alleles for each variant, we can measure statistically significant differences in the presence of the Epstein-Barr EBNA2 variant 1.2.

AIM OF THE STUDY

The general aim of this study is to pinpoint essential genetic variants within specific regions containing potential drug targets in Italian MS patients starting from genome-wide association studies (GWAS) data. We plan to achieve this by combining computational tools and *in vitro* validation methods to surpass GWAS's limitation in finding variants associated with a disease. We planned to employ a Bayesian fine mapping approach, to prioritize functional variants in selected MS-associated regions, thereby elucidating the underlying mechanisms identified in GWAS and translating these findings into potential therapeutic applications. In parallel with fine mapping applications, we planned to apply also *in vitro* validation techniques such as Massively Parallel Reporter Assay (MPRA). This application holds promise for understanding the impact of common variants on transcriptional regulation in complex diseases such as autoimmune disorders, particularly Multiple Sclerosis, where multiple genetic variants contribute to disease development. Specifically, the study planned to conduct a fine mapping on a dataset of genotyped Italian MS cases (N=4259) and controls (N=1644), focusing on 36 regions with 238 drug target genes. This involves genotype imputation and integrating genetic data with functional annotations to identify causal variants. Furthermore, using the Massively Parallel Reporter Assay (MPRA) in different cell lines we planned to investigate the functional impact and regulatory potential of these variants within five key regions, utilizing a disease-related T cell line (Jurkat). Furthermore, the study aims to explore the use of MPRA as a tool to investigate the interaction between environmental factors and genes associated with complex diseases. Specifically, we aimed to examine how genetic variants interact with environmental factors, particularly the Epstein-Barr virus (EBV), to influence MS pathogenesis. By employing the MPRA technique on the Jurkat cell line, this study is aimed to quantitatively analyze the regulatory activity of several sequence variations in response to EBV, focusing on the allele-specific regulatory effects in MS-associated regions. This research seeks to identify causative variants and their impact on gene expression, providing insights into the gene-environment interactions underlying MS.

MATERIALS AND METHODS

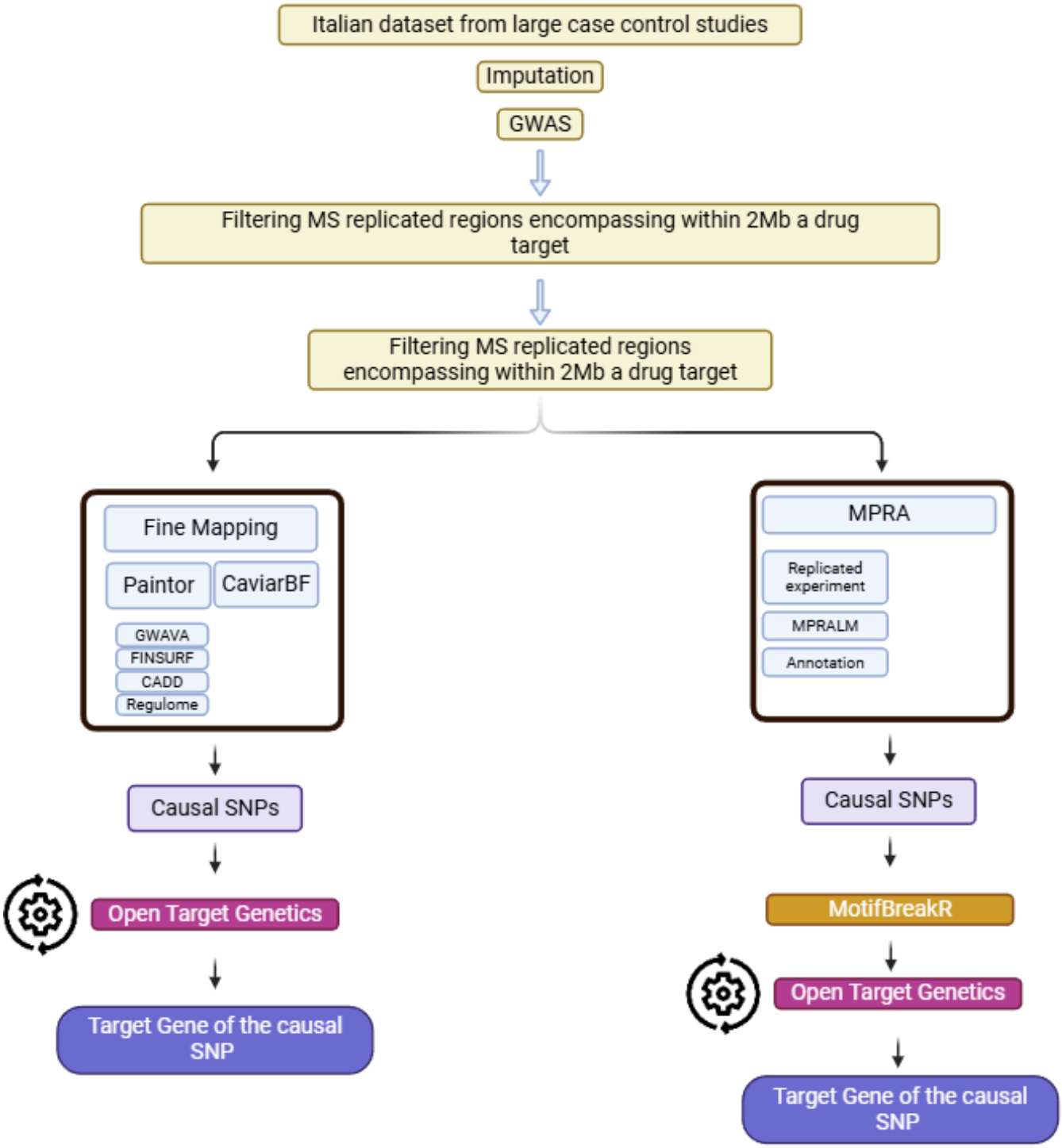


Fig.2 *Workflow of the methods used in our study*

Fine Mapping of regions containing drug target genes

To identify the main variants associated with Multiple Sclerosis within specific regions, which contain potential drug target genes we followed a systematic approach. We utilized a large cohort genotyped using array platforms with a genome-wide coverage of single nucleotide polymorphisms (SNPs). This initial dataset was then subjected to genotype imputation against the Haplotype Reference Consortium panel to generate a more comprehensive SNP map for the subsequent fine mapping. The final cohort of our study consisted of 5,903 individuals from the continental Italian population, out of which 4,259 were MS patients and 1,644 healthy controls, with 6,339,414 imputated SNPs covering the entire genome. This cohort combines two datasets which were collected by our laboratory as UPO (Progemus consortium) and San Raffaele Hospital respectively.

To identify potential drug target genes associated with MS, we cross-mapped the replicated Ms regions (statistically significant with a p-value <0.05) in the Italian cohort with the Drug-Gene interaction database V4.2. This analysis yielded 36 regions surrounding the replicated SNPs, which contained a total of 238 druggable genes.

Following this initial step, further investigation was required for the 36 selected regions to determine causality by disentangling the linkage disequilibrium (LD) effect from the statistical association between SNPs and MS. We applied two fine mapping tools: Paintor and CaviarBF. These tools utilize a Bayesian framework to assign a posterior probability of causality, known as the posterior inclusion probability (PIP), to each SNP within the 36 regions starting from the summary statistics of the original GWAS study. Additionally, a set of annotation scores was combined to perform a functionally informed fine mapping, which takes into account the biological relevance of each variant by incorporating GWAVA, CADD, FINSFURF, and RegulomeDB annotation scores in the analysis. GWAVA allows the assessment of the functional impact of non-coding variants by considering various annotations of the non-coding elements, including open chromatin, histone modifications, transcription factor binding databases, conservation, and GC content. (Ritchie et al., 2014)

CADD on the other hand, evaluates the deleteriousness and pathogenicity of variants by integrating multiple annotations and is commonly utilized in clinical settings. (Kircher et al., 2014) FINSURF has been developed to prioritize variants based on their functional impact on diseases. (Moyon et al., 2022) RegulomeDB is a database that appraises the functional impact of genetic variants in non-coding regions of the genome by considering multiple types of functional annotations, such as transcription factor binding sites, DNase I hypersensitivity sites, and histone modification marks. By taking into account these

elements, RegulomeDB assigns a regulatory score to each variant, which indicates the likelihood that a variant influences gene regulation. (Boyle et al., 2012)

Post analysis we selected the loci that exhibited a minimum 75% overlap in the credible sets identified by both tools respectively CaviarBF and Paintor, and that demonstrated an association p-value < 1e-4. Additionally, we considered a PIP >0.6 determined by the Paintor tool. After the prediction by Fine Mapping of the most probable MS-related SNPs among the 36 regions, we performed SNP-to-gene mapping using the Open Targets Genetics Database developed by Ghoussaini et al., 2021. This database is a comprehensive resource that integrates human GWAS and functional genomics data from various cell types and tissues. We retrieved specific information for each SNP of interest, including the nearest and target genes, the distances between the SNP and the transcription start site (TSS) of these genes, and the V2G global score. The V2G global score reflects the level of evidence supporting the functional impact of a SNP on a particular gene.

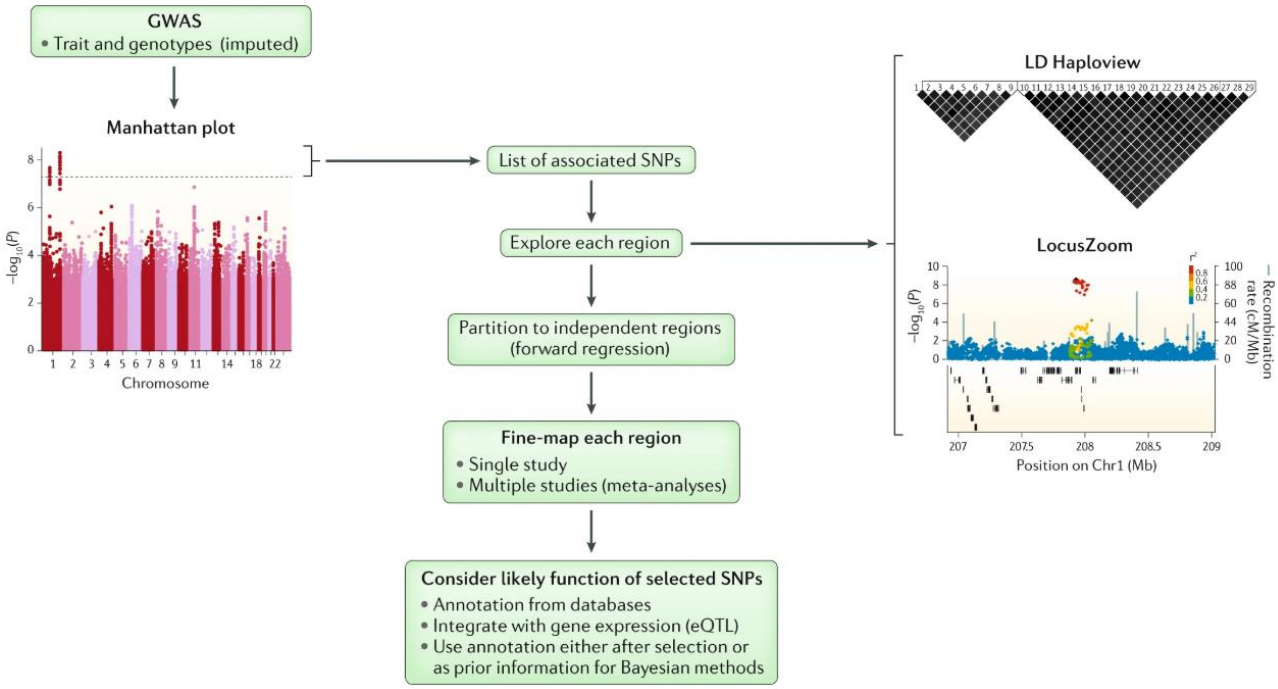


Fig.3 *Flow of a typical process from initial GWAS to annotation of the SNPs selected from fine-mapping analyses. Schaid et al.,2018 https://doi.org/10.1038/s41576-018-0016-z*

Massively Parallel Reporter Assay (MPRA)

Massively Parallel Reporter Assay (MPRA) is an in vitro technique that uses the basis of the conventional luciferase assay in a massive way, which facilitates the systematic dissection of transcriptional regulatory elements. In MPRA, microarray-synthesized DNA regulatory elements and unique sequence tags are cloned into plasmids to generate a library of reporter constructs (Melnikov et al., 2012). MPRAs measure the effect of synthetic DNA libraries on the expression of a reporter gene, typically luciferase or GFP, containing a 3' UTR barcode. Such assays have screened potential regulatory elements in diverse cellular contexts, and have applications in saturation mutagenesis or tiling along regulatory regions of interest (Ernst J et al., 2016). Instead of relying on fluorescence measurements, researchers can utilize RNA-seq to identify which regulatory elements are active and assess the number of transcripts produced from their respective reporter genes. Each transcript is distinguishable by its unique barcode, and each SNP is represented by a Reference (Ref), Alternative (Alt), and Scramble (Scr) which allows for comprehensive analysis.

Selection of Regions

Given the limitations of Fine Mapping in regions characterized by high Linkage Disequilibrium we decided to perform the MPRA as a pilot experiment for the first time in five out of the 36 regions respectively named by us as the CD40, TEC-TKX, IFNGR2, CHRNA9, PRDX5 regions. These five regions showed very complex architectural structures so we focused on variants that had an LD >0.77 with the lead SNP of the region. These selection criteria left us with 83 SNPs mapped across the 5 selected regions to test for their functional activity.

Represented below are the regional plots of the five regions selected for the MPRA analysis due to their architectural complexity.

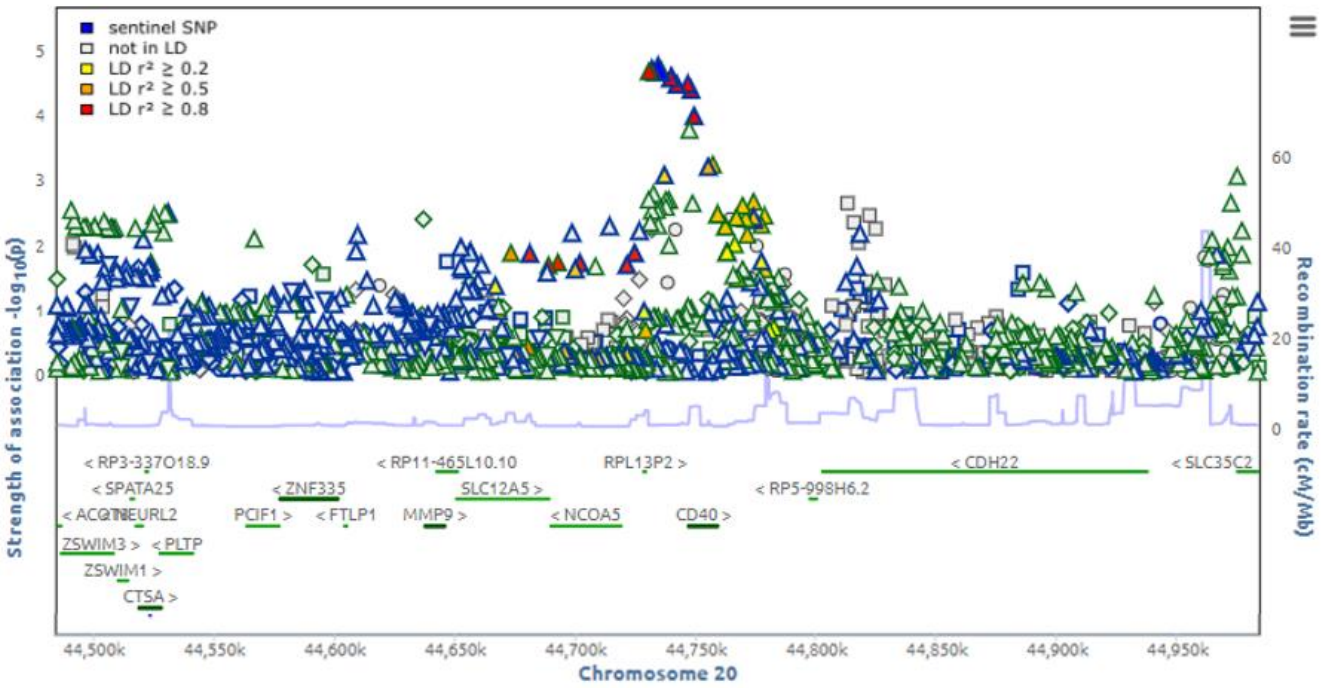


Fig 4. *Regional plot of the CD40 region.*

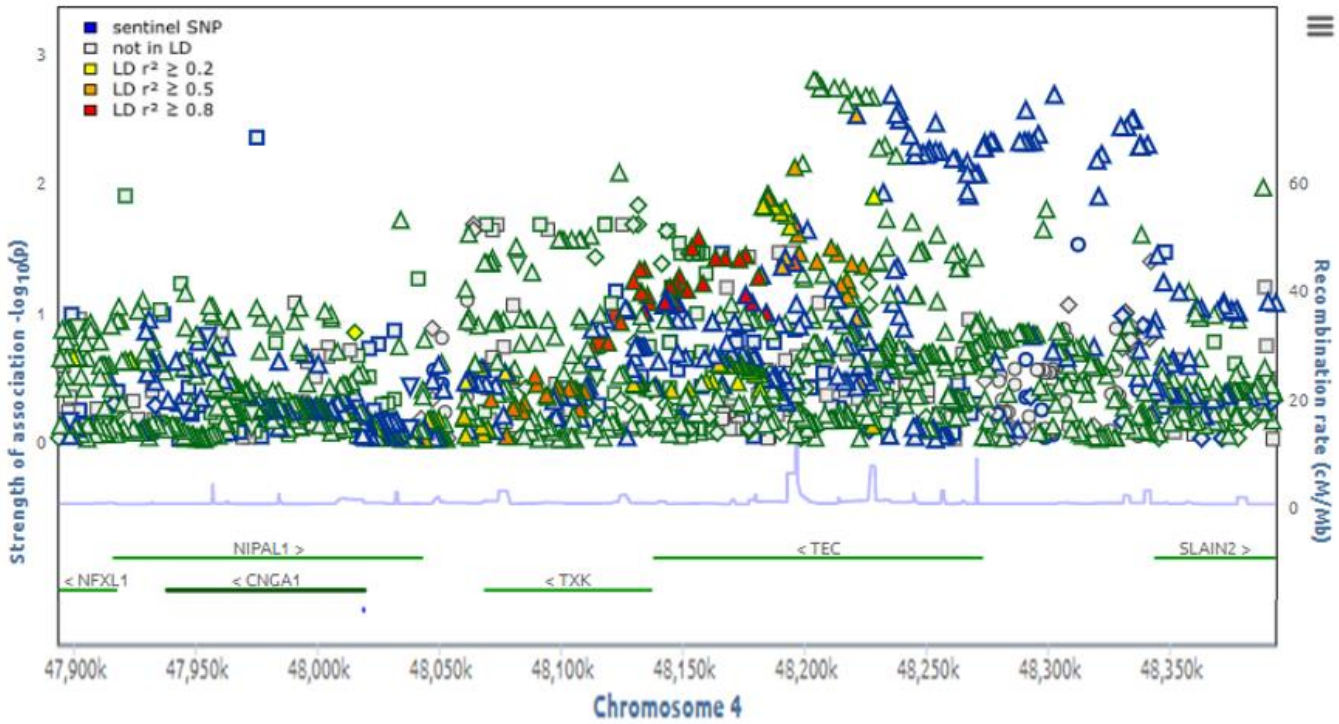


Fig 5. *Regional plot of the TEC-TKX region.*

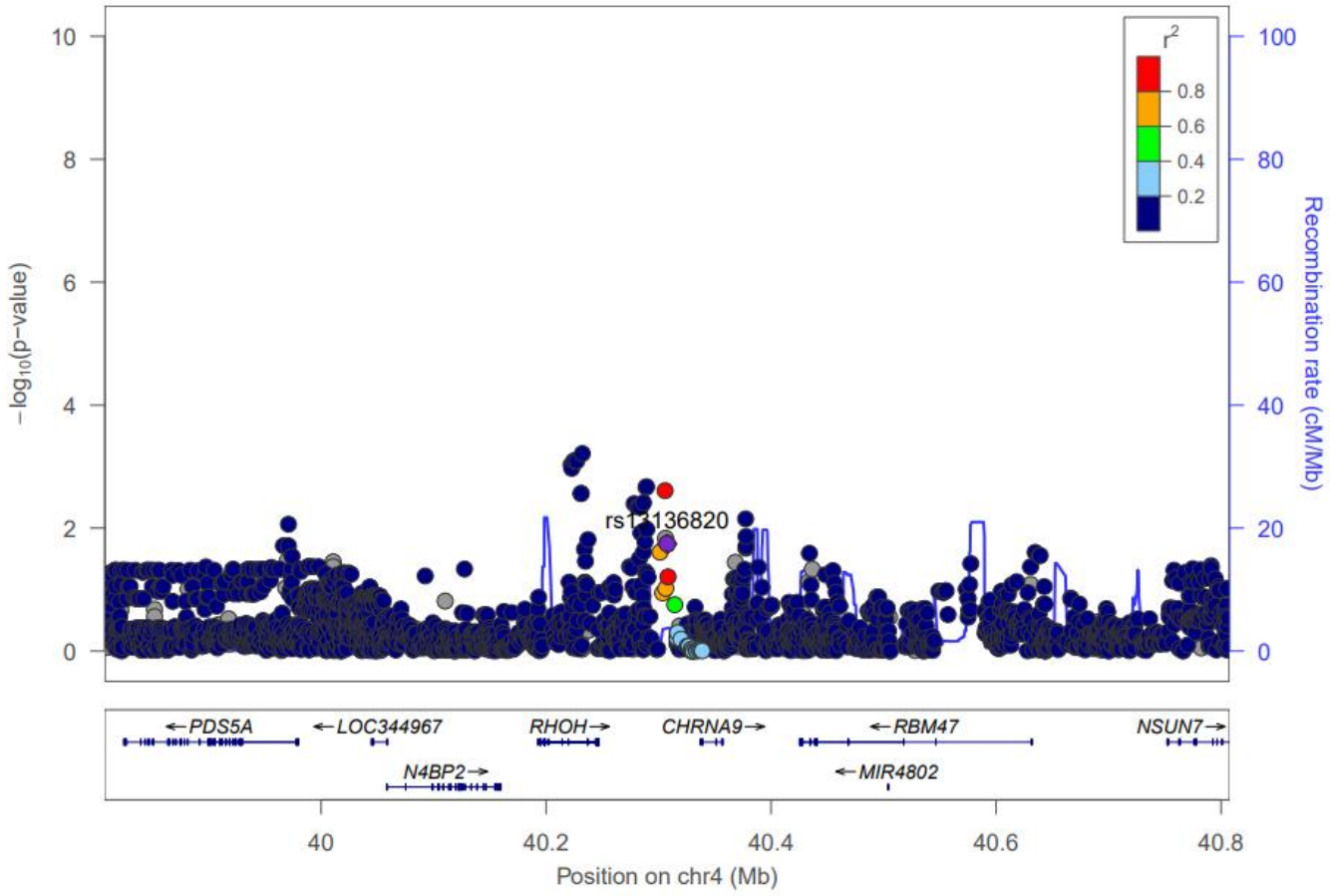


Fig 6. *Regional plot of the CHRNA9 region.*

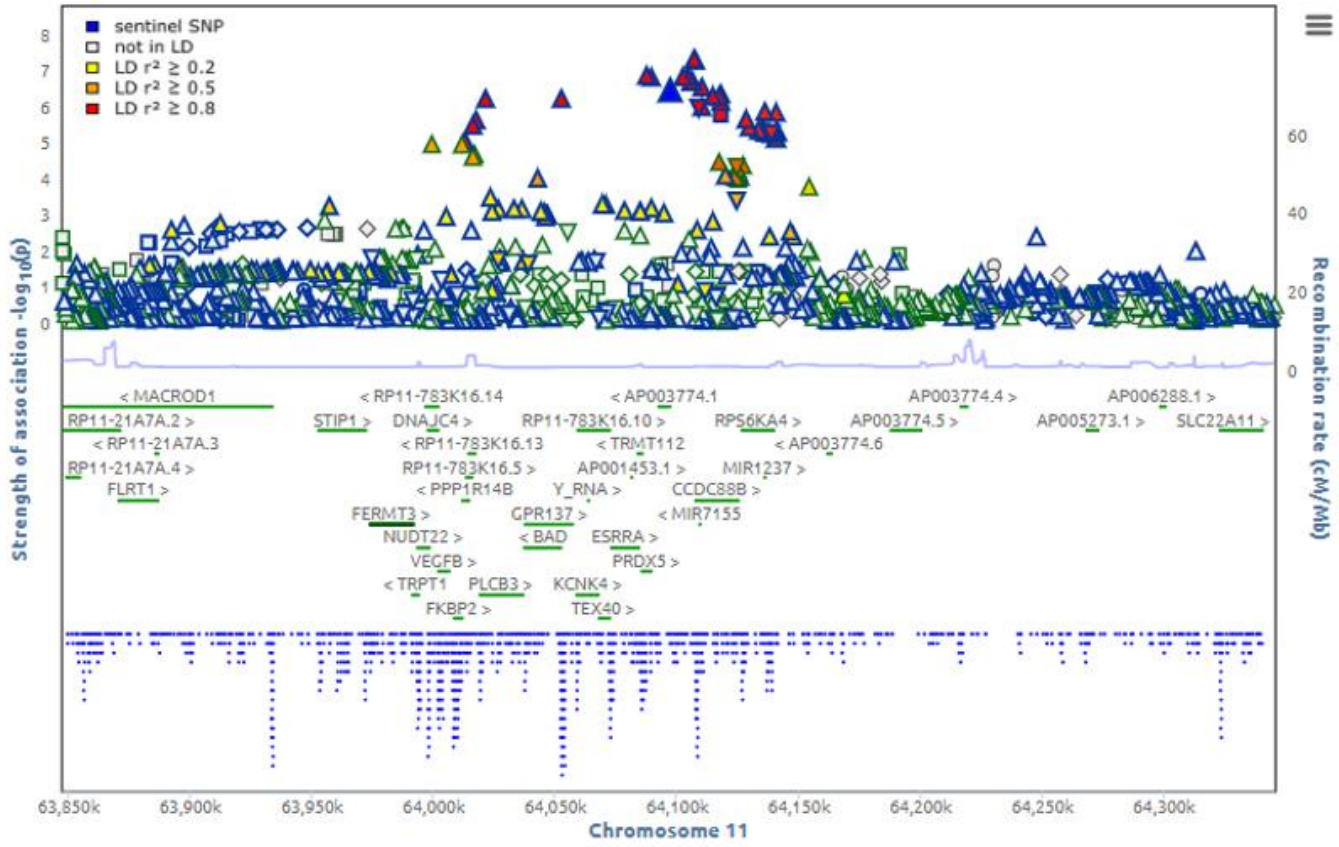


Fig 7. *Regional plot of the PRDX5 region.*

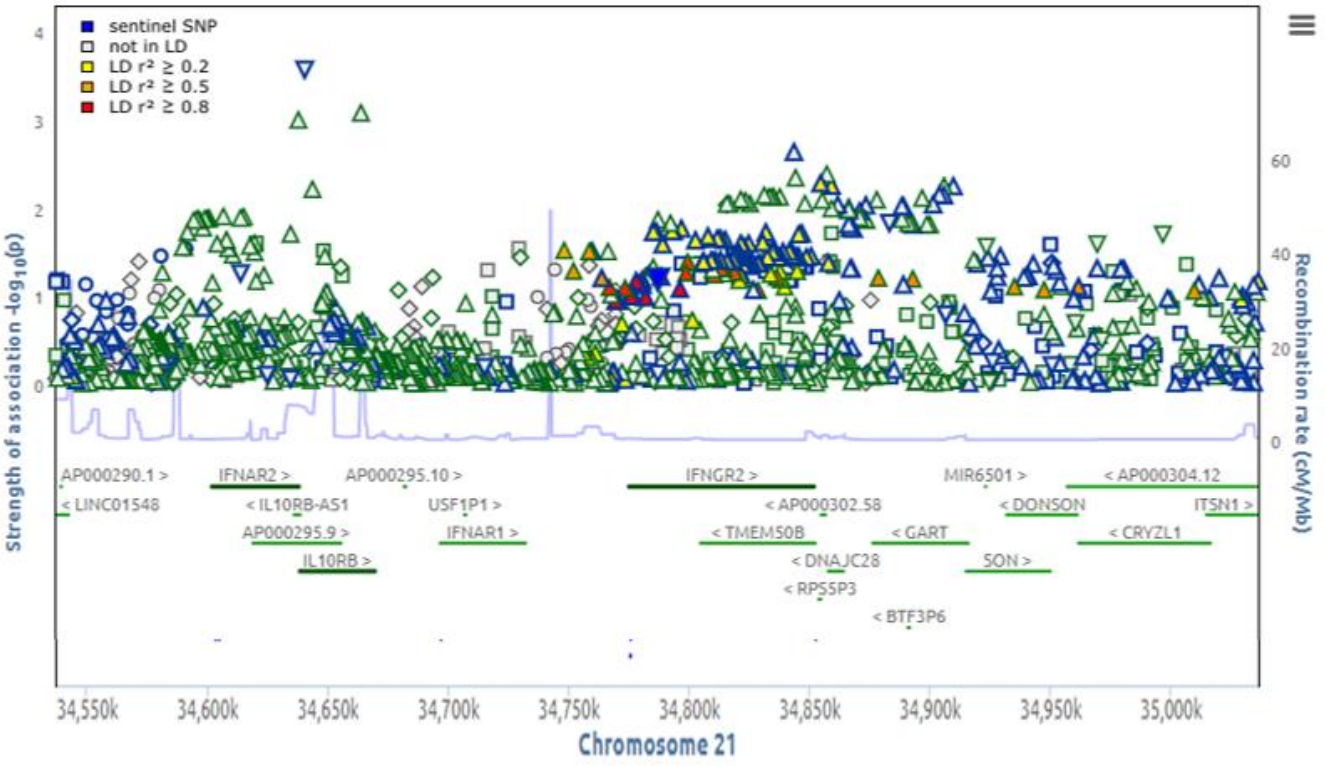


Fig 8. *Regional plot of the IFNGR2 region.*

Library Preparation

The oligo library for the MPRA assay (OLS) was designed following the guidelines from published works (Choi et al., 2020). For each of the 83 variants analyzed by this assay, we considered 145 bases encompassing the mutation to design the probes. They were followed by KpnI and XbaI restriction sites and a Tag sequence, a random 10bp region that does not affect the sequence to differentiate each probe in the later analyzing steps. At the extremities, Primer1seq and Primer2seq were added. These are sequences identical in each probe important for the following step of emulsion PCR. Indeed, thanks to these regions we were able to design primer pairs that could anneal to all the probes and add SFil restriction sites for further manipulations. At the end of the design, each probe had a length of 200 bp.

Fig.9 *Schematic representation of each probe*

Name	**Sequence**	**Length**
Primer1seq	ACTGGCCGCTTCACTG	16 bp
Variant	Variable	145 bp
KpnI site	GGTACC	6 bp
XbaI site	TCTAGA	6 bp
Tag	Variable	10 bp
Primer2seq	AGATCGGAAGAGCGTCG	17 bp
Total		**200 bp**

Table 1. *Sequences of the oligonucleotide library*

For each variant, the probe structure was designed 10 times for both the forward and the reverse strand, obtaining 20 probes for each variant differentiated by the TAG region. These probes were named alternative because they refer to the alternative SNP of the sequence. The same procedure was performed for the

reference probe, which contains the Hg38 allele. Then, for each variant 20 scramble, probes were designed following the same path, where the variant sequence contains a core of 19 bases encompassing the SNP, which acts as a negative control. This is equivalent to a total of 60 unique probes per variant, and as in the MPRA assay, we wanted to test 83 variants the library is composed of 4980 probes in total.

After receiving the library from Agilent Technologies, we performed emulsion PCR to amplify the oligo library and to add the SfiI restriction sites necessary for further steps of digestion. The emulsion PCR (ePCR) is a particular type of PCR that allows the amplification of DNA molecules in physically separated picoliter-volume water-in-oil droplets. In this way each DNA molecule is amplified independently of the others, facilitating reactions starting from complex DNA fragment mixtures with partially similar sequences (like our library) and avoiding the formation of chimeras and other artifacts between similar DNA sequences.

Two distinct vectors, namely pMPRAdonor1 and pMPRAdonor2 (Addgene vectors), were employed in the analysis, both containing an open reading frame (ORF). Donor1, which exclusively comprises the LUC gene, will be utilized to examine the effect as a promoter. On the other hand, Donor2, which contains the LUC gene coupled with a minimal promoter, will be employed to investigate the enhancer effect.

Next Generation Sequencing through Illumia DNA Prep and Nextera Index Kits

Illumina DNA Prep and Nextera XT Index Kits were chosen to perform NGS in our experiment after ePCR, to assess if all the probes are equally represented: the region analyzed is the Tag sequence, different in each probe. The first kit uses a bead-based transposome complex to tagment genomic DNA and add adapter sequences at the extremity of the fragments. Following tagmentation, a limited-cycle PCR adds the Nextera index adapter sequences to the ends of each DNA fragment. The subsequent Sample Purification Beads cleanup (SPB) purifies the libraries so that they become ready for the following loading into the MiSeq instrument.

This analysis recommends PCR amplicon > 150bp and < 500bp. Furthermore, during the tagmentation process, the adapters can’t be directly added to the distal ends of the fragments, but a drop in sequencing coverage of

~50bp from each end is expected: to ensure sufficient coverage of the amplicon target region, primers used must extend beyond the target region by 50bp per end.

Preparation of two distinct MPRA constructs

Following ePCR, the amplified library must be cut by SfiI restriction enzyme, so that it can be ligated with pMPRA vector one, which is previously, digested with SfiI, to amplify the vector in electrocompetent *E.coli* cells. When pMPRAvector1 ligated with ePCR products (corresponding to the OLS) was extracted from E.Coli cells, it then had to be digested with KpnI and XbaI restriction enzymes, whose sites are present in the original oligo library probes. pMPRAdonor1 and pMPRAdonor2 plasmids were used for MPRA assay as suggested by literature (Melnikov et al., 2014) to obtain *the LUC2* gene and *LUC2* coupled with the minimal promoter. Both plasmids were firstly amplified through PCR using a couple of primers that anneal at the extremities of XbaI and KpnI sites and then the amplicons were digested in combination with both enzymes. After the digestion, the products were run on a 1% agarose gel to confirm the digestion and then the band corresponding to *the LUC2* gene (1738bp) and *LUC2*+MiniP (1780bp) were excised and purified through NucleoSpin® Gel and PCR Clean-up kit. The products obtained from the previous digestions (pMPRAvector1+OLS, *LUC2,* and *LUC2*+MiniP) were ligated to obtain the final plasmids that will be used to transfect cell lines of interest cells: pMPRAvector1+OLS+*LUC2* and pMPRAvector1+OLS+MiniP+*LUC2* ***(Figure 10)***

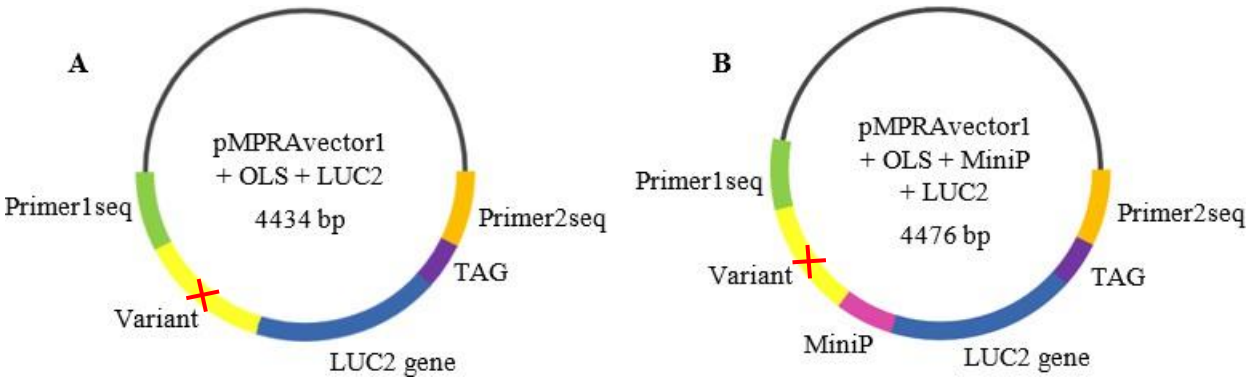


Figure 10. Schematic representation of pMPRAvector1+OLS+*LUC2* (**A**) and pMPRAvector1+OLS+MiniP+ *LUC2* (**B**). **A**: considering the variant located in a putative promoter region. **B**: considering the variant located in a putative enhancer/silencer region

Afterwards, both plasmids were chemically transfected into *E.Coli* cells for amplification.

Cell Lines Transfection

We selected HEK293T as a testing ground for our completed MPRA contructs to test the transfection efficiency. Then after confirming the efficacy of our methodological approach, we selected the Jurkat cell line as a disease-relevant cell. HEK293T (Human Embryonic Kidney 293T cells) are commonly used in gene expression studies due to their high transfection efficiency, robust growth, and ease of culture. They are widely used in functional genomics and gene editing experiments, making them an ideal choice for initial high-throughput screening. Jurkat cells are an immortalized cell line of human T lymphocyte cells. They are commonly used in immunology studies and research on T-cell signaling, making them relevant for understanding gene expression changes in immune-related contexts. The utility of the Jurkat cell line extends to studying autoimmune diseases such as multiple sclerosis (MS). MS involves aberrant immune responses, including T-cell activation and signaling. With their well-characterized signaling pathways, Jurkat cells offer a model to investigate these dysregulated processes. Understanding T cell behavior in MS through Jurkat cell studies can enhance comprehension of the disease mechanisms and aid in the development of targeted treatments (Montano, 2014). The Jurkat cells were transfected with the library-containing plasmid for both pMPRAdonor1 and pMPRA donor2. These cell lines were cultured and transfected using the Neon Transfection System (Thermo Fisher Scientific). Each experiment testing the Promoter effect and Enhancer effect was done quadruplicated for statistical optimization. Further optimizations were applied regarding the number of cells used to guarantee at least 100x representation of transfected cells relative to the number of analyzed probes. Cells were left to express the plasmid for 48 hours before harvesting.

RNA extraction and Retrotranscription

After 48 hours, we proceeded with RNA isolation using the miRNeasy tissue/cells Advanced mini-Kit. DNase treatments were performed to remove DNA residues. Reverse transcription was carried out using SuperScript™ II Reverse Transcriptase to synthesize cDNA. Before NGS, a PCR reaction with Herculase was conducted to amplify the region of interest and add sequencing adapters. Purification of PCR products was done using a Gel and PCR clean-up protocol from Macherey Nagel prior to NGS to remove excess salts.

Final Next Generation Sequencing

At the end of MPRA assay, the cDNA obtained from the transfected cells and the input DNA used for the transfection are analyzed through NGS, to determine the effect of each variant on *LUC2* expression. In this analysis, as well as for the first NGS, we focused on the TAG region, present in both the cDNA and the input DNA because cloned at the 3' end of *LUC2* gene.

For the sequencing, we used the same kits as for the first NGS, Illumina DNA Prep and Nextera XT Index.

As this is an amplicon-based analysis, we started the procedure by performing a PCR in order to amplify only the portion that we are interested in and add at the extremities of the target regions the suitable adapters for the NGS sequencing.

MPRA data analysis

Following the completion of the NGS (Next-Generation Sequencing) step, which generates FASTQ files containing the expression counts for the TAG region in each experiment, several in silico analyses were performed. Initially, an in-house script was employed to determine the frequency of occurrence for each TAG.

Next, the mpralm tool was used, which generates three distinct tables by comparing the Reference sequence with the Alternative sequence, as well as the Alternative and Reference sequences with the Scramble. This comparison allows deriving a LogFC (log-fold change) and an Adjusted P value. These values serve as the basis for filtering our probes. (Myint et al. 2019) By applying a significance filter of FDR < 0.01 and focusing solely on probes that demonstrate a significant deviation from the null (scramble), we can identify SNPs that exert a significantly different impact on gene expression between the two allelic variants (Alternative vs. Reference sequence).

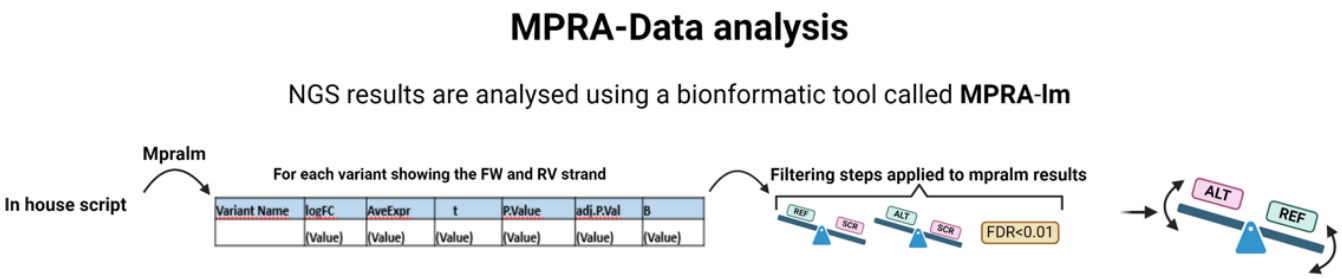


Fig.11 *Workflow of MpraLM*

MotifBreakR

MotifbreakR is a tool that enables the assessment of the similarity between the sequence surrounding a polymorphism or mutation and known transcription factor binding sites. It allows users to determine the amount of information gained or lost between different alleles of a polymorphism or between a mutation and the wildtype sequence. (Coetzee et al. 2015) MotifbreakR operates with position probability matrices (PPM) derived from position frequency matrices (PFM). PPM represents the fractional occurrence of nucleotides (A, C, G, and T) at each
position in the PFM. By using a PPM, probabilities can be generated based on the genome, allowing for the creation of position-specific scoring matrices (PSSM). These matrices utilize the information in the PPM to estimate the likelihood of observing a particular nucleotide at a specific position within a transcription factor-binding site. Figure x. Shows the R script to run the MotifbreakR package on R studio. The threshold is the maximum p-value for a match to be called. The method "default" represents the algorithm used to calculate results, which in this case is the weighted sum, so the difference in the probabilities for the two letters of the polymorphism (or variant). The for-loop allows the automatization of the process to analyze a large SNPs list.

```
library(motifbreakR)
#pca.snps.file <- system.file("extdata", "pca.enhancer.snps", package = "motifbreakR")
#pca.snps <- as.character(read.table(pca.snps.file)[,1])

library(BSgenome)
#available.SNPs()

library(SNPlocs.Hsapiens.dbSNP144.GRCh37) # dbSNP144 in hg19
library(BSgenome.Hsapiens.UCSC.hg19)     # hg19 genome

#Lista varianti da personalizzare, quelle da usare con motifbreakR:
list_variants<-c("rs1883832","rs6065926","rs2024568",
                 "rs6032662","rs6032663","rs6032660",
                 "rs6032664","rs6074022","rs1569723",
                 "rs4810485","rs4239702")

#Qui inizia il for-loop sulla lista delle varianti:
for (j in list_variants) {

print(paste0("Starting job for SNP: ",j))
snps.mb <- snps.from.rsid(rsid = j,
                          dbSNP = SNPlocs.Hsapiens.dbSNP144.GRCh37,
                          search.genome = BSgenome.Hsapiens.UCSC.hg19)

#snps.mb

#library(BSgenome)
#genomes <- available.genomes()
#length(genomes)
#genomes
#library(MotifDb)
#MotifDb
#table(mcols(MotifDb)$organism, mcols(MotifDb)$dataSource)
#data(motifbreakR_motif)
#motifbreakR_motif
results <- motifbreakR(snpList = snps.mb, filterp = TRUE,
                       pwmList = subset(MotifDb,
                                        dataSource %in% c("HOCOMOCOv11-core-A", "HOCOMOCOv11-core-B", "HOCOMOCOv11-core-C")),
                       threshold = 0.01,
                       method = "default",
                       BPPARAM = BiocParallel::SerialParam())
print(paste0("MotifBreakR for SNP ",j, " concluded. Now calculating p-value..."))
rs_variant <- results[results$SNP_id == j]
rs_variant_1 <- calculatePvalue(rs_variant, granularity = 1e-6)


#plotMB(results = rs1883832_1, rsid = "rs1883832", effect = "strong", altAllele = "C") #dà errore: Too many stacks to draw.
Either increase the device size or limit the drawing to a smaller region.

results_df<-as.data.frame(rs_variant_1,row.names=seq(length(rs_variant_1)))
results_df <- apply(results_df,2,as.character)

print(paste0("Saving results for SNP: ",j))
#Salvataggio dei file sia in formato txt che csv, potete anche escludere uno dei due:
write.table(results_df, file=paste0("results_motifs_",j,".txt"), row.names= FALSE) #aprire txt da excel
write.csv(results_df,paste0("results_motifs_",j,".csv"))

rm(snps.mb,results,rs_variant,rs_variant_1,results_df)

} #fine for-loop
```

Fig.12 *MotifBreakR script*

After generating separate Excel tables for each analyzed SNP, we applied filtering based on criteria established in the study conducted by Long et al. in 2022. Specifically, we filtered the Ref p-value and Alt p-value, ensuring that they were both below 0.01. Additionally, we considered the absolute value of

alleleDiff, ensuring it was greater than 0.7. This filtering process allowed us to select the relevant results for further analysis and interpretation. The description of the meaning of column names found in the Excel tables is shown in Table 2.

REF	The reference allele for the variant
ALT	The alternate allele for the variant
snpPos	The coordinates of the variant
MotifPos	The position of the motif relative the the variant
GeneSymbol	The geneSymbol corresponding to the TF of the TF binding motif
DataSource	The source of the TF binding motif
ProviderName, providerId	The name and id provided by the source
SeqMatch	The sequence on the 5' -> 3' direction of the "+" strand that corresponds to DNA at the position that the TF binding motif was found.
pctRef	The score as determined by the scoring method, when the sequence contains the reference variant allele, normalized to a scale from 0 - 1. If filterp = FALSE, this is the value that is thresholded
pctAlt	The score as determined by the scoring method, when the sequence contains the alternate variant allele, normalized to a scale from 0 - 1. If filterp = FALSE, this is the value that is thresholded.
ScoreRef	The score as determined by the scoring method, when the sequence contains the reference variant allele
ScoreAlt	The score as determined by the scoring method, when the sequence contains the alternate variant allele
Refpvalue	p-value for the match for the pctRef score, initially set to NA. see calculatePvalue for more information
Altpvalue	p-value for the match for the pctAlt score, initially set to NA. see calculatePvalue for more information
altPos	The position, relative to the reference allele, of the alternate allele
alleleDiff	The difference between the score on the reference allele and the score on the alternate allele
alleleEffectSize	The ratio of the alleleDiff and the maximal score of a sequence under the PWM
Effect	One of weak, strong, or neutral indicating the strength of the effect. each SNP in this object may be plotted with plotMB

Table 2 *Legend of the column names found in the excel files obtained as results from the MotifBreakR tool*

MPRA experiment on the Jurkat cell line (Replicate)

Initially, we performed the MPRA experiment on the Jurkat cell line given their relevance to the disease, which yielded significant results. However, due to the in vitro properties and cell-specific effect, to have a further affirmation of our preliminary results, the MPRA experiment in the Jurkat cell line, was replicated for further evaluation of our results. The methodology was applied in the same manner for both experiments following the guidelines mentioned (Choi et al., 2020b). Afterward, we applied statistical methods to confirm the replicability of our two distinct experiments and the selection of the probes to be further followed up coupled with different annotation scores. For the annotation of the replicated variants, we took into consideration information from UCSC, Regulomedb, and Screen Registry V3.

Complete MPRA Workflow.

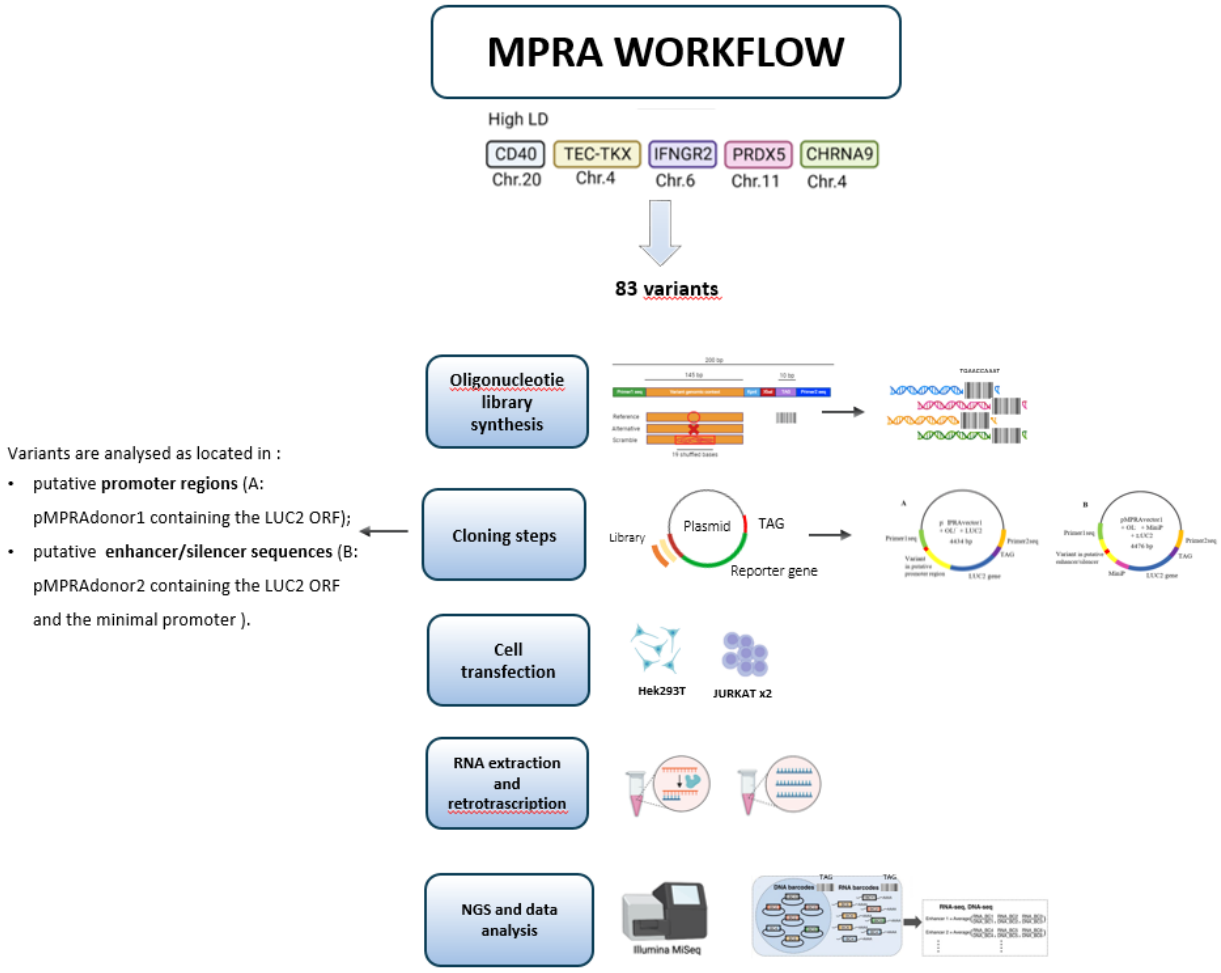


Fig.14 *MPRA workflow*

Complete MPRA Data Analysis Workflow

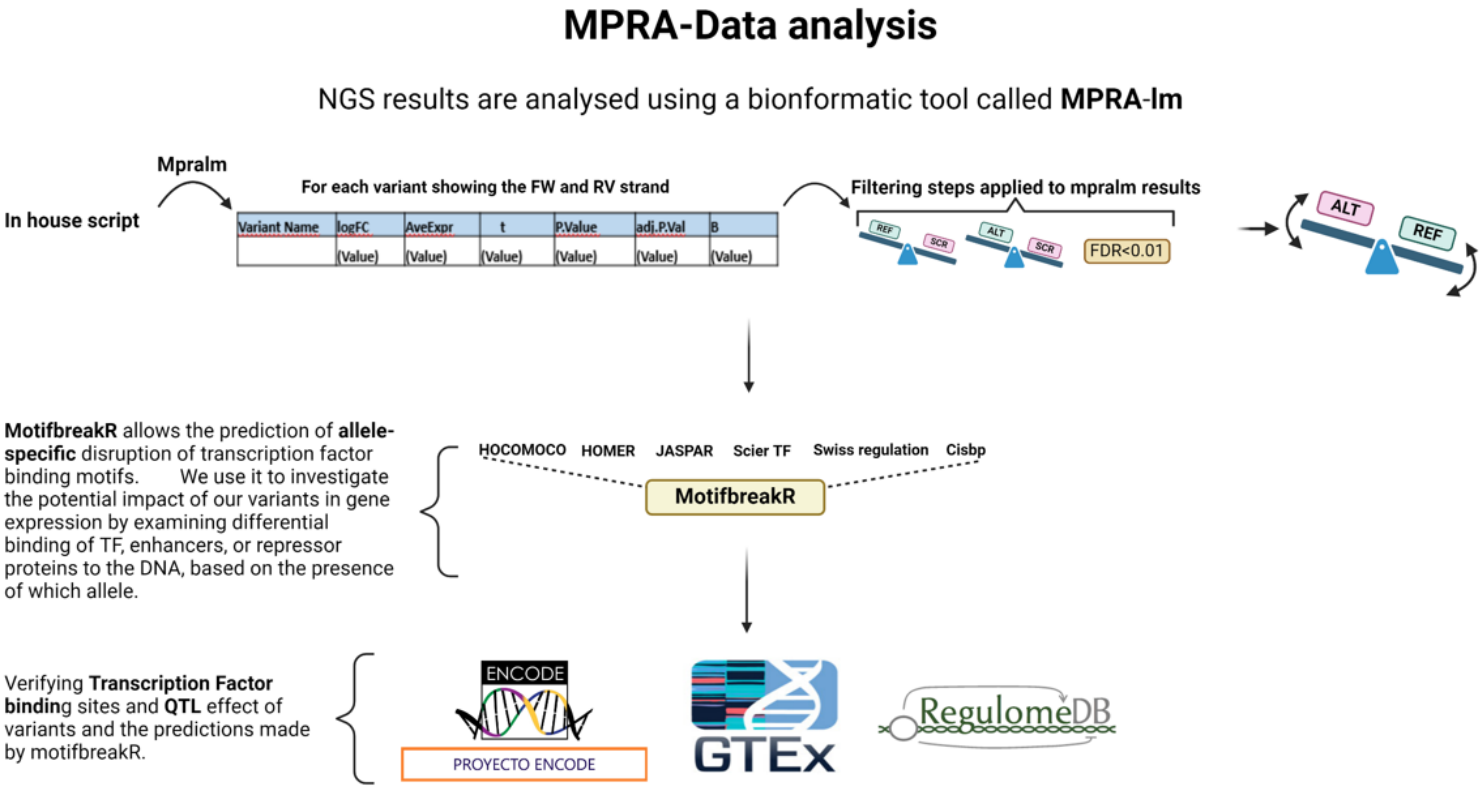


Fig.15 *MPRA data analysis workflow*

MPRA experiment with the Epstein Barr Nuclear Antigen 2 variant on the Jurkat cell line

To measure a possible effect of the EBNA2 gene, specifically the 1.2 variant, which has been described by our collaborators in previous works as a variant, that is more prominent on MS patients, we decided to perform a sequential transfection on the Jurkat cell line. Firstly, we performed transfection of the EBNA2 1.2 variant that was cloned into a plasmid containing also the GFP gene using the Neon Transfection System.

We confirmed the integration of the plasmid into the cells by checking the fluorescence of GFP. After 24 hours of the EBNA2 transfection, we sequentially performed on the same cells, the transfection with the MPRA construct, specifically the pMPRAdonor 1 construct which measures the effect of our variants as promoters. We decided to experiment following the same guidelines from our previous work in quadruplicates for statistical optimization. At the same time in parallel to this experiment to have a positive control, which would allow measuring of the effect of the EBNA2 variant, we performed a transfection with an innocuous GFP which puts the experiments in the same conditions as the EBNA2 experiment. After 24 hours we transfected the same cells that contained only the GFP with the pMPRA donor1. Transfection protocols and cell culture procedures were followed as indicated by available literature. Before harvesting, we examined the wells containing our cells (EBNA+Cells/GFP+Cells) under the microscope to detect green fluorescence, a clear indicator of successful GFP (Green Fluorescent Protein) plasmid expression, confirming robust transfection and active cell viability. 48 hours post-transfection, we performed RNA extraction using (miRNeasy tissue/cells Advanced mini-Kit) corresponding to cell count being ≥5mln cells for the sample, following the manufacturer's protocol which allowed us to obtain a sufficient quantity of RNA. Further steps including NGS, Mpralm, and MotifBreakR were followed as described before to carry on the complete MPRA analysis. Fig. 16

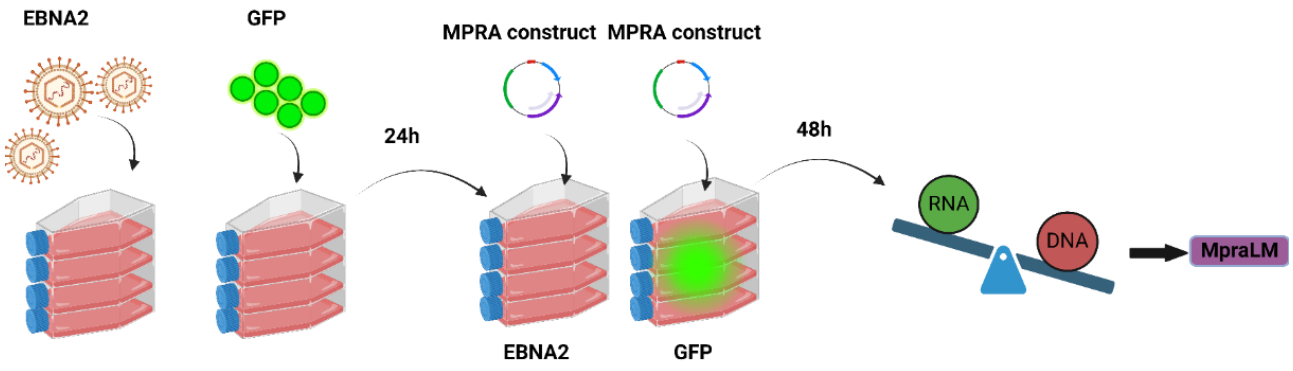


Fig. *16 Describes the Workflow of the MPRA experiment with the Epstein Barr variant 1.2.*

Evaluation of MPRA results by dual glow Luciferase assay

In order to confirm the results from our massive evaluation we decided to perform the Dual-Glo Luciferase Assay System to confirm singularly all the effects noted of the variants from MPRA. The Dual-Glo system allows fast and simple quantitation of a stable luminescent signal from two reporter genes in a single sample. The convenient protocol generates both firefly and Renilla luciferase luminescence signals from cells that have not been preconditioned ore prelysed. This allows that the internal controls can minimize sample variability by reducing false-positive and false-negative readings caused by nonspecific factors.

In order to perform this assay we had to go through the usual steps of plasmid amplification of both Renilla and the plasmid containing the variant we wanted to test. We acquired from Genscript the sequence's same as to the one tested by MPRA, with the same length and same composition for both alternative and reference. Respectively the probes named by us as 9_CD40, 6_CD40, 6_TEC-TKX, 14_IFNGR2, 5_PRDX5, 12_PRDX5, 8_PRDX5, 7_IFNGR2. Initially we performed a step of digestion using the restriction sites for NheI and BglII on all of the probes to select the sequence of our interest. This was followed by a step of ligation between the DNA fragments and a plasmid of interest, following the manufacturer's instructions NEB. Post-ligation, we confirm the ligation by gel electrophoresis. Afterwards, we amplified the ligated product using a strain of the E.coli cells (DH5α), which is followed by colony selection by picketing of which we amplified and sequenced the fragment of interest within the plasmid to confirm the correct fragment. As a final step, we amplified again only the colonies containing the correct sequence in order to perform the transfection. We chose for this experiment the Jurkat cell line in order to observe the same effect as in the MPRA experiment. We transfected 4 mln cells in triplicates for each of the fragments to be tested. We performed a co-transfection in a ratio of 3:1 where our fragment represented the majority and the rest was the renilla plasmid. We transfected 10 ug of plasmid per replicate (7.5 ug of fragment plasmid and 2.5 of Renilla plasmid) using the Neon Transfection System. We waited for 48 hours and then pelleted the cell in order to perform a step of washing. After washing, we re-suspended the cells in 100ul of PBS and added the firs reagent 100ul of Dual Glo Luciferase Reagent and we incubate for 10 minutes then perform the first read of the firefly luminescence. Then we added 100ul of the Dual Glo Stop reagent and after 10 minutes, we measured the Renilla luminescence. Statistical analysis was performed on excel.

Results

Fine mapping and MPRA in Identifying functional variants in complex regions associated with MS.

Identifying functional variants in MS associated regions carrying Drug Target genes through Fine Mapping.

Our approach in trying to identify causative variants amongst non-causative variants in MS-associated regions containing potential drug target genes started with a large cohort, which was previously genotyped using array platforms with a genome-wide coverage of SNPs. This initial dataset was then subjected to genotype imputation against the Haplotype Reference Consortium panel to generate a more comprehensive SNP map for the subsequent fine-mapping. The final cohort consisted of 5,903 individuals from the continental Italian population, including 4,259 MS patients and 1,644 healthy controls, with 6,339,414 imputed SNPs covering the entire genome. We wanted to identify potential drug target genes (genes that are targets of already approved drugs or in clinical trials) within the regions associated with MS, so we cross-mapped replicated MS regions (statistically significant with a p-value <0.05, 2Mb-wide) with genes listed in the Drug-Gene interaction database v4.2. This analysis yielded 36 regions surrounding the replicated SNPs, which contained a total of 238 druggable genes. A gene was considered druggable if it overlapped with the MS-associated region even if one base fell within the 2Mb-wide region. To further investigate these 36 regions and determine which SNP present in the regions shows evidence of "causality" and by unraveling the linkage disequilibrium (LD) effect from the statistical association between SNPs and Multiple Sclerosis, we applied two statistical fine mapping tools: Paintor and CaviarBF. Both tools utilize a Bayesian framework, which instead of assigning a p-value to a SNP as most conventional methods do, it assigns a posterior probability of causality, known also as posterior inclusion probability (PIP). Each SNP within the region starting from the summary statistics of a GWAS study has its own PIP. For a more accurate functionally informed fine mapping, we included a set of annotation scores, to take into account the biological relevance of each SNP within the region. For this purpose, we incorporated GWAVA, CADD, FINSURF, and Regulome Db annotation scores into the fine mapping analysis. After the analysis, each SNP had their respective PIP value, which allowed us to sort the SNPs by decreasing PIP order and we were able to construct the 90% credible set by summing the ranked PIP values until it reached 0.90. For more accurate analysis we selected only loci that exhibited a minimum overlap of 75% in the credible sets identified by both tools. More specifically the SNPs that were revealed to have a high PIP value from both tools and demonstrated an association p-value <1e-4.

This gave us a promising list of SNPs that are most likely to be associated with or to affect disease pathogenesis or progression as predicted by the integration of two fine mapping tools.

Once we identified the SNPs in the credible sets, we had to determine which genes they potentially influenced. For this, we conducted SNP-to-gene mapping using the Open Targets Genetics Database. Open Targets is an innovative, large-scale, multi-year, public-private partnership that uses human genetics and genomics data for systematic drug target identification and prioritization. The Open Targets Genetics is an option of the website that allows identifying targets based on GWAS and functional genomics data from various cell types and tissues. Through this, we were able to obtain specific information for each SNP of interest, including the nearest genes and target genes (genes that are affected in the expression by the presence of the SNP), the distances between the SNP and the transcription start site (TSS), and the V2G global score. The V2G global score reflects the level of evidence supporting the functional impact of a SNP on a particular gene. We obtained 19 regions that contain SNPs with compelling evidence of causality with statistical significance (Paintor PIP>0.7). Among these 19 regions, 18 displayed as a possible causal SNP a different one than the Lead SNP that is the SNP showing the most statistically significant association in the original IMSGC article. Among these, 10 regions have causal SNPs that target drug-target genes. *IFNGR1, TUBB4A, TEC, TGFBR3, IDE, and CD40* are the genes that are targets of drugs approved for other diseases but also of the SNPs that are shown as "causal". Additionally, in four regions the "causal" SNPs target potential drug target genes (*EOMES, FCRL3, CTSH,* and *ADCY3*) according to Open Targets Genetics. These genes are regarded as promising druggable targets due to their ability to encode enzymes, cell surface molecules, or their association with molecules currently undergoing preclinical investigations. For example, the EOMES gene, for which a molecule known as ntEomes-TMD has shown therapeutic potential in EAE mice, models (Shin et al. 2022). As well as the CTSH gene where among its related pathways is also involved in the Innate Immune System, and has been associated with other diseases such as Alzheimer's, which makes it worthwhile as a therapeutic targets (Nelson et al., 2020). FCRL3 interacts with target proteins of current multiple sclerosis medications, showing its involvement with the disease and suggesting that it might be a promising drug target (Jianfeng L. et al 2023).

Lead SNP	CHR	Best PIP PAINTOR	SNP	Target Gene	Drug Target
RS_12206850	Chr 6	1.00	rs13197384	AHI1	NO
RS_17066096	Chr 6	1.00	rs1327474	IFNGR1	YES
RS_438613	Chr 3	1.00	rs1813375	EOMES	POTENTIALLY
RS_3761959	Chr 1	0.99	rs945635	FCRL3	POTENTIALLY
RS_4468527	Chr 14	0.99	rs75478143	GNG2	NO
RS_1077667	Chr 19	0.98	rs73922419	TUBB4A	YES
RS_17470892	Chr 4	0.97	rs11942525	TEC	YES
RS_9843355	Chr 3	0.97	rs1132200	TIMMDC1	NO
RS_9887787	Chr 1	0.96	rs2038931	TGFBR3	YES
RS_7923837	Chr 10	0.95	rs72811230	IDE	YES
RS_58394161	Chr 1	0.93	rs58394161	RPAP2	NO
RS_10152892	Chr 15	0.87	rs876941	CTSH	POTENTIALLY
RS_12087340	Chr 1	0.86	rs11161583	BCL10	NO
RS_17119	Chr 6	0.83	rs2144684	JARID2	NO
RS_2082881	Chr 2	0.82	rs41281519	ADCY3	POTENTIALLY
RS_6832151	Chr 4	0.77	rs35818439	NSUN7	NO
RS_694739	Chr 11	0.74	rs663743	PPP1R14B	NO
RS_12928822	Chr 16	0.72	rs9923623	RMI2	NO
RS_6032552	Chr 20	0.71	rs6065926	CD40	YES

Table 3. *List of the nineteen regions that show at least one SNP with a PIP from Paintor higher than 0.77, along with their target gene.*

Identifying functional variants in MS associated regions through in vitro techniques: MPRA.

Simultaneously with the fine mapping analysis, we performed the MPRA experiment involving 5 out of the 36 MS-associated regions which contained druggable regions. Regions experimentally tested are represented below:

Region Name	Chromosome	Lead SNP	Drug Target Gene	No. of testes SNPs
CD40	Chr20	Rs6065926	CD40	12
PRDX5	Chr11	Rs694739	CCDC88B	24
IFNGR2	Chr21	Rs9808753	IFNGR2	22
TEC-TKX	Chr4	Rs17470892	NFXL1	23
CHRNA9	Chr4	Rs13136820	RHOH	5

Table 4: *Regions and thei respective SNPs selected for MPRA evaluation.*

We selected these 5 regions due to their architectural complexity due to the large number of SNPs in high linkage disequilibrium with the lead SNP of the region. This analysis was performed independently from the statistical analysis, however after the fine mapping analysis we confirmed two regions that were also within the credible set predicted by fine mapping, were also selected for the MPRA analysis, respectively the CD40 region and the TEC region. The total number of SNPs tested across these 5 regions was 83, and the filtering criteria for a SNP to be considered in high LD with the Lead SNP was r2≥ 0.77. MPRA enables the simultaneous functional testing of numerous potential regulatory elements, using the basis of the conventional luciferase assay in a high throughput manner. The basis of this technique consists in creating a library in which each tested variant is represented by an oligo sequence containing either the reference or alternative nucleotide in a region of 145 bases. The novelty of the technique lies in the fact that each sequence is represented by a 10 bp nucleotide sequence called a TAG that serves as a barcode to distinguish each oligo sequence. Each SNP to be tested is portrayed by a 145 base pair sequence containing either an Alternative or Reference allele and by a sequence with the same length that contains a 21 base pair segment surrounding the variant nucleotide, which serves as a null hypothesis sequence later on called Scramble. Each of either Reference, Alternative, or Scramble is reflected 10 times in both DNA strands, resulting in 60 representations for each SNP to be tested, leaving in total of 4980 probes each with a unique bar code. To test each of the variants, we utilized two distinct vectors, namely pMPRAdonor1 and pMPRAdonor2. Both constructs were created to contain the library of SNPs to

be tested however; pMPRAdonor1 contained only the LUC gene Open Reading Frame (ORF) and was utilized the test the variant's effect as promoters. pMPRAdonor2 contains the LUC gene coupled with a minimal promoter, which will investigate the enhancer effect of each tested variant. Once the construct was ready, it was transfected to 2 different cell lines, respectively HEK293T for transfection testing and JURKAT cell lines as a disease-relevant cell line which was performed 2 times for experimental evaluation. Cells were transfected in quadruplicates for statistical significance for both PMPRAdonor1 and pMPRAdonor2, they were left to express the construct for 48 hours before cell harvesting. Then, we extracted and reverse-transcribed RNA from the cells. Before NGS, a PCR reaction using High Fidelity Herculase was performed to amplify the region of interest and add sequencing adapters. We performed NGS using the Illumina DNA Prep and Nextera XT index kits on the cDNA obtained from the transfected cells and input DNA was used to compare the change in reads for each barcode post-transfection.

MPRA Data Analysis

After completing the NGS step, we obtained the FASQ files, which contained the expression counts for each barcode representing each sequence in each individual experiment. We performed several in silico bioinformatics analyses starting with an in-house script, which determined the frequency of occurrence for each TAG (barcode) in each FASTQ. Afterward, after a series of modifications of the tables representing each TAG count for each sequence in each replicate we applied the mpralm tool. The mpralm tool (Linear models for differential analysis of MPRA data) (Myint et al., 2019) outperforms existing statistical methods for the analysis of this datatype. It investigates the theoretical and real-data properties of barcode summarization methods and shows an unappreciated impact of the summarization method. Since each element coming from MPRA data is measured across several barcodes, it is useful to summarize this data into a single activity measure for a single element in a single sample. So after collapsing the FASTQ crude data, we are left with a series of tables that contain the TAG count for each sequence tested. We create three distinct files that respectively contain the information for each confrontation that we want to test: Reference vs Alternative, Alternative vs Scramble, and Reference vs Scramble, for each probe tested. These files are then given to the mpralm script which will compare and perform normalization based on the representatively of each probe and confrontation between cDNA and input DNA. The cDNA to DNA ratio is each summed across barcodes, which in turn gives us the log-ratio. The log-ratio of each tested variant represents the direction of the effect in the presence of the alternative allele, indicating the effect of the alternative allele on the expression of the Luc gene. For example, with the sign of the log FC is –

(negative), it means that the alternative allele decreases the expression of the Luc gene respectively to the reference allele. The same logic applies when the log FC sign is positive; it means that the alternative allele increases the expression of the LUC gene respectively to the reference allele. Each Log-FC is accompanied by an Adjusted P value which we use to apply our filters of significance in selecting the probes with the most prominent and accurate effect. The significance filter for the p-value was at an FDR<0.01. We were able to identify SNPs that exert a significantly different impact on the expression of the Luc gene between the two allelic variants (Alternative vs. Reference sequence). However, we focused solely on the probes that demonstrated a significant deviation from the null sequence (scramble) when compared to both the Alternative and the Reference sequence. We justify this by discussing that in the cases where the scramble sequence which shouldn't have any effect on the expression of the Luc2 gene, is represented in a higher number than the alternative sequence or the reference sequence, it's probably due to an overrepresentation in the plasmid pool of the scramble sequence. Represented below is an example of the mpralm tool outcome for the rs1883832 for the first MPRA experiment on the Jurkat cell line, which emerged as significant out of the 12 SNPs tested for the CD40 region tested for the promoter effect. The first column shows that in the tested variant, the log FC demonstrates the ratio of the Alternative to Reference, by comparing cDNA/DNA input counts, with a negative sign representing that in that specific probe, the alternative allele had a lower expression as compared to the reference allele. Then, the adjusted p-value shows the significance of the probe result. For this variant, both the reference and the alternative allele probes showed a significantly higher expression in comparison with the null (Scramble).

Probe name	**Log FC**	**Adj.P.Val**
9_CD40_FW	-0,79107	0,009149

Table 5*. Result table from the mpralm tool indicating that the 9th SNP out of the 12 tested in the CD40 regions surpasses the filtering criteria and demonstrating a reduced luciferase effect in the presence of the alternative allele when compared to the reference allele on the Jurkat cell line.*

Identification of functional SNPs in Gene Expression through MPRA

To obtain valid information on any of the SNPs that we had selected to test for their functionality we decided to apply the MPRA experiment on the Jurkat cell line. Jurkat cells are immortalized T lymphocytes, which are highly valued in immunological research. They are known for their robust representation of T cell biology and T cell activation (TCA) mechanisms via T cell receptor (TCR) signaling and interleukin-2 (IL-2) production. These factors make this cell line a very interesting candidate in studying the functionality of our SNPs in this invitro assay.

For the initial purpose of our assay, we prioritized the results from this cell line among others (Hek293T and SHSY5Y) given its relevance with Multiple Sclerosis. For reassurance in our results, we decided to perform the Jurkat experiment in duplicate at different times and then measure the outcomes. Following this security step only the probes that showed the same direction of effect and surpassing the filters applied to the mpralm outcome between both experiments were selected for further analysis and evaluation. We compared the Log FC as given by mpralm of both the experiments conducted on the Jurkat cell line and observed a high correlation (R^2=0.8) between both experiments, suggesting a high reproducibility in our methodological approach. ***(Fig.17)***

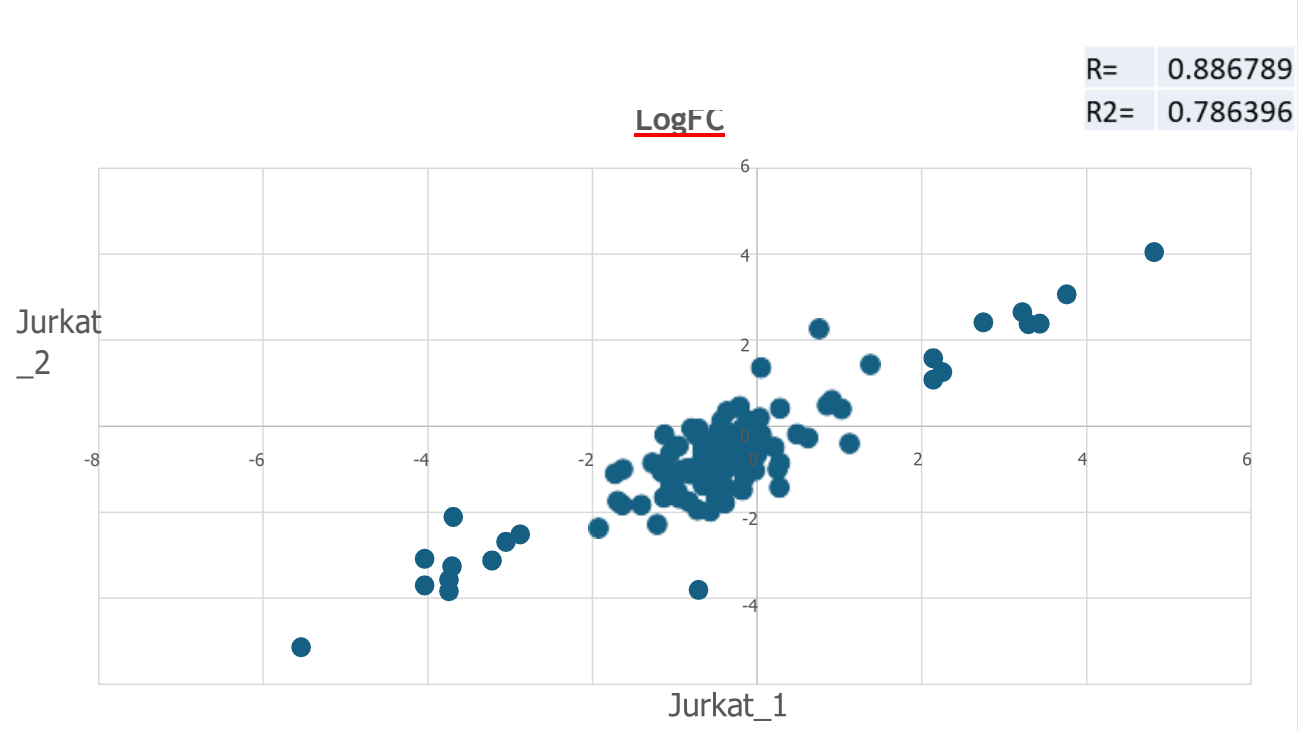


Fig.17 *Correlation graph showing the high correlation of the LOG FC between the two experiments*

This analysis proved that when we replicated the experiment the Log FC of the tested variants had a high correlation. Then to discard any probes that did not follow the same trend in both experiments we performed a simple linear regression test. The simple linear regression with a 95% confidence interval showed that the 4 probes tested were not in the confidence interval, as they exhibited a significant difference in the log FC between the two Jurkat experiments. Respectively, the four probes, which show discrepancies will not be taken into account for further evaluation as they are considered too versatile specifically; 17_ PRDX5_RV (rs4930152), 4_CD40_FW (rs2024568), 19_TEC-TKX_FW (rs17471024), 22_IFNGR2_INST_T_FW (rs17471024) and 14_PRDX5_RV (rs3815362).***(Fig.18)***

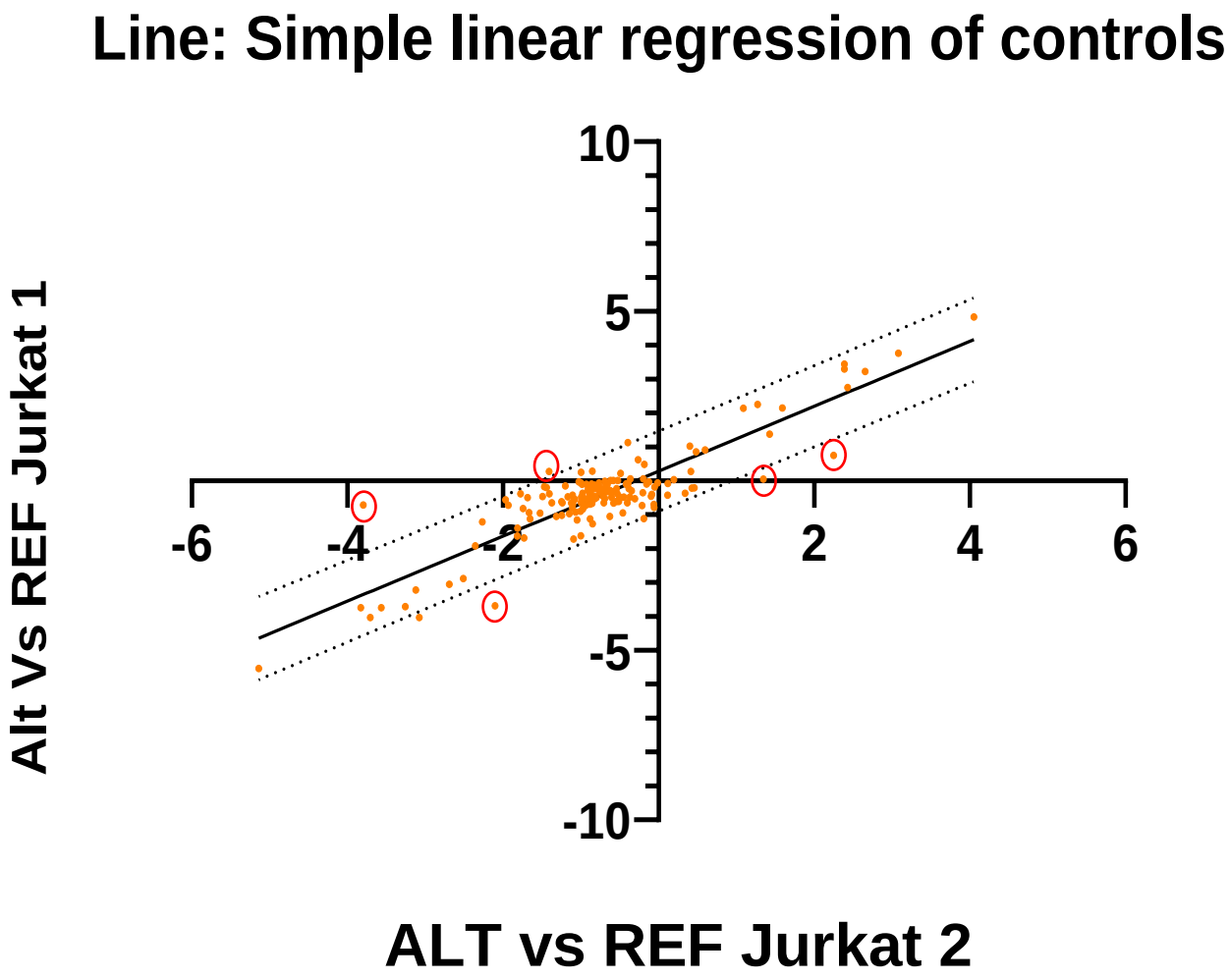


Fig.18 *Simple linear regression analysis between the Log FC of each probe tested in two different MPRA experiments on the Jurkat cell line showing the replicability of our results. Exception for 4 tested variants which show to be too versatile between experiments to be taken into account.*

Reported below are all the probes which when tested with the MPRA technique have surpassed the filtering criteria of the adjusted p-value (FDR<0.01) and show a significant difference in the expression when compared to the null hypothesis. Respectively the number of probes significant from the Jurkat 1 experiment and the Jurkat 2 experiment.

In summary, 4 out of the 5 analyzed regions (named TEC-TKX, CD40, PRDX5, IFNGR2) exhibited significant imbalances between the reference and the alternative allele probe for at least one SNP (range 1-6 SNPs for each region).

Table 6 *Number of significant probes from the 1st Jurkat experiment*

Number of significant probes on the 1st Jurkat experiment			
Region Name	**Number of significant probes**	**Promoter effect**	**Enhancer effect**
PRDX5	2	1	1
CD40	1	1	0
IFNGR2	4	1	3
TEC-TKX	1	1	0

Table 7 *Number of significant probes from the 2nd Jurkat experiment*

Number of significant probes on the 2st Jurkat experiment			
Region Name	**Number of significant probes**	**Promoter effect**	**Enhancer effect**
PRDX5	4	3	1
CD40	2	1	1
IFNGR2	6	4	2

We selected for further evaluation only the variants that showed the same direction of effect in both experiments and with a statistically significant difference from the null hypothesis and in the comparison between Alternative and Reference.

Represented below are the exact values of these variants post-data analysis showing their respective log FC. This analysis highlighted variants that show compelling potential in influencing gene expression either as promotors or enhancers.

Table 8 *Variants replicated in both experiments with the Jurkat cell line showing an effect as a Promoter.*

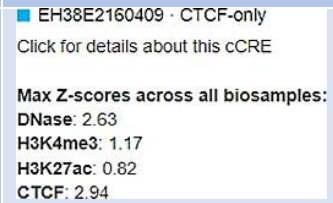
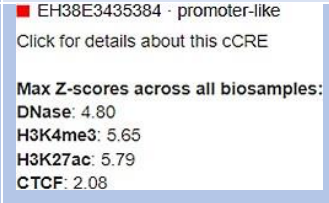
Variants Replicated with an effect as a Promoter in both experiments with Jurkat cell line						
Chromosomic position	Rs ID	Variant name	Jurkat 1		Jurkat 2	
			LogFC	Adj p-value	LogFC	Adj p-value
chr20:46118243	rs1883832	9_CD40_FW	-0,79107	0,009149	-0.05	0.8496
chr11:64317576	rs28364831	5_PRDX5_RV	-0,20274	0,334721	-1,44239	0,006306
chr21:33409989	rs28653198	14_IFNGR2_RV	-1,62785	4,82E-05	-1,81876	7,08E-05
chr11:64301969	rs72924108	8_PRDX5_RV	2,250237	8,58E-05	1,270621	0,000124
chr20:46111457	rs6074022	6_CD40_FW	-0,19725	0,332805	-0,76382	4,08E-06

Table 9. *Variants replicated in both experiments with the Jurkat cell line showing an effect as a Enhancer.*

Variants Replicated with an effect as an Enhancer in both experiments with Jurkat cell line						
Chromosomic position	Rs ID	Variant name	Jurkat 1		Jurkat 2	
			LogFC	Adj p-value	LogFC	Adj p-value
chr11:64276712	rs72922077	12_PRDX5_RV	-1,18385	0,000133	- 1.10392	0.00598
chr21:33401092	rs17880053	7_IFNGR2_INS -> G_FW	-1,24795	1,91E-05	- 0,87748	0,003449
chr4:48145192	rs17574371	6_TEC-TKX_RV	-0,49134	0,040602	- 0,76851	0,004554

These variants are selected for further evaluation using as a priority selection criteria if the variant tested in both cell lines showed the same direction of effect from Mpralm. However, variants showing the same direction of effect between both cell lines for Alternative Vs Reference demonstrated some small discrepancies (mismatches) when confronting both experiments regarding the p-value (sometimes borderline) and the filters taken into account from the null hypothesis. We hypothesize that this happens due to the nature and sensitivity of the technique which could be affected by various

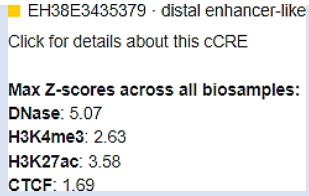
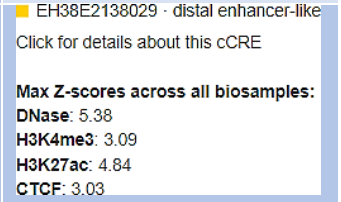
conditions, including the cell state at the time of transfection. However, the demonstration of persistent expression differences between the Alternative and the Reference alleles in these variants and fulfilling the majority of the similarities between experiments makes them worthy of follow-up. Along with this, further research using online databases such as RegulomeDB, the ENCODE project, Screen registry V3, and UCSC revealed that some of the variants that show the same direction of effect in both cell lines fall in regions of particular interest such as promoter like signatures, H3K36me3, proximal or distal enhancer signatures, etc. These data gave us confidence in following certain variants for further evaluation taking into account that they displayed the same direction of effect in the duplicated experiment and fall in promising regions. Interestingly two variants tested do not display to fall in any regions of interest in the context of annotation; however, they are the only variants replicated with a perfect score amongst all others so we decided to retain them for further evaluation. Represented below in table 10 is the annotation information of the significant variants: ***Table 10*** *Annotations of the variants significant from both MPRA experiments as Promoters on the Jurkat cell line.*

Variants significant from both Jurkat experiments as Promoters			
	UCSC	**Regulomedb**	**Screen Registry V3**
14_IFNGR2_RV_RS2865319 8 (-)	No genomic positioning information, only transcription factors data from Jaspar	RegulomeDB chromatin state in the variants' position indicating low transcriptional activity in neurons	No candidate cCREs data at this genomic position
8_PRDX5_RV_RS72924108 (+)	No genomic positioning information, only transcription factors data from Jaspar	RegulomeDB chromatin state in the variants' position indicating low transcriptional activity in neurons	
9_CD40_FW_RS1883832 (-)	Gene hancer (strong promoter), ORegAnno (TF binding sites)	cCREs-promoter like, ATAC-Seq, bivalent enhancer,	

5_PRDX5_RV_RS28364831 (-)	Encode (promoter-like signature), Genehancer promoter-like, ORegAnno TF binding site	cCREs-promoter like, (active TSS transcriptional activity)	EH38E2963009 · promoter-like Click for details about this cCRE **Max Z-scores across all biosamples:** **DNase**: 7.30 **H3K4me3**: 6.10 **H3K27ac**: 6.35 **CTCF**: 2.65
6_CD40_FW_RS6074022 (-)	Encode-(distal enhancer-like signature), Genehancer (enhancer-like activity), ORegAnno TF binding site	cCREs-distal enhancer-like, (weak enhancer)	EH38E3435379 · distal enhancer-like Click for details about this cCRE **Max Z-scores across all biosamples:** **DNase**: 5.07 **H3K4me3**: 2.63 **H3K27ac**: 3.58 **CTCF**: 1.69

Table 11 *Annotations of the variants significant from both MPRA experiments as Enhancers on the Jurkat cell line.*

Variants significant from both Jurkat experiments as Enhancers			
	UCSC	**Regulomedb**	**Screen Registry V3**
12_PRDX5_RV_RS72922077 **(-)**	-Encode cCREs available **distal enhancer like effect** -Falls in layered H3K27ac -Regulatory element from ORegAnno TF	**cCREs-Distal enhancer like** Chromatin state on neurosphere show Enhancer/strong transcription	EH38E1543964 · distal enhancer-like Click for details about this cCRE **Max Z-scores across all biosamples:** **DNase**: 2.97 **H3K4me3**: 2.71 **H3K27ac**: 4.04 **CTCF**: 1.59

6_TEC-TKX_RV_17574371 **(-)**	Encode (distal enhancer), Genehancer (enhancer), ORegAnno TF binding site	cCREs-Distal enhancer like	
7-IFNGR_INS->G_FW_RS17880053 (-)	(proximal enhancer), (strong promoter), Jaspar TF, ORegAnno, TF binding sites	cCREs-distal enhancer, low transcriptional activity	

Allele Specific Transcription Factor Prediction by MotifBreakR.

MotifbrekR is a bioinformatic tool that allows one to judge whether the sequence surrounding a polymorphism or mutation is a good match to known transcription factor binding sites, and how much information is gained or lost in one allele of the polymorphism relative to another mutation vs wildtype. motifbreakR works with position probability matrices (PPM). PPM are derived as the fractional occurrence of nucleotides A, C, G, and T at each position of a position frequency matrix (PFM). PFM are simply the tally of each nucleotide at each position across a set of aligned sequences. With a PPM, one can generate probabilities based on the genome, or more practically, create any number of position-specific scoring matrices (PSSM) based on the principle that the PPM contains information about the likelihood of observing a particular nucleotide at a particular position of a true transcription factor binding site (Coetzee et al 2015). Since it allows us to assess changes in transcription factors binding of sequence variations of interest based on which allele is present we employed the tool on the variants that emerged as functionally involved in gene expression in the Jurkat cell line. We applied filters on the results from motifbreakR selecting only TF with a difference binding between the two alleles (Alternative or Reference) conferring an Alt p-value and Ref p-value below 0.01 and selecting only TF with a "strong effect", a parameter reported by the tool that indicates the strength of the binding of the transcription factor to the sequence. Among the 8 variants passing the MPRA significance filters and that was replicated in the Jurkat experiment performed in double, we obtained a prediction of binding factors from motifbreakR tool for 4 variants, respectively; the

9_CD40_FW_RS_1883832, 5_PRDX5_RS28364831, 6_TEC_TKX_RV_RS72924108 and the 6_CD40_FW_RS6074022. ***(Table 12-13-14-15)***

Table 12 *Results from MotifbreakR on the transcription factors binding to DNA motifs of 9_CD40_FW_RS 1883832*

Promoter Effect					
9_CD40_FW_RS1883832					
Chr.	**SNP_ID**	**REF**	**ALT**	**geneSymbol**	seqMatch
chr20	rs1883832	T	C	**RUNX1**	**ctcacctcgcTatggttcgtc**
chr20	rs1883832	T	C	**MEF2B**	**ggtctcacctcgcTatggttcgtctgc**
chr20	rs1883832	T	C	TYY1	**tctcacctcgcCatggttcgtct**
chr20	rs1883832	T	C	REST	**tgccgcctggtctcacctcgcCatggttcgtctgcctctgcag**
chr20	**rs1883832**	**T**	**C**	**ZN329**	**ccgcctggtctcacctcgcCatggttcgtctgcctctgc**

Table 13 *Results from MotifbreakR on the transcription factors binding to DNA motifs of 5_PRDX5_RS28364831*

Promoter Effect					
5_PRDX5_RV_RS28364831					
Chr.	**SNP_id**	**REF**	**ALT**	**geneSymbol**	seqMatch
chr11	rs28364831	A	C	**E2F4**	**ttgtgggagctgCggtaggtaggtg**
chr11	rs28364831	A	C	**E2F1**	**tttgtgggagctgCggtaggtaggtga**
chr11	rs28364831	A	C	**ZNF436**	**cttctcgtgtttgtgggagctgCggtaggtaggtgaaagacctgtc**
chr11	**rs28364831**	**A**	**C**	**MAF**	**tcgtgtttgtgggagctgAggtaggtaggtgaaagac**

Table 14 *Results from MotifbreakR on the transcription factors binding to DNA motifs of 6_CD40_FW_RS6074022*

Promoter Effect					
6_CD40_FW_RS6074022					
Chr.	**SNP_id**	**REF**	**ALT**	**geneSymbol**	seqMatch
chr20	rs6074022	C	T	**SRBP1**	**agtgtcctcaCgacatggcag**
chr20	rs6074022	C	T	**AHR**	**tgtcctcaCgacatggc**
chr20	rs6074022	C	T	BACH2	**agtgtcctcaTgacatggcag**
chr20	rs6074022	C	T	NFE2	**tgagtgtcctcaTgacatggcagac**
chr20	**rs6074022**	**C**	**T**	**BATF3**	**gtgctgagtgtcctcaTgacatggcagacagct**

Table 15 Results from MotifbreakR on the transcription factors binding to DNA motifs of 6_TEC_TKX_RV_RS72924108

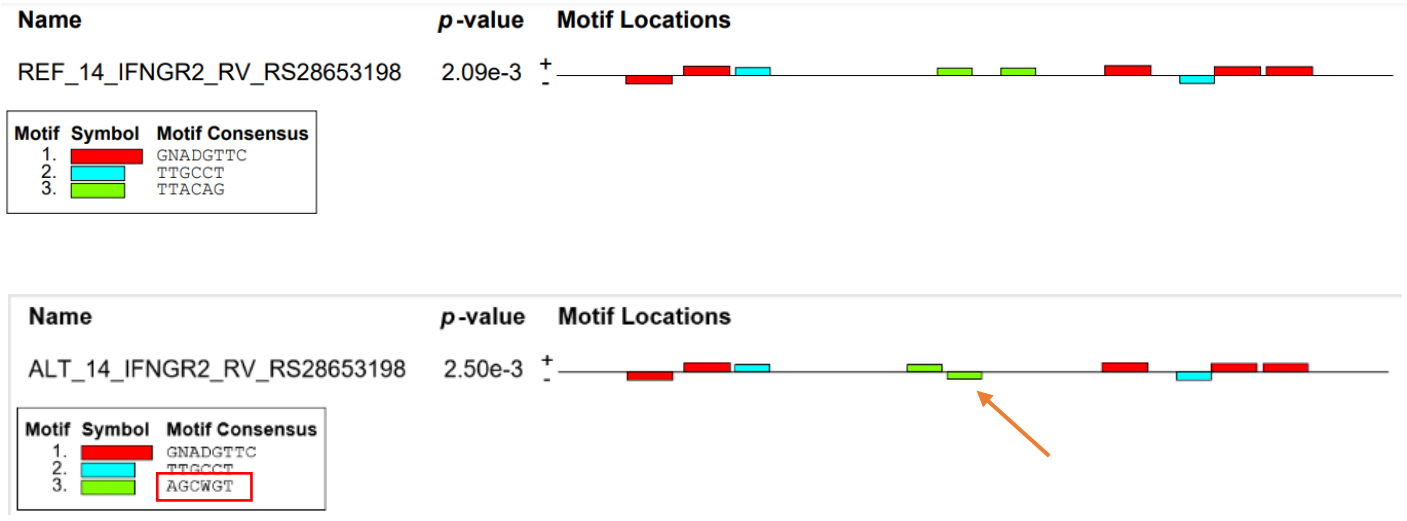
Promoter Effect					
6_TEC_TKX_RV_RS72924108					
Chr.	**SNP_id**	**REF**	**ALT**	**geneSymbol**	seqMatch
Chr4	rs17574371	T	C	**CTCF**	**taacttcgggtgtgtcagCttcctggaaggaagaca**
Chr4	rs17574371	T	C	**TBX3**	**ggtgtgtcagCttcctgggaa**
Chr4	rs17574371	T	C	DUX4	**ggtgtgtcagTttcctgggaa**
Chr4	**rs17574371**	**T**	**C**	**FOXO3**	**gtgtgtcagTttcctggga**

Motif analysis by Meme suite.

For the remaining important variants from the MPRA analysis, MotifbreakR was not able to predict any transcription factors binding to the sequence in the presence of either alternative or reference allele. We decided to utilize the MEME suite on the tested sequence of the SNPs (145 bp of the sequence tested in MPRA) to see if, in the presence of either allele, there would be the creation of any transcription factor binding motif. MEME discovers novel, ungapped motifs (recurring, fixed–length patterns) in a given sequence. MEME splits variable-length patterns into two or more separate motifs. MEME represents motifs as position-dependent letter-probably matrices that describe the probability of each possible letter at each position in the pattern. We applied MEME on the 4 variants that motifbreakR could not yield information, respectively; 14_IFNGR2_RV_RS28653198, 8_PRDX5_RV_RS72924108, 12_PRDX5_RV_RS72922077, AND 7_IFNGR2_INS->G_FW_RS17880053.

We uploaded on the MEME suite portal the sequence of each tested variant, corresponding to the 145 base pair sequence, which was also, tested by MPRA, respectively the alternative and the reference sequence, to see if, in the presence of either allele, we have the formation of a new motif. We selected any number of repetition options for the motif sites distribution in the sequence for each reference and alternative sequence and ran the tool in classic mode, represented below the results for each tested sequence.

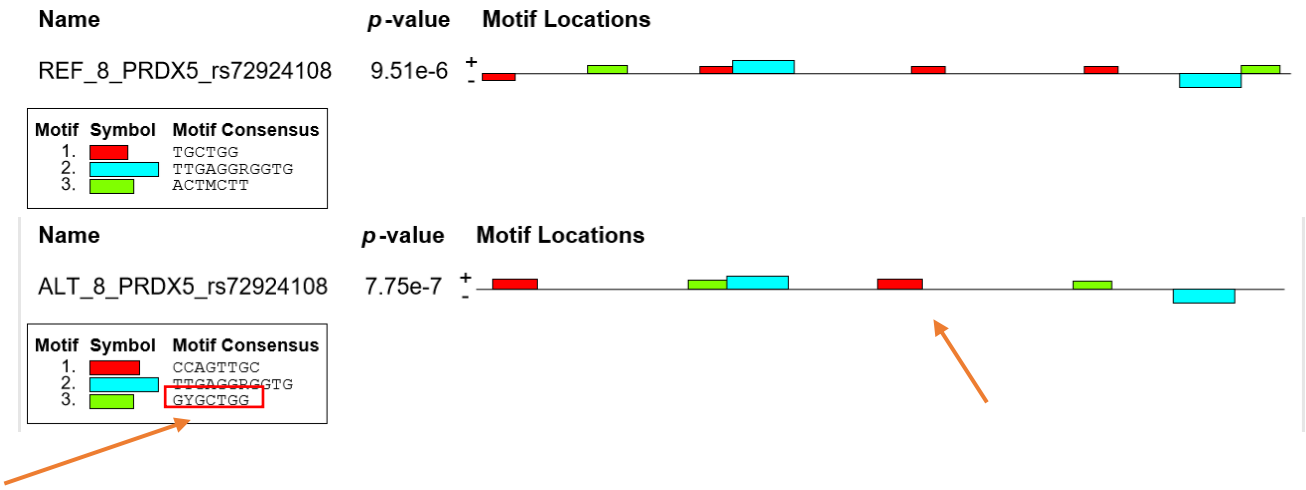
Fig 19. *MEME suite results for 14_IFNGR2_RV_RS28653198 variant*



We notice that in the presence of the alternative **C**, there is another motif, which was not present in the reference sequence, indicating the possibility that this motif could harbor a new consensus motif for other transcription factors that alter gene expression. Tomtom is a tool that compares one or more motifs against a database of known motifs. Tomtom ranks these motifs in the database and produces an alignment for each significant match based on the JASPAR CORE. We used Tomtom to identify possible transcription factors that bind to newly formed motif. Through the JASPAR2022_CORE, we were able to identify different transcription factors that bind to the motif in the presence of the alternative allele. Represented below the predicted transcription factors that bind in the newly formed motif.

Preview	Matches	List
	29	MA0665.1 (MSC), MSC_full, MA1472.2 (Bhlha15), MA1619.1 (Ptf1A), MA1641.1 (MYF5), MA0832.1 (Tcf21), MA1485.1 (FERD3L), Tcf21_DBD, MA0521.2 (Tcf12), MA0796.1 (TGIF1), TGIF1_DBD, UP00258_1 (Tgif2_3451.1), MA1967.1 (TFAP4::FLI1), MA0782.2 (PKNOX1), MA0667.1 (MYF6), MA0691.1 (TFAP4), MYF6_full, TFAP4_full, UP00122_1 (Tgif1_2342.2), TFAP4_DBD, MA1571.1 (TGIF2LX),

Fig 20. *MEME suite results for 8_PRDX5_RV_RS72924108 variant.*



When we input the alternative and reference sequence into the Meme suite database in the case of the tested 145 bp sequence of one of the 24 tested variants of the PRDX5 region, which was predicted as significate from the MPRA analysis we obtain some valid data. In the presence of the alternative **C** allele, there is a formation of a new motif at the position of the SNP, which by applying the TomTom tool reveals different transcription factors shown below.

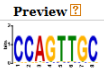
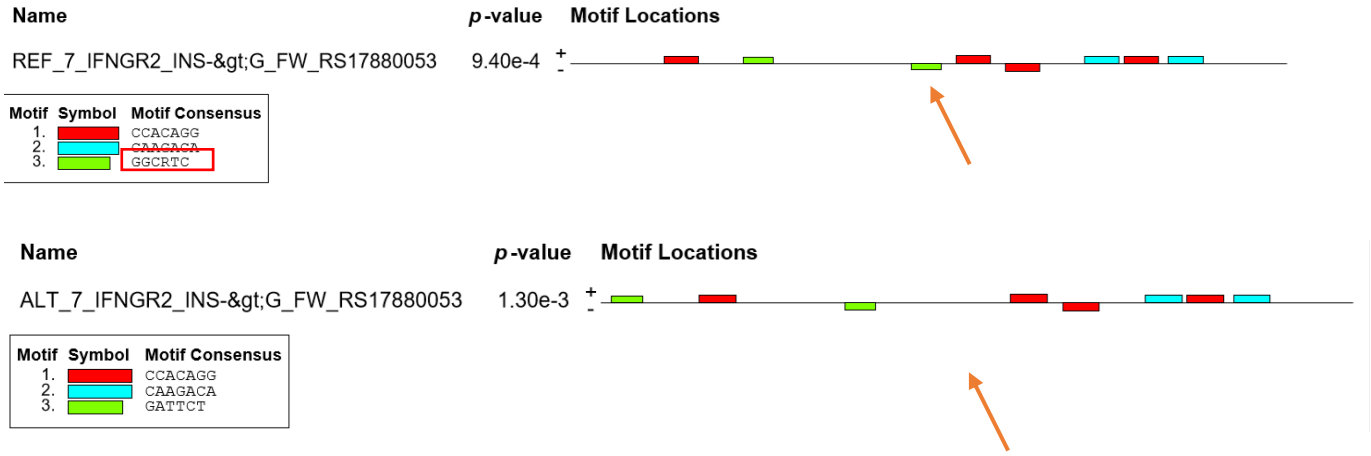
Database	ID	Alt. ID	Preview	Matches	List
query_motifs	1	CCAGTTGC		8	SCRT2_DBD, SCRT1_DBD, MA0744.2 (SCRT2), MA1635.1 (BHLHE22), TCF4_DBD, MA1997.1 (Olig2), MA0633.2 (Twist2), MA1638.1 (HAND2)

Fig 21 *MEME suite results for the 7_IFNGR2_INS->G_FW_RS17880053 variant.*



As regarding to the 7_IFNGR2_INST->G_FW we notice that in the presence of the insertion **G** in the alternative sequence we lose the motif that was present in the sequence containing the reference allele (without the insertion). These results will be further evaluated in the upcoming section. Represented below is the list of the transcription factors predicted by TomTom that bind to the motif present in the reference sequence.

Database	ID	Alt. ID	Preview	Matches	List
query_motifs	1	GGCRTC		2	MA0794.1 (PROX1), PROX1_DBD

Meme suite was not able to show any differences in the motifs present for the 12_PRDX5_RV_RS72922077, meaning that in the presence of either alternative or reference allele, there are no motifs that could bind in the sequence. However, by this procedure, we were able to obtain substantial information regarding the other four MPRA significate sequences which motifbreakR did not provide information. Either most of them lose a motif or demonstrate the creation of a new motif consensus site in the presence of either allele. However, these preliminary results need further evaluation. In our situation, we already have information on how each allele influences expression from MPRA, now we also have a possible pathway by which it could exert this function. However, in the list of SNPs that we have selected to test by MPRA, we need to associate the effect seen by MPRA, with the risk allele associated with MS.

Risk allele association to transcription factors and data interpretation.

Among the 8 variants that passed the MPRA filters and were replicated within two distinct experiments in the Jurkat cell line, we obtained the prediction of transcription factor binding from motifbreakR for 4 variants. For the remaining 4 variants we used the MEME suite website to observe the creation of new consensus motifs for transcription factor binding, which gave us promising results except for one variant the 12_PRDX5_RV_RS72924108. Represented below is a table that shows the specific information for each variant as it regards the direction of effect from MPRA for the tested variants, the risk allele, and the prediction of transcription factors from both motifbreakR and MEME suite. The risk allele is the allele, which is highly associated with the disease. It is important to refer to the risk allele when discussing the direction of effect since this allele probably contributes to disease pathogenesis or progression. (Table 16)

Variant with data from MotifbreakR	Promoter/Enhancer Effect tested	Risk allele	Reference or Alternative	MPRA effect
9_CD40_FW_RS1883832	Promoter	T	Reference	-
5_PRDX5_RV_RS28364831	Promoter	C	Alternative	-
6_CD40_FW_RS6074022	Promoter	C	Reference	-
6_TEC-TKX_RV_RS17574371	Enhancer	C	Alternative	-
Variant with data from MEME suite				
14_IFNGR2_RV_RS28653198	Promoter	C	Alternative	-
8_PRDX5_RV_RS72924108	Promoter	C	Alternative	+
7_IFNGR2_INS->G_FW_RS17880053	Enhancer	G insertion	Alternative	-
Variant with no changes in motif				
12_PRDX5_RV_RS72924108	Enhancer	A	Alternative	-

Table 16 *representing the information regarding TF for each variant.*

This table shows all the variants for which we were able to obtain information about the risk allele of the variant associated with Multiple Sclerosis. To better understand this table we need to look at the results in a chronological way that better explains the intricate methodology in evaluating each variant. Starting from the variants that emerge as significant from two distinct in vitro experiments, we apply a series of filters that leave us with a few probes that demonstrate a prominent effect among replicates. This effect is represented in the form of a negative or positive sign (-/+). The sign describes the ratio between RNA and DNA counts (as explained in the methodology) between the alternative and reference allele. A negative sign (-) shows that the variant when tested by MPRA as either promoter or enhancer indicates that the alternative alleles are associated with reduced expression levels when compared to the expression of the reference allele which is higher(each represented by

its reads post-NGS). However, in some of the variants not always the alternative allele is the risk allele. This creates the need to address the effect of the variant on the respective allele associated with the disease. As an example in the table above the 9_CD40_FW_RS1883832 variant shows a prominent effect in two distinct experiments as a promoter. The direction of effect is a negative sign (-) showing a reduced expression in the presence of the alternative allele as we in the MPRA tables primarily refer to that allele. In this case, the risk allele is the reference allele, but since the direction of the effect is an output of a Log Ratio between the alternative and the reference allele TAG count (logFC), we can say that we can obtain information indirectly for the reference allele, which is associated with an increased expression in this case. Similarly, we can use the same reasoning for further evaluations associated with the other variants. The 6_CD40_FW_RS6074022 has a similar situation, wherein in the presence of the alternative allele the expression is reduced and in the presence of the reference allele we have an increased expression. In the other cases, for those in which we were able to obtain the risk allele, we noticed that the risk allele is the alternative allele associated with a negative sign, indicating a reduced gene expression in the presence of the alternative allele when compared to the reference allele. In one case the alternative allele is associated with a positive sign indicating an increased expression of the reporter gene in the presence of the alternative allele. Once we have connected the MPRA direction of effect with the MS-associated risk allele, we can continue to further analyze specific pathways why in the presence of the risk allele there is a change in expression.

Represented below (Table 17) shows transcription factors predicted by either motifbreakR or MEME suite that bind in the presence of the risk allele.

Variant with data from MotifbreakR	Promoter/ Enhancer Effect tested	Risk allele	Reference or Alternative allele	MPRA effect based on the risk allele	Transcription factors that bind to the risk allele
9_CD40_FW_ RS1883832	Promoter	T	Reference	Positive	**RUNX1, ZN768**
5_PRDX5_RV_ RS28364831	Promoter	C	Alternative	Negative	**E2F4, E2F1, ZNF436, MAF**
6_CD40_FW_ RS6074022	Promoter	C	Reference	Positive	**SRBP1, AHR**
6_TEC-TKX_RV_ RS17574371	Enhancer	C	Alternative	Negative	**CTCF, TBX3**

Variant with data from MEME suite	Promoter/ Enhancer Effect tested	Risk allele	Reference or Alternative	MPRA effect based on the risk allele	Transcription factors that bind to the risk allele
8_PRDX5_RV_ RS72924108	Promoter	C	Alternative	Positive	**SCRT2, SCRT1, BHLHE22, TCF4, HAND2, TWIST2**
14_IFNGR2_RV_ RS28653198	Promoter	C	Alternative	Negative	**MSC, TGIF2, TFAP4, MYF5, PKNOX1, FERD3L**
7_IFNGR2_INS->G_FW_RS17880053	Enhancer	G Insertion	Alternative	Negative	**Loss of PROX1**

Table 17 *describing the variants selected to have the highest probability of being causative, along with their risk allele and the risk allele effect, associated with transcription factors that bind in the presence of the risk allele.*

9_CD40_FW_RS1883832- Transcription Factor Description

Transcription factors binding in the presence of the reference T allele include RUNX1 and MEF2B. RUNX1 (Runt-Related Transcription Factor 1), is a Core binding factor (CBF), a heterodimeric transcription factor that binds to the core element of many enhancers and promoters. The protein encoded by this gene represents the alpha subunit of CBF and is thought to be involved in the development of normal hematopoiesis. Chromosomal translocations involving this gene are well documented and have been associated with several types of leukemia. **MEF2B** (Myocyte Enhancer Factor 2B) is a Protein Coding gene associated with diseases such as Leukemia, Chronic Lymphocytic and Thrombocythemia 1. Among its related pathways are the Immune response Function of MEF2 in T lymphocytes and the fMLP Pathway. Gene Ontology (GO) annotations related to this gene include *DNA-binding transcription factor activity* and *protein dimerization activity.* When choosing which of the candidate transcription factors is more probable to be exerting the increased expression as predicted by MPRA between RUNX1 and MEF2B we decided to look for clues in open databases such as ENCODE and Regulomedb. ENCODE and Regulomedb contain data from already performed TF-Chip seq analysis. Both Regulomedb and ENCODE (performed on GM12878 Lymphoblastoid cells) demonstrate that there is enrichment of MEF2B in the genomic location of the SNP, giving us confidence in presuming the effect of the SNP is due to the new binding of MEF2B which act as a transcriptional activator.

5_PRDX5_RV_RS28364831- Transcription Factor Description

Transcription factors binding in the presence of the alternative **C** allele include two E2F family transcription factors among others. The E2F family plays a crucial role in the control of cell cycle and action of tumor suppressor proteins and is also a target of the transforming proteins of small DNA tumor viruses. E2F1 primarily functions as a transcriptional activator during the G1/S transition and binds to promoters of genes involved in DNA replication and cell cycle progression. Conversely, **E2F4** typically acts as a **repressor** in quiescent and differentiated cells by recruiting repressive complexes, including pocket proteins like p107 and p130. Studies conducted by Lee et al., 2011 confirm E2F4 function as a transcriptional repressor by conducting overexpression experiments and Chip sequencing. Also by using the Regulomedb and Encode databases, we were able to confirm by performed TF- Chip seq, enrichment of this transcription factor in the region of the SNP. Altogether, this information gives confidence to this candidate transcription factor in its role of modifying gene expression.

6_CD40_FW_RS6074022- Transcription Factor Description

Transcription factors binding in the presence of the alternative C allele include SRBP1 and AHR among others. AHR specifically seems of particular interest since the protein coded by this gene is a ligand-activated helix-loop-helix transcription factor involved in the regulation of biological response to planar aromatic hydrocarbons. Therefore, it enables cells to adapt to changing conditions by sensing compounds from the environment, diet, microbiome, and cellular metabolism, which plays an important role in development and immunity. Upon ligand binding, it translocates into the nucleus, where it heterodimerizes with ARNT and induces transcription by binding to xenobiotic response elements (XRE). It has a prominent effect on cell motility and immune modulation. These elements make AHR a very interesting candidate that can explain the increased expression observed in the presence of the alternative allele as predicted by MPRA. However, for this particular transcription factor, there were no data from Regulomedb, only from ENCODE demonstrating enrichment of the TF in the region where the SNP falls.

6_TEC-TKX_RV_RS17574371- Transcription Factor Description

Transcription factors binding in the presence of the alternative **C** allele include CTCF, TBX3, etc. Among these transcription factors, the most promising one seems to be TBX3 (T-Box Transcription factor 3). T-box genes encode transcription factors involved in the regulation of developmental processes. Diseases associated with TBX3 include Holt-Oram Syndrome and among its related pathways is the Nervous system development. It binds to the palindromic T site 5'-TTCACACCTAGGTGTGAA-3'. Its effect as a transcription repressor makes it a valid candidate that could explain the effect predicted by MPRA when in the presence of the alternative allele.

8_PRDX5_RV_RS72924108- Transcription Factor Description

Transcription factors binding in the presence of the alternative **C** allele include SCRT2, TCF4, and HAND2. Although these transcription factors show promising initiatives as they all are involved in neuronal development and autoimmune response. The most promising one seems to be TCF4. **Transcription Factor 4** encoded protein contains a motif first identified in immunoglobulin enhancers. This gene is broadly expressed and may play an important role in nervous system development. Among its related pathways, are ERK signaling and signaling by TGFBR3. It has a function as a transcription activator, making it a valid candidate for further follow-up analysis since it explains the direction of the effect predicted by MPRA. It is involved in the initiation of neuronal differentiation and it activates transcription by binding to the E box (5'-CANNTG-3'). Data from Regulomedb and Encode was not present to confirm from other TF-Chip seq analyses the enrichment of these regions by this specific transcription factor. However, this does not remove all credibility from our findings, as the TF factors predicted by the insilico method were also selected by applying a series of stringent filters.

14_IFNGR2_RV_RS28653198- Transcription Factor Description

Transcription factors binding in the presence of the alternative **C** allele include MSC, TGIF2, TFAP4, MYF5, PKNOX1, and FERD3L. Amongst these transcription factors, TGIF2 (TGF2 Induced Factor Homeobox 2) seems to be promising when considering that the protein encoded by this gene is a DNA-binding homeobox protein and a transcriptional repressor, which represses transcription by recruiting histone deacetylases to TGF beta-responsive genes. Diseases associated with TGIF2 include Holoprosencephaly and Congenital Nervous System Abnormalities which are related to a series of diseases among them also Autoimmune Disease of the Central Nervous System. TGFI2 is a transcriptional repressor, which represses transcription via the recruitment of histone deacetylase proteins, which could explain the direction of effect as predicted by MPRA in the presence of the alternative allele.

7_IFNGR2_INS->G_FW_RS17880053-Transcription Factor Description

In the case of the **G** insertion in the sequence of rs17880053 when the insertion is present, there is the loss of a consensus motif. The transcription factor that is used to bind in the absence of the insertion is PROX1. PROX1 is a member of the homeobox transcription factor family. The protein encoded by this gene is conserved across vertebrates and plays an essential role during development. As a transcription factor is involved in gene transcriptional regulation and progenitor cell regulation in a number of organs. It's a key regulatory protein in neurogenesis and it's a transcriptional activator of core clock components and metabolic genes. The absence of binding of PROX1 could explain the effect of the insertion G allele in reducing expression.

8.6 Exploring Drug Repurposing for target Genes

To predict the target gene of the carefully filtered functional variants associated with MS, we decided to employ the Open Target Genetics tool. Open Targets Genetics is a comprehensive tool, that highlights variant-centric statistical evidence to allow both the prioritization of candidate causal variants at trait-associated loci and identification of potential drug targets. It aggregates and integrates genetic associations curated from both literature and newly derived loci from UK biobank and FinnGen along with functional genomics data (chromatin conformation, chromatin interaction) and quantitative trait loci (eQTLs, pQTLs, and QTLs). Large-scale pipelines apply statistical fine-mapping across thousands of trait-associated loci to resolve association signals and link each variant to its proximal and distal target gene(s) using Locus2Gene assessment. Integrated cross-trait colocalisation analysis allows the linking of detailed pharmaceutical compounds making Open Target Genetics able to explore drug-repositioning opportunities. Starting from 8 variants that showed a consistent imbalanced effect in the presence of either Reference or Alternative allele on the expression of the reporter gene in two distinct experiments we were able to predict a possible TF binding in the sequence for at least seven of them. However, even though we were not able to predict TF for the 12_PRDX5_RV_RS72922077 variant, it doesn't mean that it's not eligible for follow-up. This variant has passed a very stringent line of filters indicating its unquestionable involvement as a functional SNP. It probably exerts its function in another pathway not influenced by transcription binding factors such as epigenetic modifications, chromatin structure, or 3D genome organization. However, for the 8 variants that emerge as functional SNPs, we were able to identify their target gene and a possible drug that has an antagonist effect to that of the effect of the SNP.

Represented below for each variant their target gene and the effect they have on the target gene as predicted by MPRA, and a drug that has an apposed effect to that predicted by MPRA as the effect predicted is related to the risk allele thus attributing to the disease.

9_CD40_FW_RS1883832- Target Gene and Drug

As predicted by Open Target Genetics this SNP affects the CD40 gene, as confirmed by various eQTL and pQTL studies present in the database of the tool. The effect predicted on the gene by Open Target is referred to as the Alternative allele (as commonly done in these databases), but in our case, we refer to the effect based on the reference allele so we will take into account the effect predicted by MPRA and not the effect predicted by the tool. MPRA shows that in the presence of the reference allele, there is an increased expression as measured by the reporter assay. It is important to mention that the effect of many SNPs tested in a reporter system depends on the tissue type and cell line, as different cells express higher or lower levels of many genes, which in turn modify the signaling network of a cell, ultimately showing an inconsistent direction of effect for a given SNP when tested in diverse conditions/cells. While in some papers, it's demonstrated an opposite effect to that predicted by us regarding the rs1883832, to our knowledge we are the first to apply this experimental flow in the Jurkat cell line. We confirm an effect in the presence of the reference allele (Risk Allele) which is associated with an increased expression as measured in an invitro reporter system. By Open Target Genetics we know that the gene with the highest probability to be a target of rs1883832 is the CD40, a gene highly involved in the immune system and inflammatory responses including T cell-dependent immunoglobulin class switching, memory B cell involvement, etc. Again using the Open Target Genetics we were able to predict a drug that has an antagonist effect to that of the SNP; presumably, if the SNP increases the expression of the CD40 gene, we can attribute this as something that promotes disease pathogenesis. Therefore, we focused on drugs that reduce the expression of the CD40 gene, such as Bleselumab, and Iscalimab. There is a long range of drugs that target and inhibit the functions of CD40; however, these two drugs have completed phase II clinical trials making them even more promising. Iscalimab (CFZ533) is a monoclonal antibody its mechanism of action is to inhibit the tumor necrosis factor receptor superfamily member 5. It has been explored across multiple studies for its role as an inhibitor of CD40 in many autoimmune conditions and other contexts. Some of the most notable uses and discoveries regarding Iscalimab are in the efficacy of Sjogren's Syndrome, which demonstrated that it could significantly reduce disease activity over 24 and 48 weeks. It has also shown an immunosuppression profile in rhesus monkeys by demonstrating selective suppression

of B-cell mediated immunity without affecting T-cell function and avoiding thrombolic complications commonly associated with similar therapies. (Flandre et al., 2022)

6_CD40_FW_RS6074022- Target Gene and Drug

Prediction done by Open Target Genetics shows that rs6074022 affects the CD40 gene, similarly to another SNP in the same region that we discussed above rs1883832. The effect predicted by Open Target Genetics refers to the alternative allele of this SNP, but as in the case of rs1883832, the risk allele is the reference. MPRA shows that in the presence of the reference allele, there is an increased expression as measured by the reporter assay. Similarly, to the rs1883832, this SNP has also been mentioned in numerous studies conducted on GWAS data studying complex diseases, which are caused by an interplay between genetic and environmental factors. The effect of this SNP in the presence of the risk allele trends towards an increased luciferase activity, implying an increased expression of CD40. Therefore, we will focus on drugs that reduce the expression of the CD40 gene, such as Iscalimab, and Bleselumab.

5_PRDX5_RV_RS28364831- Target Gene and Drug

As predicted by the Open Target Genetics database the target gene of this SNP is the VEGF-B gene. The effect of the rs28364831 variant on the gene as predicted by Open Target Genetics based on the alternative allele shows a negative Beta value of -0.102 with a p value of 1.5e-33 meaning that this variant has a negative effect on the expression of the VEGF-B gene. This is also confirmed functionally by our experimental evaluation where we see that in the presence of the alternative allele, there are reduced luciferase expression levels. VEGF-B is a member of the PDGF/VEGF family, which regulates the formation of blood vessels and is involved in endothelial cell physiology. VEGF-B plays an important role in several types of neurons showing a protective role of neurons in the retina and the cerebral cortex and of motor neurons such as amyotrophic lateral sclerosis. It is essential for normal vascular development and homeostasis and is also implicated in neurodegeneration and it may play a protective role in Alzheimer's disease (AD) since it is reduced in AD as demonstrated by patient's serum *in vivo* (Storkebaum et al., 2004). Our data suggest that there is a reduction of VEGF-B expression in the presence of rs283664831, indicating a possible favorable role in MS pathogenesis. However, since VEGF-B is mostly studied in the cancer context, being almost always upregulated, most drugs shown by Open Target Genetics serve as inhibitors of the VEGF-B gene. This led us to

follow two routes in identifying drugs with possible agonist roles to that of VEGF-B. We weighted to target downstream signaling pathways of the VEGF-B gene, which when bound to its receptor and co-receptors such as NRP-1, activates the PI3K pathway, which recruits the AKT cell membrane. AKT is then activated through phosphorylation by PDK1 and mTORC. Among its receptors and these downstream signaling pathways, only mTORC had a drug that had an activating role, which could diminish the presumed negative effect of the VEGF-B under expression. The monoclonal antibody is called INDOXIMOD with an effect as an mTORC1 activator, used in metastatic prostate cancer. However, overexpressing such an important serine/threonine kinase such as mTORC could have its repercussions, and VEGF-B probably exerts its negative impact through more intricate and fine pathways. For this we decided to focus on the second approach; proposing the use of recombinant human VEGF165 (rhVEGF165) which would enhance angiogenesis and improve blood-brain barrier with better functional neurological outcome (Zhang et al., 2000).

12_PRDX5_RV_RS72922077- Target Gene and Drug

The predicted target gene of this SNP from the Open Target Genetics database similarly to the other SNP in the same region is the VEGFB gene. The effect predicted by both Open Target Genetics and MPRA is that this SNP has a negative effect reducing expression when the alternative allele is present which is also the risk allele. Similarly, to the other SNP in this region that affects the VEGFB gene the drug that targets the gene with an agonist effect would be the use of recombinant human VEGF165.

8_PRDX5_RV_RS72924108- Target gene and Drug

The predicted target gene of this SNP, similarly to the other SNPs in the same region is the VEGF-B gene. The effect of this SNP on the target gene as anticipated by the Open Target Genetics database shows that it should reduce the expression of the VEGF-B gene. However, the effect predicted by MPRA in the presence of the alternative allele has a positive sign, indicating an elevated expression of the reporter gene in the functional assay when the risk allele is present. This is controversial to the other two SNPs in the same region; however, this contrast better highlights the difficulties in pinpointing causative variants and their effect on target genes, emphasizing the need for follow-up studies. Nevertheless, this variant supposedly increases the expression of the VEGFB gene, implying that this effect contributes to disease progression, the drug's required outcome should be an inhibitor of the VEGFB gene. CONBERCEPT is a protein with a vascular endothelial growth factor B inhibitor effect, which has completed the clinical trial phase IV.

14_IFNGR2_RV_RS28653198- Target gene and Drug

The predicted target gene of this variant is the IFNGR2 gene (Interferon Gamma Receptor 2) based on the V2G score. The Open Target Genetics database shows that the eQTL effect of this variant in the presence of the alternative allele on the gene should increase its expression. However, as mentioned before we are the only ones to have conducted this kind of study and experimental pipeline using this sort of cell line, therefore we will follow the prediction based on our in vitro functional analysis, which evidently should represent a closer biological view of the effect of SNP in gene expression. Besides, the nature of the disease we are researching is characterized by an utterly large number of features and complex infrastructure given its autoimmune nature, leaving space for innovative findings and novel ways to look at the attribution of diverse pathways to its manifestation. That being said, the effect of the SNP on the IFNGR2 gene in the presence of the risk alternative allele is negative, meaning it reduces its expression as measured by the reporter gene. Estimating that this outcome contributes to disease pathogenesis, a drug that has an agonistic effect on the IFNGR2 gene is INTERFERON GAMMA-1B. INTERFERON GAMMA-1B is a protein with a mechanism of action as an interferon-gamma receptor agonist with a phase III completed status in clinical trials.

7_IFNGR2_INS->G_FW_RS17880053- Target gene and Drug

Likewise to the other SNP in the same region tested by MPRA, the target gene of rs17880053 is the IFNGR2 gene. Additionally, the effect of the SNP as predicted by Open Target Genetics should increase its expression as predicted by the beta value on the IFNGR2 gene, but we will focus on the effect remarked when in the presence of the alternative allele as measured by the reporter gene in our functional assay. The effect of this SNP corresponds to the other SNP in the same region, predicting a reduced expression of their target gene; *IFNGR2*. A drug with an agonist role to the IFNGR2 gene is INTERFERON GAMMA-1B.

6_TEC-TKX_RV_RS17574371- Target gene and Drug

The target gene predicted by Open Target Genetics for this variant is the TEC gene (TEC protein tyrosine kinase) of the TEC family which also includes Btk, Itk, Rlk, and Bmx. The effect predicted by both the tool and MPRA suggests a negative effect of the SNP on the TEC gene signifying a reduced expression of the gene when the risk allele is present. This might seem argumentative and polemic since recently there have been Btk inhibitors in clinical trials as possible drugs for multiple sclerosis. Nevertheless, there have been studies indicating that the loss of TEC, leads to increased B cell activation through enhanced activation of the Akt pathway. Also, TEC protein tyrosine kinase is described as a negative regulator IL2RA (CD25) expression induced by TCR-cross-linking. These results will be discussed more in detail in the discussion section. Most drugs available for TEC are inhibitors given recent favorable results in MS by using BTK inhibitors. Since Akt is activated in TEC deficient mice leading to increased activation of mature B cells, promoting inflammation and immune response, a hypothetical drug to be used as an inhibitor of the Akt pathway is Capivasertib (AZD5363). Capivarsertib is a serine/threonine kinase inhibitor used to treat hormone receptor-positive, HER2-negative breast cancer, which could potentially be repurposed in autoimmune treatment in appropriate dosage and after clinical trial assessment.

Using MPRA as a tool to study the interaction between environmental factors and genes associated with complex diseases

Our hypothesis to test the effect of the Epstein-Barr EBNA2 variant 1.2 using the MPRA technique followed the same experimental pipeline as described before in the case of the MPRA experiment dissecting functional variants associated with MS in complex regions. For this, we performed a sequential transfection on the Jurkat 2 cell line, firstly with a construct that contains the EBNA2 1.2 variant, and after 24 hours with the MPRA construct to observe the effect in the expression of the sequences as measured by the reporter gene. The outcome of each SNP is measured by comparing tag counts for each probe after transcription, using the MPRAlm tool to compare RNA reads to input DNA. The resulting counts are applied to quantify the activity of a given putative regulatory sequence. Then to explain why the effect of the SNP on the reporter gene had changed we applied MotifBreakR, which predicted a series of Transcription Factors (TF) that were distorted in the presence of viral agents. To assess the target gene of these SNPs we applied the Open Target Genetics tool, which can predict the target gene of the tested SNPs by showing the eQTL effect.

The mpraLM tool was utilized after the final NGS to perform a ratio between RNA reads to DNA input similarly to as described above. It generated three distinct tables by comparing the difference in expression of each Reference and Alternative sequence, also respectively the Reference and the Alternative sequence to the Scramble sequence, which later served as a null hypothesis in the filtering steps. This comparison allowed us to attain a LogFC (Log-fold-change) and an Adjusted P value for each variant tested. We applied a significance filter of FDR<0.01 and focused solely on probes that demonstrate a significant deviation from the null hypothesis (Scramble sequence); we identified the SNPs that exert a significantly different impact on expression between the two allelic variants (Alternative Vs Reference sequence). Variants that were discarded from the replicability experiment on the Jurkat cell line mentioned before were not taken into account in this experiment either since they are considered too versatile to be kept in the account, namely: 17_ PRDX5_RV (rs4930152), 4_CD40_FW (rs2024568), 19_TEC-TKX_FW (rs17471024), 22_IFNGR2_INST_T_FW (rs17471024) and 14_PRDX5_RV (rs3815362). When we applied our experimental pipeline we performed a sequential transfection and each experiment was performed in quadruplicates to enhance statistical significance. Firstly we transfected the Jurkat cells with the pCDNA 1.2 plasmid which contains around 1602 bp of the EBNA2 variant coupled with the GFP sequence using the Neon Transfection System (Thermo Fisher Scientific). In parallel as a positive control we transfected the same number of Jurkat cells with a basic GFP construct which is innocuous on the cells. After 24 hours, we transfected all the replicates with the MPRA library construct Donor 1, which measures the effect of the SNPs as promoters based on the expression of the Luc2 gene.

After 48 hours we extracted the RNA from the cells using the miRNeasy tissue/cells Advanced mini-Kit (Qiagen) and we performed DNase treatment to remove DNA residues. Reverse transcription was carried out using SuperScript™ II (Thermo Fisher) Reverse Transcriptase to synthesize cDNA. Before NGS, a PCR reaction with a High Fidelity Herculase was conducted to amplify the region of interest and add sequencing adapters. We performed NGS on the cDNA obtained from transfected cells and input DNA using Illumina DNA Prep and Nextera XT Index Kits. The number of tags represented for each probe in both cDNA and input DNA indicated the effect of each variant on LUC2 expression. Each step was performed on the cells that were exposed to the pathogen and to those that served as a positive control.

NGS generated the FASTQ files containing the expression counts for the TAG region in each experiment; and after several in silico analyses, we determined the frequency of occurrence for each TAG. Next by using mpralm, we obtained the Log Fc and P values corresponding to each tested SNP. Our hypothesis is based on this crucial step where we want to compare the Log FC of each variant tested individually whether it has significant changes in the direction of effect. Therefore, we measure the effect on the expression of the reporter gene in the presence of the EBNA2 variant 1.2 for each reference or alternative allele and compared it to the effect on the expression of the reporter gene in the presence of an innocuous GFP plasmid as a substitution for EBNA variant 1.2. To have statistical confidence in our results we compared the LogFC of each tested probe with a simple linear regression test between the experiment with the exposure to the pathogen and the control. We selected a 95% confidence interval, which revealed 8 probes with significant changes in the effect measured by the reporter gene when exposed to EBNA variant 1.2. **(Fig.22)**

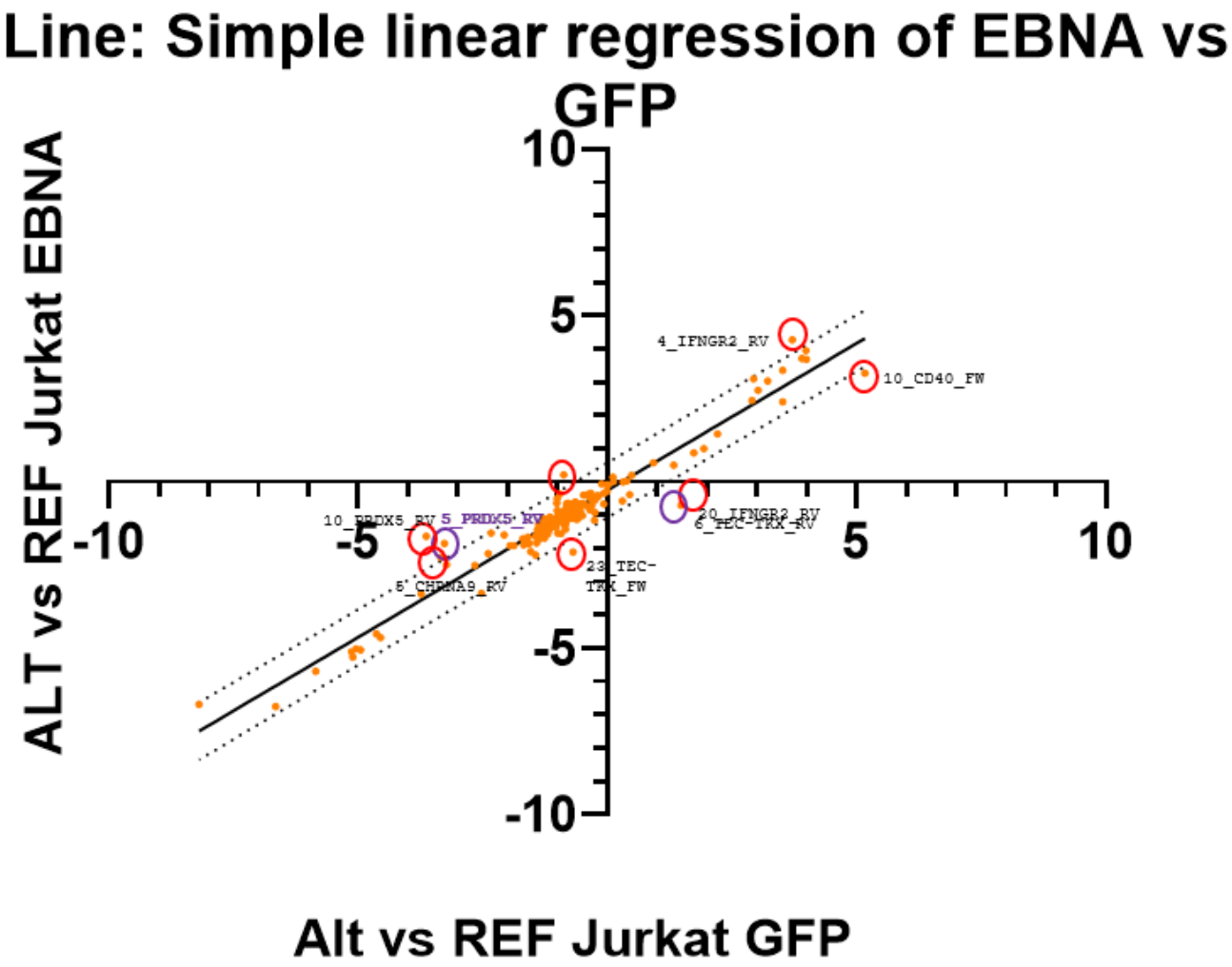


Fig.22 *Simple linear regression analysis of the Log FC of each probe tested with the MPRA experiment on the Jurkat cell line with and without EBNA2 variant 1.2.*

From our analysis, we obtained 8 probes with potential differential expression outside of the confidence interval when exposed to the EBNA2 gene variant 1.2 demonstrating a credible effect in the presence of the pathogen. Out of the 8 probes, only a few fulfilled the predetermined MPRA filters, which are crucial since they facilitate the finding of sequences tested that have a prominent and robust effect linked to a significant p-value. Represented below are the exact values of the 8 probes for each experiment. **(Table 18)**

Values of the probes outside the confidence interval from the EBNA variant 1.2 experiment								
Chr. Pos	Rs ID	Variant name	ALtvsRef	P-Val	RefvsScr	P-val	AltvsScr	P-val
chr11:64277888	rs56057146	10_PRDX5_RV	-1,63	0,00013	-0,30154	0,1165	-1,94	4,06-09
chr11:64317576	rs28364831	5_PRDX5_RV	-1,85	0,0024	5,70	2,58-12	3,82	7,27-06
chr4:48156152	rs10021919	23_TEC-TKX_FW	-2,11	3,79-06	-0,79	2,91-06	-2,91	7,60-08
chr4:48145192	rs17574371	6_TEC-TKX_RV	0,208	0,01	0,73	4,79-06	0,942	1,98-08
chr21:33406121	rs17879003	20_IFNGR_RV	0,70	0,0019	-0,685	0,000118	-1,3966	1,80-05
chr21:33403894	rs9974603	4_IFNGR_RV	4,26	3,93-11	-7,49	5,10-12	-3,231	5,68-08
chr20:46119208	rs4810485	10_CD40_FW	3,25	8,69-09	-3,32	1,97-11	-0,069	0,548
chr4:40306251	rs7691190	5_CHRNA9_RV	-2,27	1,08-07	1,31	1,40-06	-0,967	1,01-05

Table 18 *The values of each variant in the experiment with the presence of EBNA variant 1.2 indicating the two variants that surpass the MPRA filters*

Values of the probes outside the confidence interval from the GFP experiment								
Chr. Pos	Rs ID	Variant name	ALtvsRef	P-Val	RefvsScr	P-val	AltvsScr	P-val
chr11:64277888	rs56057146	10_PRDX5_RV	-3,633	3,7E-08	1,9504	1,1042E-07	-1,6864	1,4531E-07
chr11:64317576	rs28364831	5_PRDX5_RV	-3,271	8,219E-07	6,8721619	2,783E-14	3,5876	2,8426E-08
chr4:48156152	rs10021919	23_TEC-TKX_FW	-0,6884	0,00570	-0,6742	0,00021	-1,3631	0,0000366
chr4:48145192	rs17574371	6_TEC-TKX_RV	-0,87609	0,00093686	1,95869	2,5959E-07	1,0806	3,8198E-06
chr21:33406121	rs17879003	20_IFNGR_RV	1,4778	0,00881394	-0,343016	0,03852620	1,13347	0,039095
chr21:33403894	rs9974603	4_IFNGR_RV	3,7070	0,0000412	-7,36497	3,2687E-10	-3,661	1,451E-07
chr20:46119208	rs4810485	10_CD40_FW	5,1669	2,316E-10	-3,383656	3,7877E-12	1,784	4,083E-07
chr4:40306251	rs7691190	5_CHRNA9_RV	-3,347	0,00000129	2,514	7,906E-08	-0,836	0,0001

Table 19 *The values of each variant in the experiment with the presence of GFP indicating the two variants that surpass the MPRA filters*

Only two probes, which are outside the confidence interval, also fulfil the MPRA filters namely the 5_PRDX5_RV (rs28364831) and 6_TEC-TKX_RV (rs17574371) probes. (Table 20)

Values of the probes that fulfill the MPRA filtering criteria and are outside the confidence interval						
Chromosomic position	Rs ID	Variant name	EBNA experiment		GFP experiment	
			LogFC	Adj p-value	LogFC	Adj p-value
chr11:64317576	rs28364831	5_PRDX5_RV	-1,85	0,0024	-3,271	8,219E-07
chr4:48145192	rs17574371	6_TEC-TKX_RV	0,208	0,01	-0,87609	0,00093686

Table 20 *Variants that surpass all MPRA filters and are outside the confidence interval*

We hypothesize that there are cellular mechanisms that are influenced by the presence of the EBNA2 variant 1.2 which in turn influence the level of expression of the reporter gene, as impacted by each probe represented by logFC. When comparing the Log Fc of these two probes, we notice that there are significant changes between the experiment done with the EBNA variant 1.2 and the control experiment. In particular, the 5_PRDX5_RV variant in the absence of EBNA exhibits a logFC -3.27, which shows a reduced expression of the reporter gene in the presence of the alternative allele. Instead, when exposed to EBNA2 variant 1.2 the logFC is further increased to -1.85, which follows the same trend in reducing the expression in the presence of the Alternative allele but with a less prominent effect. The 6_TEC-TKX_ RV, in the absence of the EBNA gene variant, demonstrates a LogFC of -0.87 indicating a decreased expression in the presence of the Alternative allele. On the other hand, when exposed to the EBNA variant, the tested sequence displays a LogFC of 0.208, which portrays an increased expression in the presence of the alternative allele, possibly an effect due to a differential binding of transcription factors, which are more prominent when the cell is exposed to the EBNA2 gene variant.

This shows compelling evidence of a possible differential effect on expression in the presence of the EBNA2 gene variant 1.2 in variants associated with MS.

Allele Specific Transcription Factor Prediction by MotifBreakR

In order to identify whether there were any Transcription Factors that preferably bind the presence of either alternative or reference allele, we applied MotifBreakR similarly as mentioned before. We were able to obtain information for both variants that changed significantly their effect on the reporter gene. Represented below a table with the information for each SNP.

Table 21 *Results from MotifbreakR on the transcription factors binding to DNA motifs of 5_PRDX5_RS28364831*

Promoter Effect					
5_PRDX5_RV_RS28364831					
Chr.	**SNP_id**	**REF**	**ALT**	**geneSymbol**	seqMatch
chr11	rs28364831	A	C	**E2F4**	**ttgtgggagctgCggtaggtaggtg**
chr11	rs28364831	A	C	**E2F1**	**tttgtgggagctgCggtaggtaggtga**
chr11	rs28364831	A	C	**ZNF436**	**cttctcgtgtttgtgggagctgCggtaggtaggtgaaagacctgtc**
chr11	**rs28364831**	**A**	**C**	**MAF**	**tcgtgtttgtgggagctgAggtaggtaggtgaaagac**

Table 22 *Results from MotifbreakR on the transcription factors binding to DNA motifs of 6_TEC_TKX_RV_RS72924108*

Promoter Effect					
6_TEC_TKX_RV_RS72924108					
Chr.	**SNP_id**	**REF**	**ALT**	**geneSymbol**	seqMatch
Chr4	rs17574371	T	C	**CTCF**	**taacttcgggtgtgtcagCttcctggaaggaagaca**
Chr4	rs17574371	T	C	**TBX3**	**ggtgtgtcagCttcctgggaa**
Chr4	rs17574371	T	C	DUX4	**ggtgtgtcagTttcctgggaa**
Chr4	**rs17574371**	**T**	**C**	**FOXO3**	**gtgtgtcagTttcctggga**

Risk allele association to transcription factors and data interpretation.

For the two variants that showed significant changes in the expression of the reporter gene, we saw that there is differential binding of a transcription factor in the presence of the alternative allele, which in this case for both variants is the risk allele. From now on, we will focus and refer only to the effect predicted by MPRA based on the risk allele, and transcription factors that bind on the same allele.

Variant with data from MotifbreakR	Promoter/Enhancer Effect tested	Risk allele	Reference or Alternative	MPRA effect
5_PRDX5_RV_RS28364831	Promoter	C	Alternative	-
6_TEC-TKX_RV_RS17574371	Enhancer	C	Alternative	-

Table 23 *Risk allele and MPRA predicted effect for the significate variants*

This table shows the two variants with a modified effect on the expression of the reporter gene in the presence of the EBNA2 variant 1.2 corresponding to the risk allele. The effect predicted by MPRA shows a decreased expression in the presence of the alternative allele in the absence of the pathogen, which is modified when the EBNA2 variant is presented to the cells.

5_PRDX5_RV_RS28364831-Transcription Factor Description

Transcription factor binding in the presence of the alternative C allele includes the E2F1. The E2F family plays a crucial role in the control of the cell cycle and action of tumor suppressor protein but is also a target of the transforming protein of small DNA tumor viruses. The E2F1 transcription factor is a transcription activator that binds to DNA cooperatively with DP proteins through the E2 recognition site. However, the E2F1 transcription factor has been studied extensively in the biology of Epstein-Barr virus (EBV). Previous studies have described the involvement of EBV nuclear antigen 3C (EBNA3C) to modulate E2F1 through its interaction with E2F6, inhibiting the E2F1 activity to promote cellular proliferation (Pei et al., 2016). Additionally, (Zhu et al., 2014) reveal that the EBV protein Rta represses interferon regulatory factor 3(IRF-3) by enhancing E2F1 binding to the IRF-3 promoter, effectively diminishing host antiviral defenses. These data support our hypothesis that in the presence of EBNA, the E2F1 enhances its transcriptional activity leading to an increased expression of the reporter gene when in the presence of the alternative allele on the rs28364831.

6_TEC-TKX_RV_RS17574371- Transcription Factor Description

Transcription factor binding in the presence of the alternative C allele includes the chromatin boundary factor CTCF that has been studied in regulating Epstein-Barr latency, particularly how it influences the expression of EBNA2, noted as an important viral antigen. EBNA2 levels influence CTCF levels, as shown by Charles M et al., 2006 by either direct stimulation or indirect stimulation via c-MYC. EBNA2 can directly upregulate CTCF mRNA levels, or indirectly increase its levels by stimulating the expression of the cellular protein c-MYC, which in turn enhances CTCF transcription. Since CTCF levels are increased in the presence of EBNA2, and this transcription factors preferentially binds to the sequence of interest in the presence of the alternative allele, we have reason to believe that this binding influences the expression of the target gene of the SNP. CTCF has been shown to have a function as a transcriptional activator, which could explain the change on the LogFC obtained by MPRA in its presence that is influenced by the EBNA2 variant 1.2.

Exploring Drug Repurposing for target Genes.

Subsequently, we wanted to predict the target gene of the variants that have a modified effect In the presence of a known MS-associated pathogen. Similarly to before we applied the Open Target Genetics tool, which allows the identification of a target gene of the SNPs and a potential drug that targets that gene.

Represented below for each of the two variant's target gene and the effect they have on the target gene as predicted by MPRA, and a drug that has an apposed effect to that predicted by MPRA as the effect predicted is related to the risk allele.

5_PRDX5_RV_RS28364831- Target Gene and Drug.

Prediction done using the Open Target Genetics tool shows that the target gene of this SNP as foreseen by the V2G score is the VEGF-B gene. The effect of the rs28364831 variant on the gene as predicted by Open Target Genetics based on the alternative allele shows a negative Beta value of -0.102 with a p value of 1.5e-33 meaning that this variant has a negative effect on the expression of the VEGF-B gene. This is also confirmed functionally by our experimental evaluation where we see that in the presence of the alternative allele, there are reduced luciferase expression levels. These levels are increased when the cells are introduced to the EBNA2 variant 1.2. As mentioned before the VEGF-B gene is a member of the PDGF/VEGF family, which regulates the formation of blood

vessels and is involved in endothelial cell physiology. To assess an appropriate drug that has an opposing effect to that of the SNP to the gene we have to mention the complex nature of our experimental evaluation. We are evaluating the effect of different SNPs associated with MS on the expression of the reporter gene, which serves as a way to investigate functional variants amongst others. However, all these SNPs are already associated with the disease since they pass the MPRA filters, and the effect on the reporter gene in the presence of the environmental factors in this case shows the same direction of effect. This serves as testimony, to prove subtly that the presence of the EBNA2 variant, alters the expression of some genes in a sensitive and non-robust way, which could lead to other pathways that negatively contribute to patients with the disease. Having stated this, since in the presence of the EBNA2 variant 1.2 the sign of direction (effect on the reporter gene) remains the same we will focus on a Drug that has an agonistic effect, so would help the expression of the VEGFB gene. Using the recombinant human VEGF165 (rhVEGF165) would enhance angiogenesis and improve the blood-brain barrier with better functional neurological outcomes (Zhang et al., 2000).

6_TEC-TKX_RV_RS17574371- Target Gene and Drug

The target gene predicted by Open Target Genetics for this variant is the TEC gene (TEC protein tyrosine kinase) of the TEC family, which also includes Btk, Itk, Rlk, and Bmx. The effect predicted by both the tool and MPRA in the absence of the EBNA2 variant 1.2 suggests a negative effect of the SNP on the TEC gene signifying a reduced expression of the gene when the risk allele is present. However, in the presence of the EBNA2 variant, there is a complete change of Log FC corresponding to this variant. This means that in the presence of the alternative allele, some possible transcription factors binding to that allele are modified in the presence of a known environmental factor associated with MS. This leads to a positive Log FC that means the effect of the variant has changed indicating an increased expression of the TEC gene as predicted by the Open Target Genetics tool. This result is promising since the over-expression of the TEC gene leads to an overrepresentation of immune responses and T-cell activation. A promising drug that has an effect as an inhibitor of the negative effect of the EBNA2 variant 1.2, meaning that reduces the expression of the TEC gene is RITLECITINIB a small molecule witch is a TEC family kinase inhibitor.

Dual Glo Luciferase evaluation

After obtaining the reads from the Tecan microplate reader we compared the number of reads for each alternative and reference fragment that we tested and normalized based on the renilla reads. Our results vary with some confirming the prediction done by the MPRA assay, whereas others show no robust changes between alternative and reference sequences. When regarding to the 9_CD40 probe (RS1883832) and 14_IFNGR2 (RS28653198) represented on **fig.23** we observe that for rs1883832 there is an overexpression of the alternative sequence which is contrary to what we see from the MPRA result. As regarding to rs28653198, we obtained statistically significant result (p.val=0.001) which shows the same trend as predicted by MPRA.

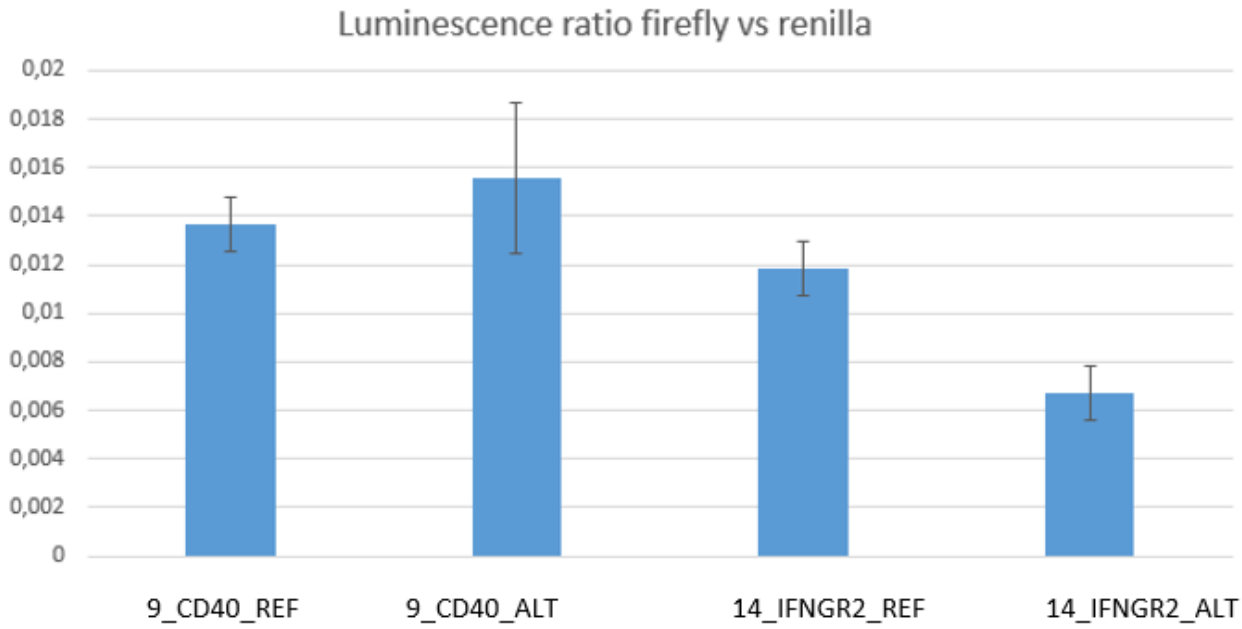


Fig.23 *Graphical representation of the results obtained from the Dual Glo luciferase assay for rs 1883832 and rs28653198*

As regarding the result for the 6_CD40 (RS6074022) and 6_TEC-TKX (RS17574371) we observe that rs6074022 demonstrates a statistically significant result (p.val =0.0025) that goes in the same direction as predicted by the MPRA assay. The rs17574371 shows an opposite effect to that of MPRA. **Fig.24**

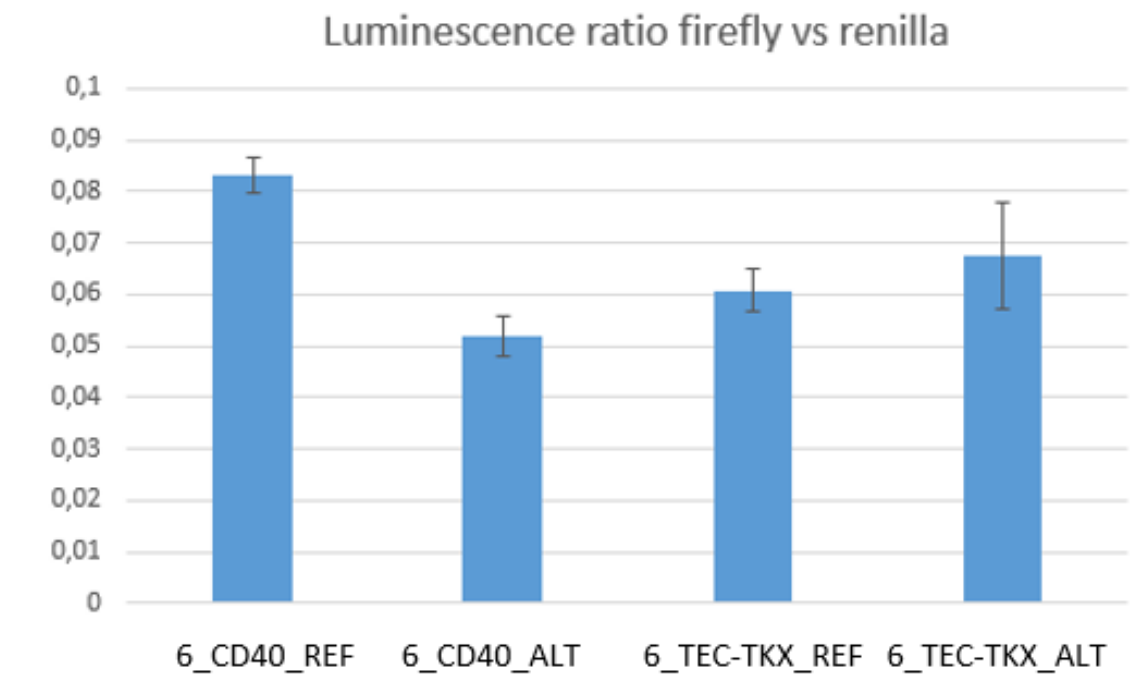


Fig.24 *Graphical representation of the results obtained from the Dual Glo luciferase assay for rs 6074022 and rs17574371*

As regarding to the probes 5_PRDX5 (RS28364831) and 8_PRDX5 (RS72924108) from the dual Glo Luciferase assay we observe that rs28364831 shows an opposite direction as compared to the MPRA results, the same is shown for rs72924108 . **Fig.25**

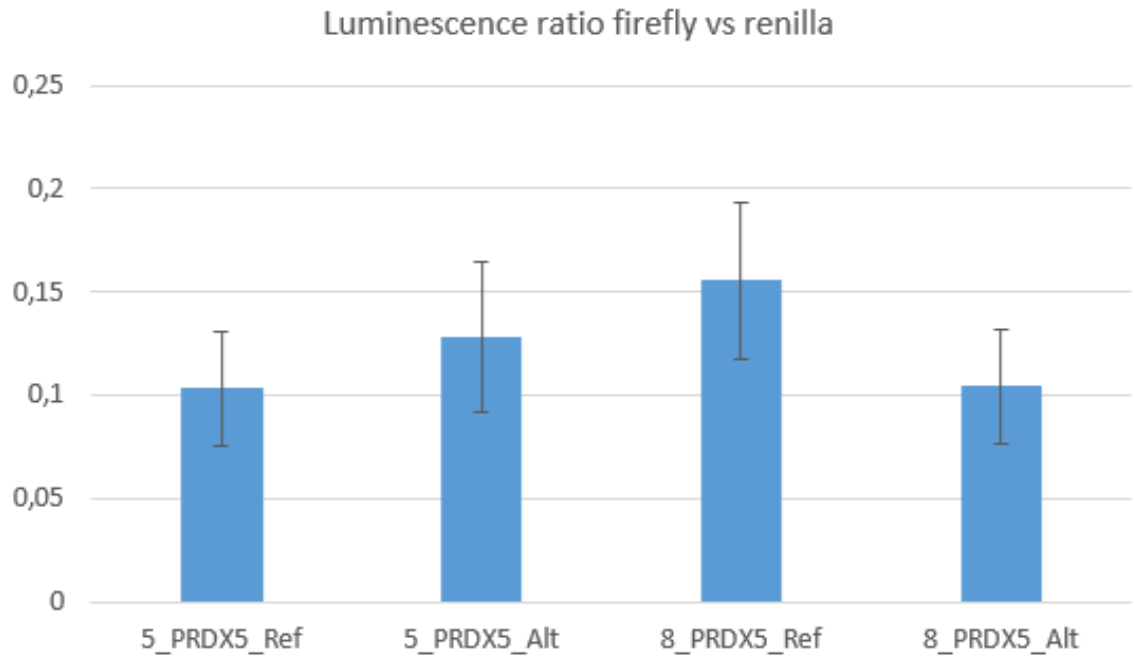


Fig.25 *Graphical representation of the results obtained from the Dual Glo luciferase assay for rs 28364831 and rs72924108*

Finally as regarding to the probes 12_PRDX5 (RS72922077) and 7_IFNGR_INS->G (RS17880053) the results from the dual Glo Luciferase assay show that rs72922077 does not follow the same trend as the MPRA result but rs17880053 follows the same trend as the MPRA result with borderline statistical significance (p.val=0.051). **Fig.26**

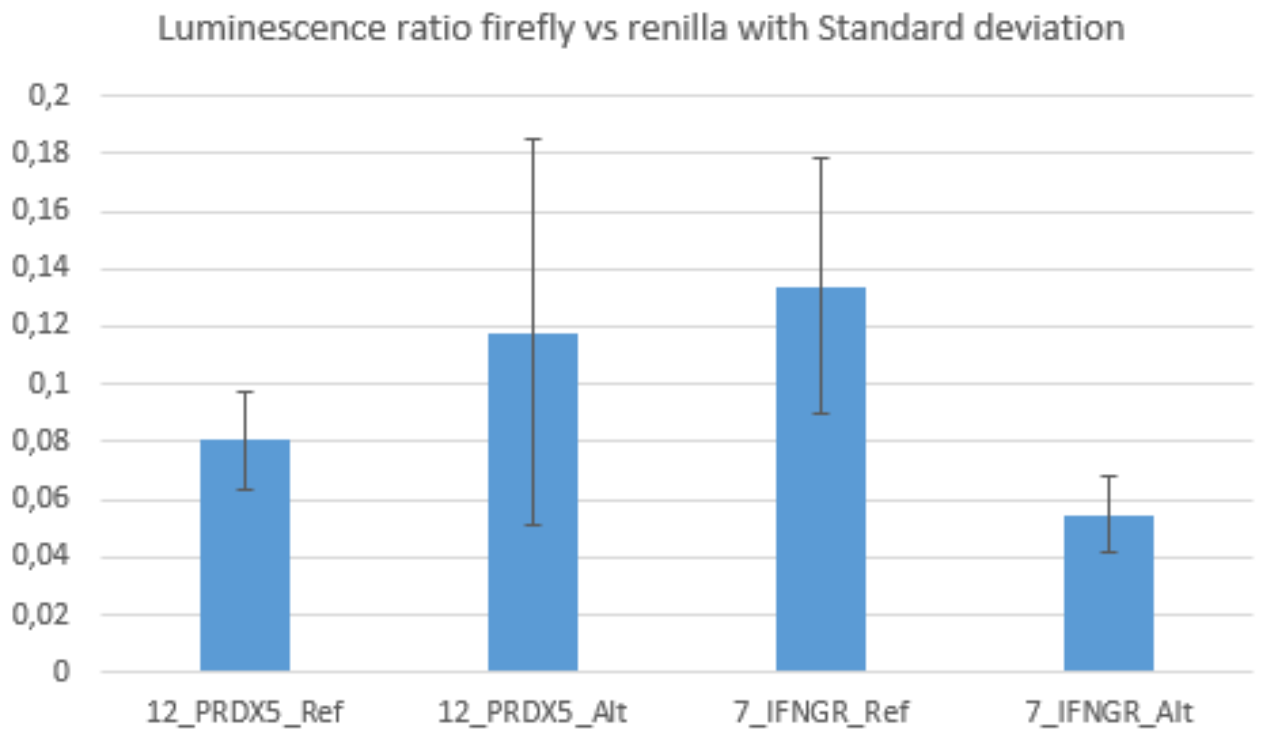


Fig.26 *Graphical representation of the results obtained from the Dual Glo luciferase assay for rs72922077 and rs17880053*

Discussion

Multiple Sclerosis (MS) is a complex autoimmune disease of the central nervous system (CNS) that results in significant neurodegeneration in most of those affected and is a common cause of chronic neurological disability in young adults. As a disease, MS manifests itself in a relapsing-remitting pattern that could be resolved completely or leave deficits. The etiology of MS remains unknown; however, it is assumed to be a close interplay between genetic and environmental factors. Among the genetic factors, the HLA locus, specifically the DRB1*15:01 haplotype, has been identified as the most associated with the disease risk (Sawcer et al., 2011). The development of Genome-Wide Association Studies (GWAS) allowed the simultaneous identification of hundreds of thousands of SNPs, spaced across the entire genome for the association with a particular trait giving new insights into genetic variants that contribute to the disease. The 2019 GWAS study (IMSGC, 2019) increased the number of statistically independent associations with MS susceptibility to 233. They identified 200 risk loci in the autosomal non-major histocompatibility complex (non-MHC) associated with susceptibility to the disease. Additionally, 30 HLA markers, excluding HLA-DRB1*15, and one locus on chromosome X were discovered to be linked to MS. The genome-wide and suggestive effects jointly could explain about 48% of the estimated heritability. As regards the environmental factors associated with MS, the accumulation of epidemiological serological and virological data have increasingly supported the involvement of the Epstein-Barr virus in the genesis of MS. The Epstein-Barr virus is a type of gamma herpesvirus, which results in long-term infection in over 90% of the population worldwide (Keane et al., 2021).

Our study aimed to identify common variants implicated with Multiple Sclerosis and to associate them to a drug, based on the gene they target. In an Italian cohort by utilizing genome-wide association studies, (GWAS) data we wanted to apply two approaches to pinpoint causative variants in highly complex regions characterized by Linkage Disequilibrium. Our two approaches comprise functionally informed Fine Mapping and MPRA (Massively Parallel Reporter Assay). Starting from a large cohort of 5,903 individuals from the continental Italian population, including 4,259 MS patients and 1,644 healthy controls, which, represented around 6,339,414 imputed SNPs covering the entire genome, we wanted to identify potential drug target genes within regions associated with MS. After applying a series of significance filters along with selecting regions associated with MS which were targetable by drugs we were left with 36 regions surrounding the replicated SNPs, which contained a total of 238 genes. Fine mapping analysis conducted using two of the most favorable tools: Paintor and Caviar BF, revealed 19 regions that contain SNPs with compelling evidence of causality with statistical significance. Among these 19 regions, 18 displayed as a possible causal SNP a different one than the lead SNP, which highlights the importance of fine mapping studies. These regions contain SNPs that could contribute to MS by targeting important genes and therefore

influencing their expression. By utilizing Open Target Genetics, we were able to predict the target genes of the SNPs as predicted by the V2G score. These genes are regarded as promising druggable targets due to their abilities to encode cell surface molecules, signaling receptors or co-receptors, or their involvement with immune system regulation. The CD40 gene has been implicated with MS in various GWAS studies, where it has been shown that different variants in this gene are thought to affect immune cell activation and inflammatory processes, both being heavily associated with the disease. As regards the other genes that are not so profoundly studied in association with MS, the function of the encoded proteins could be studied for a possible association with MS. The IFNGR1 gene encodes the interferon-gamma receptor, which has been implicated in various autoimmune diseases given its role in regulating immune responses. The TEC (Tec Protein Tyrosine Kinase) relates heavily with intracellular signaling mechanisms of cytokine receptors, lymphocyte surface antigens, and integrin molecules, all perfect candidates that could indirectly influence MS pathogenesis. TGFBR3 encodes a receptor for the TGF-beta, a cytokine involved in immune regulation, which is associated with the regulation of encephalitogenic and regulatory T cells in MS (Lee W et al 2017). Mutations in the TUBB4A gene have been linked to several neurological disorders but are not known for their involvement with MS, which could leave an open path for further studies. IDE (Insulin Degrading Enzyme) is a protein-coding enzyme that plays a role in the cellular breakdown of insulin, thereby playing a role in intercellular peptide signaling. Diseases associated with IDE include Alzheimer's Disease and Glucose Intolerance, however, yet again the involvement of this gene in MS is the subject of further evaluation. The prediction done by fine mapping is however limited due to several factors. Firstly, the dependence on Functional Annotations where current functional datasets are incomplete and biased due to cell types, tissues, or different experimental conditions, which might be misleading in the context of the disease of interest. Also, many regulatory elements are active only under specific environmental or developmental conditions which might not be captured in the available datasets. Another prominent issue that functionally informed fine-mapping has to face is also Linkage Disequilibrium (LD) which makes it difficult to distinguish the actual functional variant in a region with high linkage disequilibrium. Functionally informed methods also may over-prioritize variants that are associated with well-studied pathways, potentially overlooking discoveries of novel mechanisms. Also, fine mapping it's not able to predict a probable effect of a functional variant in a target gene, it can just predict an involvement of the variant with the disease. These limitations are increased in our case being that the tissue and cell-type specificity requirements in the case of MS are very difficult to pinpoint considering that it involves multiple cell types such as T cells, B cells, and microglia. Nonetheless, even though there are several limitations to fine mapping, its application has been crucial in dissecting data coming from large

GWAS studies, but to have confidence in the results predicted by fine mapping there is the need for the integration of this approach with other in vitro techniques that can surpass the limitations presented. MPRA (Massively Parallel Reporter Assay) can surpass some of the limitations presented by fine mapping, specifically it can be applied in relevant cell types and it can surpass the limitations imposed by Linkage Disequilibrium. We applied the MPRA experiment on 5 out of the 36 MS-associated regions as a pilot experiment, selecting as plausible candidates the regions that showed the highest complexity due to a high number of variants in Linkage Disequilibrium in their respective regions. Specifically, the regions selected to be tested comprise the variants that fall in the CD40 region, TEC-TKX region, PRDX5 region, IFNGR2 region, and the CHRNA9 region. We applied MPRA in a compressive number of 83 SNPs across 5 regions on the Jurkat cell line as a disease-relevant cell line. We applied the experiment two times in the same cell line for added accuracy in our results, and we selected for further evaluation only the variants that showed the same direction of effect on both cell lines and surpassed the MPRA methodology filters. This analysis left us with 8 variants across 4 regions with a statistically significant effect as either promoter or enhancer while also predicting the effect of the variant on the reporter gene which is presumed to follow the same direction in biological conditions. Through the use of MotifBreakR and Meme suite, we were able to predict possible transcription factors that preferentially bind the risk allele of the SNPs. We then performed a careful inspection of the TF that had the highest probability to bind to the risk allele, by employing other tools such as Regulomedb and ENCODE that showed proof of physical binding for some of the significative variants. The transcription factors were then examined to select the ones that exert a function that goes in the same direction as the variant as measured by MPRA. By using Open Target Genetics we were able to predict the target gene of these variants and connect the target gene with the effect of the variant to select a drug that shows an opposing effect to that of the variant on the gene.

To our knowledge, we are the first to have applied the MPRA technique as regards to Multiple Sclerosis on the Jurkat cell line. This is of importance since the effect noticed of each variant is very cell-dependent and context-specific. For this purpose, we will go into detail about the predicted effect of each variant on the target gene and how this can be of relevance regarding MS.

As regarding the CD40 gene, which is the target gene of two SNPs in high LD with the lead SNP located in the region of the CD40 gene, shows an increased expression of this same gene in the presence of the risk allele of these two variants. CD40 is a membrane-bound costimulatory protein and is a member of the tumor necrosis factor receptor (TNFR) superfamily. In normal conditions, CD40 is constitutively expressed by B cells and dendritic cells, but when the cells are activated the protein is then expressed also on other cell types such as hematopoietic cells, including T cells,

monocytes, and macrophages. The classical ligand for CD40 is the tumor necrosis factor (TNF) family member CD40 ligand (CD40L), which is expressed on both T cells and platelets. The interaction between CD40 with its cognate ligand, CD40L (CD154), is required for the primary immune response and accompanying inflammatory responses. Recognition of the specific peptide antigen in the context of the major histocompatibility complex (MHC) on Antigen antigen-presenting cells (APC) by the CD4+ T cells leads to the transient expression of CD40L, which interacts with CD40, resulting in the upregulation of accessory molecules such as CD80 and CD86 on APC. CD80/CD86 in turn provide sufficient co-stimulatory signals for the activation of naïve T cells and differentiation to effector cells. This interaction facilitates the activation and proliferation of T cells, which can migrate to the CNS and initiate an immune response that leads to demyelination and other inflammatory dysregulations. Consequently, exaggerated CD40 signaling has been associated with the pathogenesis of various chronic inflammatory and autoimmune diseases. Autopsy studies in MS patients have revealed that monocytes, macrophages, and activated microglia are the main cell types expressing CD40 in the CNS. (Gerritse K et al 1996) In MS patients there is a high heterogeneity of cells that express CD40 involved in the dysregulation of the immune response. Starting from macrophages and microglia where there is a mixed population of proinflammatory M1 macrophages and anti-inflammatory M2 macrophages that represent the extremities of the spectrum in vivo. (Vogel et al ., 2014) The interplay between CD40 as an M1 marker for perivascular macrophages and other M1-markers along with the CD40L-induced activation by these co-activated cells results in the secretion of M1-associated cytokines and chemokines, including interleukin (IL)-1, IL-6, IL-12, IL-18. All these co-activated cells and molecules contribute in favor of the ongoing inflammation in the CNS. However, 70% of the CD40+ cells, also express M2 markers, suggesting a mixed M1/M2 phenotype exists in MS lesions. (Vogel et al., 2014) Along, with this it is stated by previous studies that the amount of CD40 present in the CNV correlates with the expression of various inflammatory cytokines such as IL-12, IFN- γ, and TNF-α. (Issazadeh S et al., 1998) As regards B cells, the activation of CD40 influences the activation of naïve B cells and memory B cells characterized by CD69 expression, also their proliferative response is increased. In another pathway, the antigen-presenting capacity of B cells is also affected by the ligation of CD40 and CD40L, where it has been seen from MS patients that they exhibited an increased expression of MHC class I and II, which in turn also increases the proliferation of T cells. (Harp CT et al., 2008; Arbour N et al 2006). Following CD40 stimulation, memory and naïve B cells from MS patients showed a significantly higher level of NF-kB activation which correlates with the increased inflammatory response. The role of CD40 expression in T cells is mostly seen in CD4+ and CD8+ cells that initially express low levels of CD40 mRNA but which are increased after activation. (Vaitaitis et al., 2017) Recent further insights in

Genome-Wide Association Studies data have been able to identify a correlation between SNPs in immune-related loci, including the CD40 locus, and the incidence of MS (IMGSC et al., 2013). In particular, there is substantial evidence of the involvement of the rs1883832 C->T and rs 6974022 T->C in the CD40 gene which is associated with an increased risk of MS. These are the two main high-risk SNPs associated with MS, with an influence on the CD40 gene expression. Not coincidently these are the same two SNPs predicted by MPRA as functional SNPs in a region of 12 SNPs tested characterized by high LD. This demonstrates the first positive checkpoint of our analysis where the MPRA technique was able to pinpoint functional variants amongst others with a high probability of affecting gene expression. This is a very crucial point in our study since we can confirm the usefulness of our methodological approach and also of our filtering and selecting criteria to finely select only the variants with a robust effect on gene expression. The second checkpoint MPRA allows us to fulfill is to predict the effect of the allelic variant on the target gene. If either allele has an effect in increasing gene expression or decreasing it. As regards this topic, there are different published works, on different populations, on different cell lines, that attribute the effect of these SNPs in different ways to gene expression. As regards rs1883832 C->T polymorphism, it coincided with the Kozak consensus sequence (GCCACCATGG), which compiled from 699 vertebrate genes, consisted of 6-8 nucleotides before and after the initiation codon (ATG), and flanked the starting methionine (AUG) codon. (Kozak et al., 1987, 1991) The CD40 C->T polymorphism change from a C allele to a T allele (non-conservative change) could cause major alterations in the initiation of the translation of a gene. Functional studies conducted by Jacobson et al., 2005; Skibola et al., 2008) have shown that individuals that carry the T genotypes of the CD40 rs1883832 SNP have lower levels of CD40 on the surface of monocyte-derived activated dendritic cells and B cells. Similarly, Eric M et al., 2005 showed that the same SNP associates also with Graves' disease, where the effect in the presence of the T allele was bound to a reduced expression of the CD40 levels on B cells but did not affect transcription since the mRNA levels of CD40 were not modified. Other studies such as that from Zhang et al., 2013 showed that the distribution of the TT allele was associated with a decreased CD40 mRNA in patients with cerebral infarction. However, most studies confirm the effect of the SNP in downregulating or overexpressing cellular markers such as CD40 (decreased) or CD27 (increased) in B-cell subtypes but are limited to the effect of the SNP on the CD40 gene expression (Orrù V et al., 2020). This shows that the effect of rs1883832 in gene expression as well as cellular markers expression is very cell type dependent and tissue-specific, and since the T/C alleles are both implicated with relatively equal frequencies between them it's hard to attribute a specific effect to the variant. In our study, we have clear evidence of the implication of rs1883832 on luciferase gene expression as measured by our in vitro assay, and this variant is a strong candidate given its effect on

CD40 which has heavily been associated with MS, Grave's disease, and rheumatoid arthritis (Raychaudhuri S et al., 2008). Li G et al 2018., demonstrate similar results showing that in the presence of the T allele, there is increased luciferase activity as opposed to the C allele in a resembling experiment. Similarly, rs6074022 T->C has been implicated genetically with MS (Sokolova EA et al., 2013), but understudied regarding the effect on gene expression. Studies conducted by Li G et al., 2018 have proven similar results in an in vitro way, where they confirm increased luciferase activity in the presence of the risk allele of rs6074022. Both variants are in linkage disequilibrium with each other, which when inherited together could increase the risk of MS. Since from our results we have evidence that both variants influence gene expression by increasing expression as measured by the reporter assay, and have predicted the CD40 gene as their target gene. Exaggerated CD40 signaling has been associated with the pathogenesis of various chronic inflammatory and autoimmune diseases. It has been documented that the blockade of the CD40-CD40L interaction can protect against the progression of antibody-and cell-mediated autoimmune diseases. In this case, a possible drug would be Iscalimab (CFZ533), which is a monoclonal antibody, fully human; pathway blocking that inhibits the tumor necrosis factor receptor superfamily member 5. By interfering with the interaction between CD40 and its ligand, CD154, Iscalimab exerts its therapeutic effects through pathway blockade without depleting CD40. The CD40-CD154 (CD40 ligand; CD40L) co-stimulatory pathway plays a critical role in T-cell-dependent humoral immune responses, the development of human memory, and the function of antigen-presenting cells. CFZ533, by effectively blocking the CD40 pathway, has the potential to inhibit CD40 pathway-dependent effector functions across various cell types, if it achieves sufficient receptor occupancy. These unique mechanisms position Iscalimab as a promising candidate for modulating immune responses in various diseases. Given its successful development in other autoimmune and inflammatory conditions, there is potential for Iscalimab to be repurposed as a therapeutic option for Multiple Sclerosis allowing for targeted modulation of immune pathways associated with MS pathogenesis (Ristov et al., 2018)(Kahaly et al., 2020).

As regards the PRDX5 region 3 out of the 24 tested SNPs showed compelling evidence in influencing gene expression as measured by the MPRA. Since these are common SNPs there is no evidence in literature as to their involvement with Multiple Sclerosis. The target gene of all three of the significative variants is the VEGF-B gene. VEGF-B is a member of the PDGF/VEGF family, which regulates the formation of blood vessels and is involved in endothelial cell physiology. VEGF-B plays an important role in several types of neurons showing a protective role of neurons in the retina and the cerebral cortex and of motor neurons such as in amyotrophic lateral sclerosis. It is essential for normal vascular development and homeostasis and is also implicated in neurodegeneration it may play a protective role in Alzheimer's disease (AD) since it is reduced in AD as demonstrated by

patients serum *in vivo*. Recent genetic studies have revealed that reduced VEGF levels cause neurodegeneration in part by impairing neural tissue perfusion (Storkebaum et al., 2004).

More specifically, microglia and astrocytes are known to modulate inflammation and neurodegeneration in the central nervous system (Haim L et al 2017). Rothhammer et al., 2019 have reported that TGF-α and VEGF-B produced by microglia regulate the pathogenic activities of astrocytes in the experimental autoimmune encephalomyelitis (EAE) of MS. VEGF-B produced by microglia can trigger FLT-1 signaling in astrocytes worsening EAE symptoms. Transcriptional analysis suggests that VEGF-B regulates NF-kB in astrocytes together with TGF-α, which are known to drive their pathogenic activities during CNS inflammation during MS flare-ups. The interplay between TGF-α and VEGF-B regulates astrocytic NF-kB in an opposing manner, where VEGF-B activates it and TGF- α inhibits it. Thus, the levels of VEGF-B by itself are not sufficient to determine MS severity, since its levels vary throughout MS subtypes. Cirac et al., 2021 demonstrated that the levels of TGF-α were reduced in patients with RRMS (relapsing-remitting) and SPMS (secondary progressive), but not in PPMS (primary progressive), controversially, VEGF-B levels were increased in patients with SPMS, but not in RRMS and PPMS. Therefore, to account for these observations, they determined the TGF-α/VEGF-B ratio, which is decreased in chronic MS lesions and represents a marker for neurodegeneration and astrogliosis. Their data suggests that the ratio between TGF-α and VEGF-B is altered in various subtypes of MS. Our data has been able to pinpoint three SNPs with a supposed effect in the VEGF-B gene as predicted by Open Target genetics, respectively rs28364831, rs72922077, rs72924108. RS28364831, which shows reduced luciferase activity in the presence of the risk allele similar to rs72922077. Controversially, rs72924108 exhibits higher luciferase activity in the presence of the risk allele, whereas Open Target Genetics demonstrated VEGF-B as a target gene of the variant. We hypothesize that these variants which are in high LD with one another, together influence the expression of the VEGF-B gene, thus altering the ratio between VEGF-B and TGF-α, contributing in still unknown ways to MS progression. Possible therapeutic approaches could be taken into account once the relation between MS-associated variants and the TGF-α/VEGF-B ratio is established to find the correct therapeutic approach, however, VEGF165 has an agonistic role to that of VEGF-B and could be used as an agonistic drug, or CONBERCEPT, which is a protein with a VEGF-B inhibitor effect.

As for the IFNGR2 region MPRA predicted 2 variants out of the 22 tested replicated in both cell lines with a significant influence on luciferase activity as promoter or enhancer, all targeting IFNGR2. IFN-γ is a soluble cytokine that is secreted by different immune cells. Functionally it can bind to the IFNGR1/IFNGR2 to activate the Janus kinase (JAK) signal transducer and the transcription protein (STAT) pathway to coordinate different cell functions such as immune regulation, leukocyte

transportation, etc. Importantly, while IFNGR1 is required for ligand binding, IFNGR2 is crucial for initiating downstream activity. IFNGR2 plays a crucial role in regulating the immune response by its involvement in both innate and adaptive immunity. However, IFNGR2 has been studied intensively in autoimmune diseases playing a pro-inflammatory role through T cell activation, macrophage activation, and cytotoxicity. Regarding Multiple Sclerosis IFNGR2 has been shown to contribute to neuroinflammation by mediated IFNG-induced activation, leading to activation of Th1 cells, which infiltrate the CNS. Overexpression of IFNGR2 is commonly associated with excessive proinflammatory response, however in relapsing-remitting autoimmune disorders immune cells are constitutively and constantly exposed to waves of significantly “high” or “low” levels of type 1 and type 2 INFs during distinct periods (Zhang Q et al., 2016; Banchereau, J et al.,2006). This pre-exposure to low sub-activating concentrations of IFNs sensitizes cells to produce enhanced responses to extracellular stimuli that induce IFNs as well. This process known as priming is characterized by the accumulation of STATs. Primed immune cells such as M1 macrophages are the predominant phenotype at sites of inflammation for diseases such as MS. More in detail in mouse models, it is known that IFNG-negative mice are highly susceptible to EAE. This highlights the dual role of IFNG in autoimmune disorders since targeting its functions at different stages is essential to understand its role. (Bettelli et al., 2004) In our case, both variants pinpointed by MPRA demonstrate a reduced luciferase activity in the presence of the risk allele for rs17880053 and rs28653198 with the IFNGR2 gene as a target, possibly attributing to a consistent dysregulation of INFGR2 expression which ultimately attributes to disease progression. However, IFN signaling is largely controlled in a multidimensional manner, where timing, exposure levels, target organs, and cellular environment determine the immune response to IFN alteration. As regards possible therapeutic approaches INTERFERON GAMMA-1B could be a possible candidate that exerts an antagonist effect of that of the SNPs on the gene. However, interventions altering IFNGR2 expression should be applied with wariness since it can influence the balance of IFNs in protecting or developing autoimmune disorders. Regarding the TEC-TKX region out of the 23 tested SNPs, MPRA revealed only one variant as likely causative showing a decreased luciferase activity in the presence of the risk allele, specifically the rs17574371 variant. The target gene of this variant is the TEC gene (Tec protein tyrosine kinase) of the TEC family kinases which also includes BTK, ITK, RLK, and BMX. They are non-receptor tyrosine kinases involved in signaling pathways downstream of various receptors. In the context of T cells, three TEC kinases, namely ITK, TEC, and RLK, are expressed and activated downstream of the T cell receptor (TCR). RLK and ITK have been specifically implicated in T-helper cell development. Maintaining a proper balance between the two subsets of T helper cells is crucial for mounting effective immune responses against pathogens. Conversely, imbalances between these

subsets have been associated with disease conditions, such as autoimmune disorders characterized by an excess of TH1 cells and hypersensitivity disorders characterized by an excess of TH2 cells. The effect predicted by MPRA of the variant indicates a reduced expression of the TEC gene in the presence of the risk allele which is controversial given that lately there have been Btk inhibitors approved for Multiple Sclerosis. However, previous studies have shown that Btk and Tec may have redundant roles in B cell activation and that Tec can partially compensate for the loss of Btk. (Ellmeier et al., 2000). Marjolein et al.,2017 provide evidence that loss of Tec leads to increased B cell activation through enhanced AKT activation, whereby increased phosphorylation is dependent on BTK kinase activity. They show that Tec-deficient mice mature B cells show increased activation, proliferation, and survival upon anti-CD40 stimulation in vitro. Moreover, Tec-deficient mice have enhanced humoral immunity upon immunization and develop mild autoimmune phenotype upon aging. This data shows new insight into the role of Tec in B cell activation. Most drugs available for TEC are inhibitors given recent favorable results in MS by using BTK inhibitors. Since Akt is activated in TEC-deficient mice leading to increased activation of mature B cells, promoting inflammation and immune response, a probable drug to be used as an inhibitor of the Akt pathway is Capivasertib (AZD5363). Capivarsertib is a serine/threonine kinase inhibitor used to treat hormone receptor-positive, HER2-negative breast cancer, which could potentially be repurposed in autoimmune treatment in appropriate dosage and after clinical trial assessment.

Our hypothesis to test the effect of the Epstein-Barr EBNA2 variant 1.2 in changing the effect on luciferase activity of SNPs associated with MS using the MPRA technique revealed only two variants that showed statistically significant changes in expression. Respectively rs28364831 on the PRDX5 region and rs17574371 on the TKX-TKX region. The Epstein-Barr virus is a gammaherpesvirus that establishes persistent infection in more than 90% of the global population and it has been identified as a risk factor for developing MS (Kaene et al.,2021). Recent extensive studies suggest that Epstein-Barr infection is likely a necessary factor for the development of the disease (Soldan and Lieberman, 2023). Even though more than 200 genetic variations associated with the risk of MS have been identified, and 47 of these are linked to functions of the Epstein-Barr virus, the interplay between these risk-associated genetic variations and EBV that might influence the susceptibility to MS is still not well established. Epstein-Barr Nuclear Antigen 2 (EBNA2) is essential for maintaining the latency III growth phase of EBV and acts by regulating both viral and cellular genes. There has been data proving an association between several autoimmune disorders including MS, which have an over-representation of EBNA2 binding sites at disease-risk loci in EBV-infected B cells (Harley et al., 2018). Published work from our collaborators Mechelli et al., 2015 have studied five major alleles of the EBV type 1 strain, the most frequent strain in the Caucasian population, which were identified

based on the nucleotide variation within the most variable region of EBNA2. Specifically, they showed that the MS risk significantly correlates with an excess of the 1.2 allele of the EBNA2 gene (odds ratio (OR) =5.13; 95% confidence interval (CI) 1.84-14.32; p=0.016). Our results suggest that the presence of the 1.2 allele of the EBNA2 gene influences the effect of both variants as measured by luciferase activity. Specifically, rs28364831 in the PRDX5 region in the presence of the 1.2 allele of the EBNA2 gene shows an increased luciferase activity regarding the risk allele. The target gene of this variant is the VEGF-B gene which Rang et al., 2022 have identified as an MS-related gene regulated by EBV-encoded micro RNAs, suggesting a complex interplay between viral infection and gene regulation in MS. Similarly, rs17574371 on the TEC-TKX region shows prominent increased luciferase activity in the presence of the pathogen respective to the MS risk allele as well. The target gene of this variant is the TEC gene, which when overexpressed leads to an increased proinflammatory response attributing to MS. We attribute these differences in expression to the preferential binding of transcription factors, which could be more abundant in the presence of the pathogenic environmental factor. Analysis conducted with motifbreakR predicted the E2F1 as a probable TF binding to the risk allele of rs28364831, which is influenced by the presence of the pathogen. In detail, the E2F1 transcription factor is a transcription activator that binds to DNA cooperatively with DP proteins through the E2 recognition site. E2F1 transcription factor has been studied extensively in the biology of Epstein-Barr virus (EBV). Previous studies have described the involvement of EBV nuclear antigen 3C (EBNA3C) to modulate E2F1 through its interaction with E2F6, inhibiting the E2F1 activity to promote cellular proliferation (Pei et al., 2016). Additionally, (Zhu et al., 2014) reveals that the EBV protein Rta represses interferon regulatory factor 3(IRF-3) by enhancing E2F1 binding to the IRF-3 promoter, effectively diminishing host antiviral defenses. These data support our hypothesis that in the presence of EBNA, the E2F1 enhances its transcriptional activity leading to an increased expression of the PRDX5 gene when in the presence of the alternative allele on the rs28364831. Analysis conducted with MotifBreakR demonstrates a high number of transcription factors that preferentially bind to the alternative allele of rs17574371. One promising candidate is the chromatin boundary factor CTCF which has been studied in regulating Epstein-Barr latency, particularly how it influences the expression of EBNA2, noted as an important viral antigen. EBNA2 levels influence CTCF levels, as shown by Charles M et al., 2006 by either direct stimulation or indirect stimulation via c-MYC. EBNA2 can directly upregulate CTCF mRNA levels, or indirectly increase its levels by stimulating the expression of the cellular protein c-MYC, which in turn enhances CTCF transcription. Since CTCF levels are increased in the presence of EBNA2, and this transcription factors preferentially binds to the studied sequence in the presence of an alternative

allele, we have reason to believe that this binding influences the expression of luciferase activity, since CTCF has been shown to have a function as a transcriptional activator.

These data show compelling evidence for using MPRA as a technique to test changes in the luciferase activity of disease-associated variants in the presence of known environmental factors. Following our pipeline MPRA can be used to test the involvement of other environmental factors in influencing the effect of disease-associated SNPs.

As regards to the results obtained from the luciferase assay, we attribute the discrepancies for a few of the probes from this experiment to the results from the MPRA technique mainly to the fact that some of the variants tested don't have such a robust effect as compared to the variants that follow the same trend with statistical significance. Specifically, two variants on the IFNGR2 region follow the same trend as to that predicted by MPRA respectively rs28653198 (intron variant 6,676 bp distant from start site) and rs17880053 (2,222 bp upstream of the IFNGR2 gene) with statistical significance. **Fig.27** shows the elevated LD between the Lead SNP and the two variants that have emerged as significant from the MPRA analysis along with their architectural complexity. These variants are confirmed through multiple techniques to be functional SNPs and by the single luciferase assay. This makes this variants highly interesting for follow up techniques in discovering the exact pathway by which they influence gene expression.

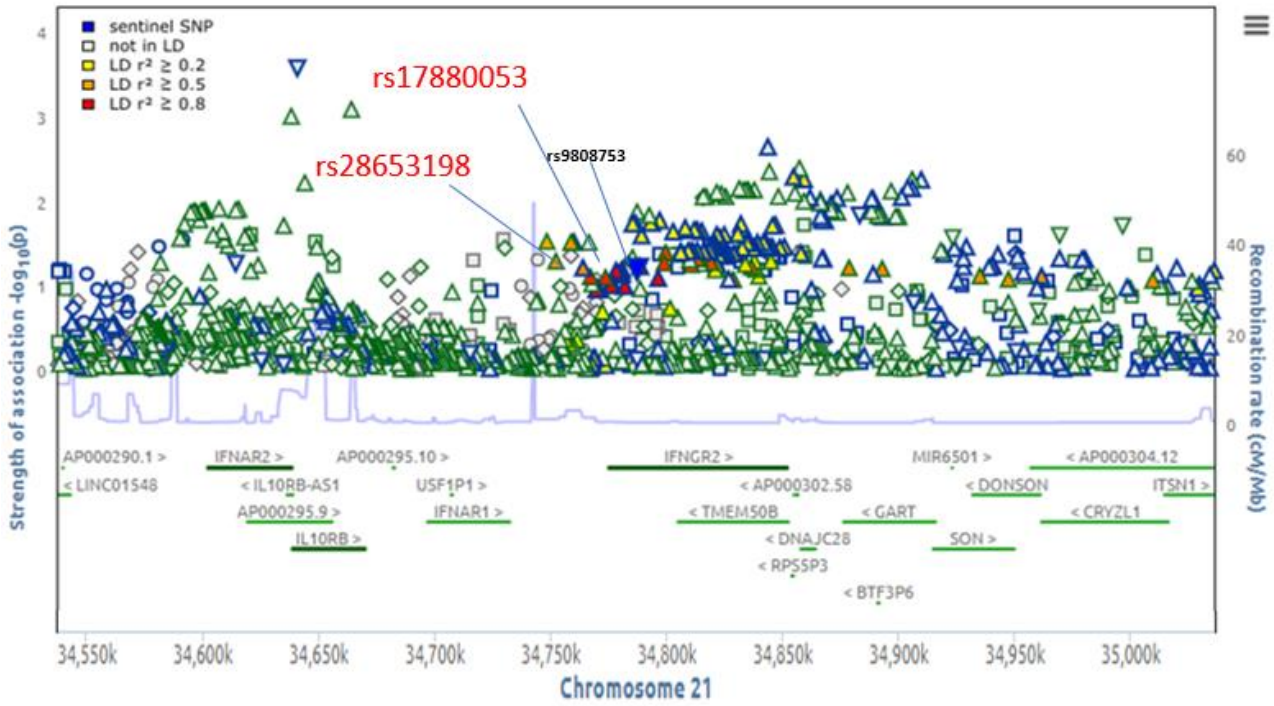


***Fig.27** Regional plot of the IFNGR2 region and location of functional SNPs.*

The CD40 region is one of the regions that fine mapping predicted rs6065926 as a functional SNP with the highest probability. However, when we tested by MPRA all of the variants in high LD with the lead SNP comprising rs6065926, only rs1883832 and rs6074022, emerge as functional variants. This shows the relevance of in vitro assays to pinpoint real functional variants by surpassing the limitations of analysis as fine mapping. When we then tested both variants in a single luciferase assay, we observed that only rs6074022 located 6,757 bp upstream of the CD40 gene shows persistent direction of effect to that of MPRA with a statistical significance. This is probably due to the robustness of the effect that this SNP shows. All three of these SNPs (rs1883832, rs6074022 and rs6065926) are in high LD to the lead SNP in the complicated region as demonstrated in **fig.28,** but only rs6074022 shows a very reliable constant effect as predicted throughout the entire experimental approach. This makes this SNP highly valuable for further follow up as mentioned also above, to understand completely how this functional SNP influences gene expression.

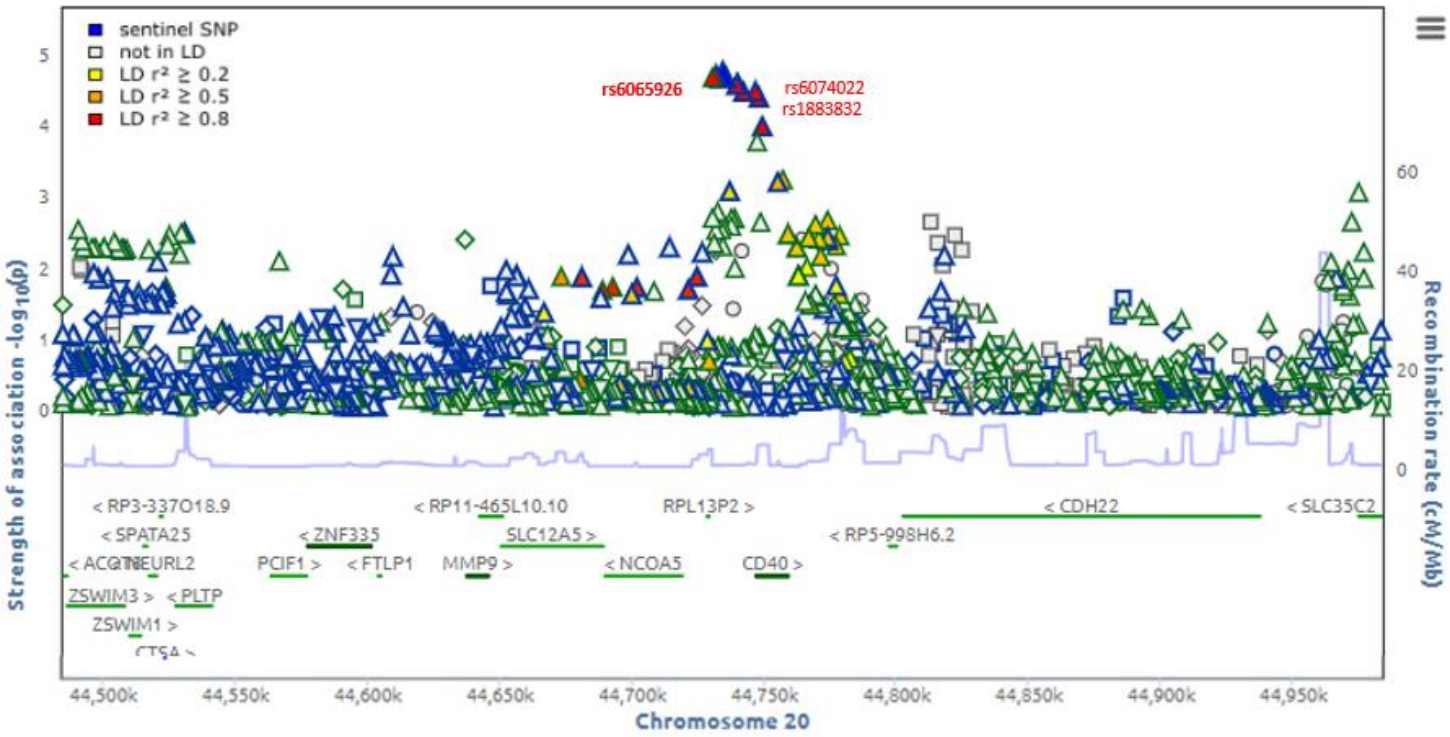
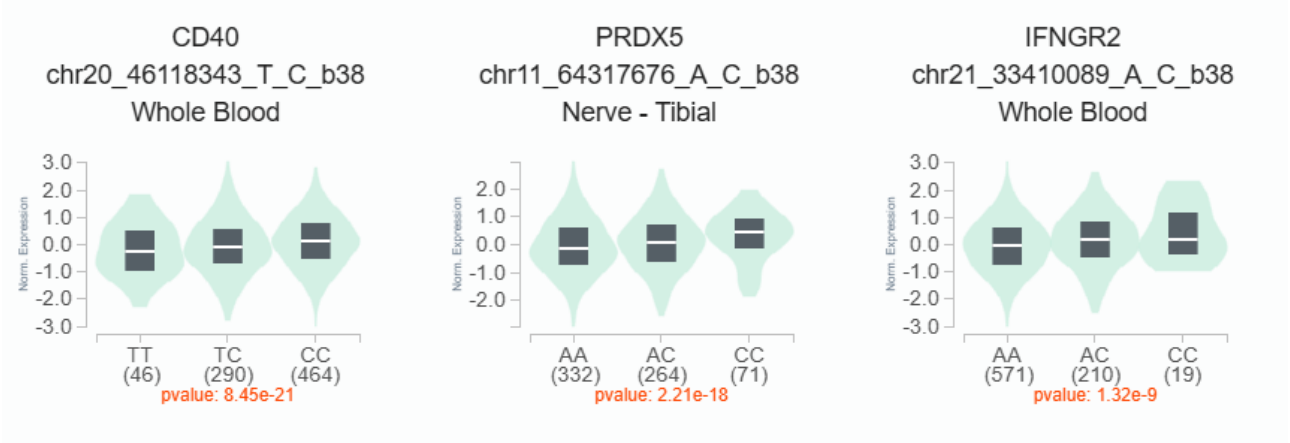
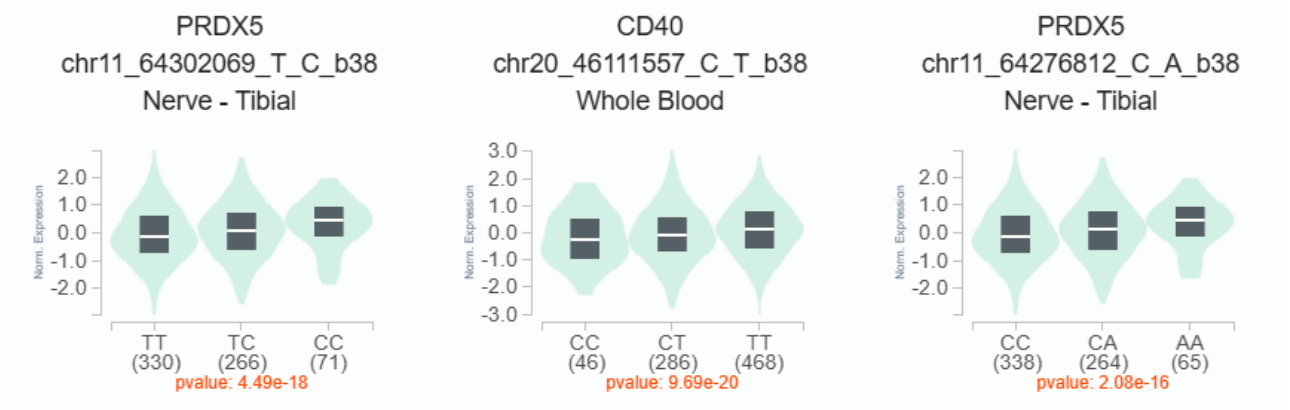


Fig.28 *Regional plot of the CD40 region and location of functional SNPs*

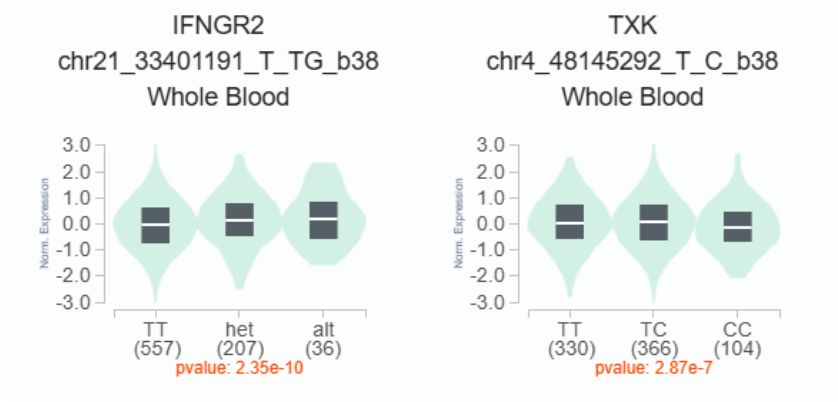
Represented on **Fig.29** the eQTL prediction on various tissues for the 8 MPRA significative variants. Regarding to the rs1883832 on the CD40 region the effect observed on the presence of the risk **T** allele is in the same direction to that of MPRA showing a decreased expression. In addition, about rs6074022 on the same region, when in the presence of the risk allele **C** the eQTL data shows results in the same direction as MPRA, so a decreased expression on the presence of the C. On the PRDX5 region, rs28364831 in the presence of the MS-risk allele **C** shows an opposing effect to that observed by MPRA, this is probably due to the cell type specific effect. The same situation is shown regarding rs72922077 on the same region, where in the presence of the A risk allele there is an increased expression, different from the MPRA results. Another variant on the PRDX5 region, rs72924108 shows an eQTL effect in the same direction as to that predicted by MPRA, so increasing the expression of the gene in the presence of the **C** risk allele. Regarding the two variants on the IFNGR2 region rs28653198 risk allele **C**, and rs17880053 with risk allele the insertion of a **G**, the effect observed is opposed to what MPRA predicts, likely due to cell type specific effect. Lastly, rs17574371 on the TEC region, in the presence of the rick allele C, shows an eQTL effect similar to that of MPRA.



Rs1883832 **Rs28364831** **Rs28653198**



Rs72924108 **Rs6074022** **Rs72922077**



Rs17880053 **Rs17574371**

Fig.29 *eQTL results on various tissues for the MPRA significant variants*

Conclusions and Future Perspectives

Our project gives novel insight into MS-associated regions coming from GWAS data through the application of a specific pipeline, which allowed us to pinpoint variants with an effect on gene expression associated with MS among redundant ones.

Starting from 36 MS-associated regions, we applied our methodological approaches to pinpoint real causative variants amongst unrelated ones in high linkage disequilibrium between each other. Fine mapping analysis were able to predict 19 regions out of which 18 displayed as a possible causal SNP a different one than the lead SNP showing the most statistically significant association in the original IMSGC article out of which 10 regions have causal SNPs that target the drug target gene. Fine mapping studies have shown to be of pivotal importance in the follow-up of GWAS studies since they facilitate the identification of exact genetic factors contributing to a disease thus giving rise to new treatment strategies. However, given the limitations presented by LD fine mapping studies require further evaluation with in vitro techniques. We successfully applied the MPRA technique on multiple cell lines for 5 of the most architecturally complex regions amongst the 36 regions associated with MS. Results from MPRA confirm at least one disease-related functional variant which alters luciferase activity with a statistical significance for 4 out of the 5 tested regions. Follow-up annotation demonstrated that the variants that were predicted to influence gene expression by MPRA are located in enhancer and promoter regions, adding more confidence to our findings. Further evaluation through the use of allele-specific transcription factor prediction tools revealed several TF of interest that could explain the effect of the variant in the presence of the risk allele. Additional assessment by relying on literature and online databases such as Open Target Genetics allowed us to predict a target gene of these functional variants and its possible implication with MS. In conclusion, through the integrated use of fine mapping and MPRA, we were able to accurately predict common SNPs associated with MS that have a possible role in gene expression. We also predicted possible transcription factors that could help facilitate the disease's favorable effect of the variants which are more bound to bind to the sequence in the presence of the risk allele. Within the target genes of the successfully analyzed variants are genes that are known to be involved in MS such as the CD40 gene, TEC gene, and IFNGR2 whereas others are less studied such as the VEGF-B gene. Still, the effect of the variants predicted by MPRA gives novel insights that have not been documented before or to which there is little evidence, showing intricate novel ways by which functional variants influence the expression of genes that ultimately lead to disease onset or progression.

We were also able to successfully use the MPRA technique as a vessel to measure the changes on luciferase activity of MS-associated variants in the presence of known MS-associated environmental factors such as the EBNA2 allele 1.2. Out of 83 tested SNPs were able to pinpoint two variants that had a statistically significant change in their luciferase activity in the presence of the MS-associated

pathogen. Following the same pipeline and relying upon literature we discovered preferential binding of TF that are influenced by the presence of the EBNA2 allele 1.2 sequence which could explain the change in the effect of these variants.

However, most of our results are preliminary and require further evaluation with more specialized analysis. Our future perspectives are based on applying more in vitro techniques that physically demonstrate the binding of the Insilco-predicted TF to the sequence of interest. A possible follow-up path would be the application of the type IIS enzyme restriction technique, which allows for a more specific investigation of the physical binding of TF without having a candidate TF. This would allow for the unbiased selection of TF that preferentially binds to the risk allele and not to the non-risk allele in a physical manner. A potential approach for validation of our findings is also the one employed by Long et al., where a combination of fine-mapping, MPRA analysis, and motifbreakR was followed by a validation step using CRISPRi. This validation method involves the targeted manipulation of a limited number of variants using CRISPRi technology to assess their functional impact. By directly modulating the expression of specific variants, this approach provided additional evidence of their functional significance in the context of the studied disease. CRISPRi utilizes the dCas9-sgRNA complex to regulate gene expression through two mechanisms. In certain genomic regions, it can block transcription elongation by binding to the non-template DNA strand, preventing RNA polymerases from progressing. Alternatively, when targeting the promoter sequence or transcription factor binding sites, it can hinder transcription initiation by obstructing the binding of RNA polymerase or transcription factors. Importantly, the silencing effect is not influenced by the targeted DNA strand. (Larson et al. 2013; Long et al. 2022) In addition, ChIP-seq is a valuable technique for the genome-wide identification of protein binding sites. It enables the precise mapping of DNA regions bound by specific proteins throughout the entire genome. This method has emerged as the preferred approach for comprehensive profiling of protein-DNA interactions across various organisms. (Bansal et al. 2015) After conducting statistical fine-mapping, another promising method that can be utilized is STING- seq. STING-seq is a high-throughput technique that leverages CRISPRi to identify functional variants identified through GWAS. Since it was recently published in 2023, its strengths and limitations are yet to be fully characterized. However, given its potential, STING-seq could become an eligible technique in the investigation of functional variants associated with complex diseases. (Morris et al. 2023)

Bibliography

- Abecasis G., Altshuler D., Auton A., Brooks L., Durbin R., et al. , 2010. A map of human genome variation from population-scale sequencing. Nature 467(7319): 1061–1073
- Afrasiabi A, Parnell GP, Fewings N, Schibeci SD, Basuki MA, Chandramohan R. Evidence from genome wide association studies implicates reduced control of Epstein-Barr virus infection in multiple sclerosis susceptibility. Genome Med. 2019;11(1):1–13. doi: 10.1186/s13073-019-0640-z
- Arbour N, Lapointe R, Saikali P, McCrea E, Regen T, Antel JP. A new clinically relevant approach to expand myelin specific T cells. J Immunol Methods (2006) 310(1–2):53–61. 10.1016/j.jim.2005.12.009
- Ascherio A, Munger KL. Simon KC. Vitamin D and multiple sclerosis. Lancet Neurol. 2010;9:599–612. doi: 10.1016/S1474-4422(10)70086-7
- Bansal M, Mendiratta G, Anand S, et al (2015) Direct ChIP-Seq significance analysis improves target prediction. BMC Genomics 16:S4. https://doi.org/10.1186/1471-2164-16-S5-S4
- Barcellos LF, Oksenberg JR, Begovich AB, Martin ER, Schmidt S, Vittinghoff E, et al. HLA-DR2 dose effect on susceptibility to multiple sclerosis and influence on disease course. Am. J. Hum. Genet. 2003;72:710–716. doi: 10.1086/367781
- Barnett MH. Prineas JW. Relapsing and remitting multiple sclerosis: pathology of the newly forming lesion. Ann. Neurol. 2004;55:458–468. doi: 10.1002/ana.20016
- *Bar-Or A, Pender MP, Khanna R, Steinman L, Hartung HP, Maniar T. Epstein–Barr Virus in Multiple Sclerosis: Theory and Emerging Immunotherapies. Vol. 26. Trends in Molecular Medicine. Elsevier Ltd. 2020:296–310. doi: 10.1016/j.molmed.2019.11.003.*
- Beecham, A. H. et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat. Genet. 45, 1353–1360 (2013)
- Belbasis L, Bellou V, Evangelou E, Ioannidis JP, Tzoulaki I. Environmental risk factors and multiple sclerosis: an umbrella review of systematic reviews and meta-analyses. Lancet Neurol. 2015 Mar;14(3):263-73. doi: 10.1016/S1474-4422(14)70267-4. Epub 2015 Feb 4. PMID: 25662901.
- Berer, K. *et al.* Commensal microbiota and myelin autoantigen cooperate to trigger autoimmune demyelination. *Nature* 479, 538–541 (2011).
- Berkovich R. Treatment of acute relapses in multiple sclerosis. Neurotherapeutics. 2013;10:97–105. doi: 10.1007/s13311-012-0160-7.

- Boyle AP, Hong EL, Hariharan M, et al (2012) Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 22:1790–1797. https://doi.org/10.1101/gr.137323.112
- Brynedal, B. et al. HLA-A confers an HLA-DRB1 independent influence on the risk of multiple sclerosis. PLoS ONE 2, e664 (2007).
- Choi J, Zhang T, Vu A, Ablain J, Makowski MM, Colli LM, Xu M, Hennessey RC, Yin J, Rothschild H, Gräwe C, Kovacs MA, Funderburk KM, Brossard M, Taylor J, Pasaniuc B, Chari R, Chanock SJ, Hoggart CJ, Demenais F, Barrett JH, Law MH, Iles MM, Yu K, Vermeulen M, Zon LI, Brown KM. Massively parallel reporter assays of melanoma risk variants identify MX2 as a gene promoting melanoma. Nat Commun. 2020 Jun 1;11(1):2718. doi: 10.1038/s41467-020-16590-1. PMID: 32483191; PMCID: PMC7264232.
- Cohen JA, Barkhof F, Comi G, Hartung HP, Khatri BO, Montalban X, et al. Oral fingolimod or intramuscular interferon for relapsing multiple sclerosis. N. Engl. J. Med. 2010;362:402–415. doi: 10.1056/NEJMoa0907839
- Ebers GC, Sadovnick AD (1994) The role of genetic factors in multiple sclerosis susceptibility. J Neuroimmunol 54:1–17. https://doi.org/10.1016/0165-5728(94)90225-9
- Ernst J, Melnikov A, Zhang X, Wang L, Rogov P, Mikkelsen TS, Kellis M, Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. Nat. Biotechnol. 34, 1180–1190 (2016)
- Fischbach F, Richter J, Pfeffer LK, Fehse B, Berger SC, Reinhardt S, Kuhle J, Badbaran A, Rathje K, Gagelmann N, Borie D, Seibel J, Ayuk F, Friese MA, Heesen C, Kröger N. CD19-targeted chimeric antigen receptor T cell therapy in two patients with multiple sclerosis. Med. 2024 Jun 14;5(6):550-558.e2. doi: 10.1016/j.medj.2024.03.002. Epub 2024 Mar 29. PMID: 38554710.

  framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315. https://doi.org/10.1038/ng.2892
- Frohman EM, Racke MK. Raine CS. Multiple sclerosis–the plaque and its pathogenesis. N. Engl. J. Med. 2006;354:942–955. doi: 10.1056/NEJMra052130.
- Fujinami RS. Oldstone MB. Amino acid homology between the encephalitogenic site of myelin basic protein and virus: mechanism for autoimmunity. Science. 1985;230:1043–1045. doi: 10.1126/science.2414848.
- Gerritse K, Laman JD, Noelle RJ, Aruffo A, Ledbetter JA, Boersma WJ, et al. CD40-CD40 ligand interactions in experimental allergic encephalomyelitis and multiple sclerosis. Proc Natl Acad Sci U S A (1996) 93(6):2499–504. 10.1073/pnas.93.6.2499

- Ghoussaini, M., Mountjoy, E., Carmona, M., Peat, G., Schmidt, E. M., Hercules, A., Fumis, L., Miranda, A., Carvalho-Silva, D., Buniello, A., Burdett, T., Hayhurst, J., Baker, J., Ferrer, J., Gonzalez-Uriarte, A., Jupp, S., Karim, M. A., Koscielny, G., MacHlitt-Northen, S., … Dunham, I. (2021a). Open Targets Genetics: Systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Research*, *49*(D1), D1311–D1320. https://doi.org/10.1093/nar/gkaa840
- Gibbs R. A., Belmont J. W., Hardenbol P., Willis T. D., Yu F., et al. , 2003. The international HapMap project. Nature 426(6968): 789–796
- Gold R, Linington C. Lassmann H. Understanding pathogenesis and therapy of multiple sclerosis via animal models: 70 years of merits and culprits in experimental autoimmune encephalomyelitis research. Brain. 2006;129:1953–1971. doi: 10.1093/brain/awl075
- Gold R. Wolinsky JS. Pathophysiology of multiple sclerosis and the place of teriflunomide. Acta Neurol. Scand. 2011;124:75–84. doi: 10.1111/j.1600-0404.2010.01444.x. Hemmer B, Archelos JJ. Hartung HP. New concepts in the immunopathogenesis of multiple sclerosis. Nat. Rev. Neurosci. 2002;3:291–301. doi: 10.1038/nrn784
- Guan, Y. & Stephens, M. Bayesian variable selection regression for genome-wide association studies, and other large-scale problems. *Ann. Appl. Statist.* 5, 1780–1815 (2011).
- Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, De Jager PL, et al. Risk alleles for multiple sclerosis identified by a genome wide study. N. Engl. J. Med. 2007;357:851–862. doi: 10.1056/NEJMoa073493
- Handel, A. E. et al. Smoking and multiple sclerosis: an updated meta-analysis. PLoS ONE 6, e16149 (2011).
- Harley JB, Chen X, Pujato M, Miller D, Maddox A, Forney C, Magnusen AF, Lynch A, Chetal K, Yukawa M, Barski A, Salomonis N, Kaufman KM, Kottyan LC, Weirauch MT. Transcription factors operate across disease loci, with EBNA2 implicated in autoimmunity. Nat Genet. 2018 May;50(5):699-707. doi: 10.1038/s41588-018-0102-3. Epub 2018 Apr 16. PMID: 29662164;
- Harp CT, Lovett-Racke AE, Racke MK, Frohman EM, Monson NL. Impact of myelin-specific antigen presenting B cells on T cell activation in multiple sclerosis. Clin Immunol (2008) 128(3):382–91. 10.1016/j.clim.2008.05.002
- Hartl D.L., Clark A.G. Sinauer Associates; Sunderland: 1997. Principles of population genetics.
- Hawkes, C. H. Smoking is a risk factor for multiple sclerosis: a metanalysis. Mult. Scler. 13, 610–615 (2007).

- Hedstrom, A. K., Lima Bomfim, I., Hillert, J., Olsson, T.& Alfredsson, L. Obesity interacts with infectious mononucleosis in risk of multiple sclerosis. Eur. J. Neurol. 22, 578–e538 (2015)
- Hedstrom, A. K. et al. Smoking and two human leukocyte antigen genes interact to increase the risk for multiple sclerosis. Brain 134, 653–664 (2011).
- Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. Identifying causal variants at loci with multiple signals of association. Genetics. 2014 Oct;198(2):497-508. doi: 10.1534/genetics.114.167908. Epub 2014 Aug 7. PMID: 25104515; PMCID: PMC4196608.
- International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. Science. 2019 Sep 27;365(6460):eaav7188. doi: 10.1126/science.aav7188. PMID: 31604244; PMCID: PMC7241648.
- Issazadeh S, Navikas V, Schaub M, Sayegh M, Khoury S. Kinetics of expression of costimulatory molecules and their ligands in murine relapsing experimental autoimmune encephalomyelitis in vivo. J Immunol (1998) 161(3):1104–12.
- Jacobson EM, Concepcion E, Oashi T, Tomer Y. A Graves' disease-associated Kozak sequence single-nucleotide polymorphism enhances the efficiency of CD40 gene translation: a case for translational pathophysiology. Endocrinology. 2005 Jun;146(6):2684-91. doi: 10.1210/en.2004-1617. Epub 2005 Feb 24. PMID: 15731360.
- Kahaly, G. J., Stan, M. N., Frommer, L., Gergely, P., Colin, L., Amer, A., Schuhmann, I., Espie, P., Rush, J. S., Basson, C., & He, Y. (2020). A Novel Anti-CD40 Monoclonal Antibody, Iscalimab, for Control of Graves Hyperthyroidism—A Proof-of-Concept Trial. The Journal of Clinical Endocrinology & Metabolism, 105(3), 696–704. https://doi.org/10.1210/CLINEM/DGZ013
- Kantarci OH. Genetics and natural history of multiple sclerosis. Semin. Neurol. 2008;28:7–16. doi: 10.1055/s-2007-1019125.
- Kibinge NK, Relton CL, Gaunt TR, Richardson TG. Characterizing the Causal Pathway for Genetic Variants Associated with Neurological Phenotypes Using Human Brain-Derived Proteome Data. Am J Hum Genet. 2020 Jun 4;106(6):885-892. doi: 10.1016/j.ajhg.2020.04.007. Epub 2020 May 14. PMID: 32413284; PMCID: PMC7273531.
- Kichaev G, Yang WY, Lindstrom S, Hormozdiari F, Eskin E, Price AL, et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. 2014;10:e1004722.

- Kircher M, Witten DM, Jain P, et al (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46:310–315. https://doi.org/10.1038/ng.2892
- Klareskog, L., Catrina, A. I., & Paget, S. (2009). Rheumatoid arthritis. *Lancet (London, England)*, *373*(9664), 659–672. https://doi.org/10.1016/S0140-6736(09)60008-8
- Kozak, M., 1987. An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res. 15, 8125–8148.
- Kozak, M., 1991a. An analysis of vertebrate mRNA sequences: intimations of translational control. J. Cell Biol. 115, 887–903.
- Larson MH, Gilbert LA, Wang X, et al (2013) CRISPR interference (CRISPRi) for sequence-specific control of gene expression. Nat Protoc 8:2180–2196. https://doi.org/10.1038/nprot.2013.132
- Lee BK, Bhinge AA, Iyer VR. Wide-ranging functions of E2F4 in transcriptional activation and repression revealed by genome-wide analysis. Nucleic Acids Res. 2011 May;39(9):3558-73. doi: 10.1093/nar/gkq1313. Epub 2011 Jan 18. PMID: 21247883; PMCID: PMC3089461.
- Lee PW, Severin ME, Lovett-Racke AE. TGF-β regulation of encephalitogenic and regulatory T cells in multiple sclerosis. Eur J Immunol. 2017 Mar;47(3):446-453. doi: 10.1002/eji.201646716. Epub 2017 Feb 10. PMID: 28102541; PMCID: PMC5499671.
- Levin, L. I., Munger, K. L., O'Reilly, E. J., Falk, K. I. & Ascherio, A. Primary infection with the Epstein-Barr virus and risk of multiple sclerosis. *Ann. Neurol.* 67, 824–830 (2010).
- Lewontin, R. C. & Kojima, K. The evolutionary dynamics of complex polymorphisms. *Evolution* 14, 458–472 (1960)
- Li G, Martínez-Bonet M, Wu D, Yang Y, Cui J, Nguyen HN, Cunin P, Levescot A, Bai M, Westra HJ, Okada Y, Brenner MB, Raychaudhuri S, Hendrickson EA, Maas RL, Nigrovic PA. High-throughput identification of noncoding functional SNPs via type IIS enzyme restriction. Nat Genet. 2018 Aug;50(8):1180-1188. doi: 10.1038/s41588-018-0159-z. Epub 2018 Jul 16. PMID: 30013183; PMCID: PMC6072570.
- Lin J, Zhou J, Xu Y. Potential drug targets for multiple sclerosis identified through Mendelian randomization analysis. Brain. 2023 Aug 1;146(8):3364-3372. doi: 10.1093/brain/awad070. PMID: 36864689; PMCID: PMC10393411.
- Long E, Yin J, Funderburk KM, et al (2022) Massively parallel reporter assays and variant scoring identified functional variants and target genes for melanoma loci and highlighted

cell-type specificity. Am J Hum Genet 109:2210– 2229. https://doi.org/10.1016/j.ajhg.2022.11.006

- Lublin FD, Reingold SC. Defining the clinical course of multiple sclerosis: results of an international survey. National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis. Neurology. 1996;46:907–911. doi: 10.1212/wnl.46.4.907
- Maglione A, Zuccalà M, Tosi M, Clerico M, Rolla S. Host Genetics and Gut Microbiome: Perspectives for Multiple Sclerosis. Genes (Basel). 2021 Jul 29;12(8):1181. doi: 10.3390/genes12081181. PMID: 34440354; PMCID: PMC8394267.
- Maller, J. B. et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* 44, 1294–1301 (2012).
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. Nature. 2009 Oct 8;461(7265):747-53. doi: 10.1038/nature08494. PMID: 19812666; PMCID: PMC2831613.
- Manuel AM, Dai Y, Jia P, Freeman LA, Zhao Z. A gene regulatory network approach harmonizes genetic and epigenetic signals and reveals repurposable drug candidates for multiple sclerosis. Hum Mol Genet. 2023 Mar 6;32(6):998-1009. doi: 10.1093/hmg/ddac265. PMID: 36282535; PMCID: PMC9991005.
- Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG Jr, Kinney JB, Kellis M, Lander ES, Mikkelsen TS. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol. 2012 Feb 26;30(3):271-7. doi: 10.1038/nbt.2137. PMID: 22371084; PMCID: PMC3297981.
- Montano, M. (2014). Model systems. Translational Biology in Medicine, 9–33. https://doi.org/10.1533/9781908818652.9
- Morris JA, Caragine C, Daniloski Z, et al (2023) Discovery of target genes and pathways at GWAS loci by pooled single- cell CRISPR screens. Science 380:eadh7699. https://doi.org/10.1126/science.adh7699
- Moyon L, Berthelot C, Louis A, et al (2022) Classification of non-coding variants with high pathogenic impact. PLOS Genet 18:e1010191. https://doi.org/10.1371/journal.pgen.101019

- Munger, K. L., Levin, L. I., Hollis, B. W., Howard, N. S. & Ascherio, A. Serum 25-hydroxyvitamin D levels and risk of multiple sclerosis. JAMA 296, 2832–2838 (2006).
- Myint L, Avramopoulos DG, Goff LA, Hansen KD (2019) Linear models enable powerful differential activity analysis in massively parallel reporter assays. BMC Genomics 20:209. https://doi.org/10.1186/s12864-019-5556-x
- Nielsen, T. R. et al. Effects of infectious mononucleosis and HLA-DRB1*15 in multiple sclerosis. Mult. Scler. 15, 431–436 (2009).
- O'Connor KC, Bar-Or A. Hafler DA. The neuroimmunology of multiple sclerosis: possible roles of T and B lymphocytes in immunopathogenesis. J. Clin. Immunol. 2001;21:81–92. doi: 10.1023/a:1011064007686
- Oksenberg JR, Seboun E, Hauser SL (1996) Genetics of Demyelinating Diseases. Brain Pathol 6:289–302. https://doi.org/10.1111/j.1750-3639.1996.tb00856.x
- Olsson T, Barcellos LF, Alfredsson L. Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis. Nat Rev Neurol. 2017 Jan;13(1):25-36. doi: 10.1038/nrneurol.2016.187. Epub 2016 Dec 9. PMID: 27934854.
- Orrù V, Steri M, Sidore C, Marongiu M, Serra V, Olla S, Sole G, Lai S, Dei M, Mulas A, Virdis F, Piras MG, Lobina M, Marongiu M, Pitzalis M, Deidda F, Loizedda A, Onano S, Zoledziewska M, Sawcer S, Devoto M, Gorospe M, Abecasis GR, Floris M, Pala M, Schlessinger D, Fiorillo E, Cucca F. Complex genetic signatures in immune cells underlie autoimmunity and inform therapy. Nat Genet. 2020 Oct;52(10):1036-1045. doi: 10.1038/s41588-020-0684-4. Epub 2020 Sep 14. Erratum in: Nat Genet. 2020 Nov;52(11):1266. doi: 10.1038/s41588-020-00718-6. PMID: 32929287; PMCID: PMC8517961.
- Pei Y, Banerjee S, Sun Z, Jha HC, Saha A, Robertson ES. EBV Nuclear Antigen 3C Mediates Regulation of E2F6 to Inhibit E2F1 Transcription and Promote Cell Proliferation. PLoS Pathog. 2016 Aug 22;12(8):e1005844. doi: 10.1371/journal.ppat.1005844. PMID: 27548379; PMCID: PMC4993364.
- Rang, X., Liu, Y., Wang, J., Wang, Y., Xu, C., & Fu, J. (2022). Identification of multiple sclerosis-related genes regulated by EBV-encoded microRNAs in B cells. *Multiple Sclerosis and Related Disorders,59*. https://doi.org/10.1016/j.msard.2022.103563
- Ransohoff RM. Natalizumab for multiple sclerosis. N. Engl. J. Med. 2007;356:2622–2629. doi: 10.1056/NEJMct071462
- Raychaudhuri S, Remmers EF, Lee AT, Hackett R, Guiducci C, Burtt NP, Gianniny L, Korman BD, Padyukov L, Kurreeman FA, Chang M, Catanese JJ, Ding B, Wong S, van der

Helm-van Mil AH, Neale BM, Coblyn J, Cui J, Tak PP, Wolbink GJ, Crusius JB, van der Horst-Bruinsma IE, Criswell LA, Amos CI, Seldin MF, Kastner DL, Ardlie KG, Alfredsson L, Costenbader KH, Altshuler D, Huizinga TW, Shadick NA, Weinblatt ME, de Vries N, Worthington J, Seielstad M, Toes RE, Karlson EW, Begovich AB, Klareskog L, Gregersen PK, Daly MJ, Plenge RM. Common variants at CD40 and other loci confer risk of rheumatoid arthritis. Nat Genet. 2008 Oct;40(10):1216-23. doi: 10.1038/ng.233. Epub 2008 Sep 14. PMID: 18794853; PMCID: PMC2757650.

- Ristov, J., Espie, P., Ulrich, P., Sickert, D., Flandre, T., Dimitrova, M., Müller-Ristig, D., Weider, D., Robert, G., Schmutz, P., Greutmann, B., Cordoba-Castro, F., Schneider, M. A., Warncke, M., Kolbinger, F., Cote, S., Heusser, C., Bruns, C., & Rush, J. S. (2018). Characterization of the in vitro and in vivo properties of CFZ533, a blocking and non-depleting anti-CD40 monoclonal antibody. American Journal of Transplantation, 18(12), 2895–2904. https://doi.org/10.1111/ajt.14872
- Ritchie GRS, Dunham I, Zeggini E, Flicek P (2014) Functional annotation of noncoding sequence variants. Nat Methods 11:294–296. https://doi.org/10.1038/nmeth.2832
- Rudick RA, Cohen JA, Weinstock-Guttman B, Kinkel RP. Ransohoff RM. Management of multiple sclerosis. N. Engl. J. Med. 1997;337:1604–1611. doi: 10.1056/NEJM199711273372207
- Sadovnick AD, Armstrong H, Rice GPA, et al (2004) A population-based study of multiple sclerosis in twins: Update. Ann Neurol 33:281–285.
- Sandberg, L. et al. Vitamin D and axonal injury in multiple sclerosis. Mult. Scler. 22, 1027–1031 (2015)
- Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, Moutsianas L, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature. 2011;476:214–219. doi: 10.1038/nature10251.
- Sawcer, S. et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476, 214–219 (2011).
- Schaid, D.J., Chen, W. & Larson, N.B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* 19, 491–504 (2018). https://doi.org/10.1038/s41576-018-0016-
- Selter CR. Hemmer B. Update on immunopathogenesis and immunotherapy in multiple sclerosis. Immuno. Targets Therapy. 2013;2:21–30. doi: 10.2147/ITT.S31813.

- Sesia M, Katsevich E, Bates S, Candès E, Sabatti C. Multi-resolution localization of causal variants across the genome. Nat Commun. 2020 Feb 27;11(1):1093. doi: 10.1038/s41467-020-14791-2
- Shin B-Y, Lee S-H, Kim Y, et al (2022) Interactomic inhibition of Eomes in the nucleus alleviates EAE via blocking the conversion of Th17 cells into non-classic Th1 cells. Immunol Med 45:119–127. https://doi.org/10.1080/25785826.2022.2031812
- Simon G. Coetzee, Gerhard A. Coetzee, Dennis J. Hazelett, *motifbreakR*: an R/Bioconductor package for predicting variant effects at transcription factor binding sites, *Bioinformatics*, Volume 31, Issue 23, December 2015, Pages 3847–3849, https://doi.org/10.1093/bioinformatics/btv470
- Skibola, C.F., et al., 2008. A functional TNFRSF5 gene variant is associated with risk of lymphoma. Blood 111, 4348–4354.
- Sobel R. Moore W. Vol. 2. London: UK Oxford Univ. Press; 2008. pp. 1513–1608. Demyelinating diseases. Greenfield's neuropathology
- Sokolova EA, Malkova NA, Korobko DS, Rozhdestvenskii AS, Kakulya AV, Khanokh EV, Delov RA, Platonov FA, Popova TY, Aref' eva EG, Zagorskaya NN, Alifirova VM, Titova MA, Smagina IV, El' chaninova SA, Popovtseva AV, Puzyrev VP, Kulakova OG, Tsareva EY, Favorova OO, Shchur SG, Lashch NY, Popova NF, Popova EV, Gusev EI, Boyko AN, Aulchenko YS, Filipenko ML. Association of SNPs of CD40 gene with multiple sclerosis in Russians. PLoS One. 2013 Apr 22;8(4):e61032. doi: 10.1371/journal.pone.0061032. PMID: 23613777; PMCID: PMC3632563.
- Storkebaum E, Carmeliet P. VEGF: a critical player in neurodegeneration. J Clin Invest. 2004 Jan;113(1):14-8. doi: 10.1172/JCI20682. PMID: 14702101; PMCID: PMC300888.
- Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, Andersen KG, Mikkelsen TS, Lander ES, Schaffner SF, Sabeti PC. Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. Cell. 2016 Jun 2;165(6):1519-1529. doi: 10.1016/j.cell.2016.04.027. Erratum in: Cell. 2018 Feb 22;172(5):1132-1134. doi: 10.1016/j.cell.2018.02.021. PMID: 27259153; PMCID: PMC4957403.
- Thierry, Flandre., Keith, Mansfield., Pascal, Espié., Tina, Rubic-Schneider., Peter, Ulrich. (2022). Immunosuppression Profile of CFZ533 (Iscalimab), a Non-Depleting Anti-CD40 Antibody, and the Presence of Opportunistic Infections in a Rhesus Monkey Toxicology Study. Toxicologic Pathology, 50(5):712-724. doi: 10.1177/01926233221100168

- Vaitaitis GM, Yussman MG, Waid DM, Wagner DH, Jr. Th40 cells (CD4+CD40+ Tcells) drive a more severe form of experimental autoimmune encephalomyelitis than conventional CD4 T cells. PLoS One (2017) 12(2):e0172037. 10.1371/journal.pone.0172037
- Van de Schoot, R., Depaoli, S., King, R. *et al.* Bayesian statistics and modelling. *Nat Rev Methods Primers* 1, 1 (2021). https://doi.org/10.1038/s43586-020-00001-2
- Vogel DY, Heijnen PD, Breur M, de Vries HE, Tool AT, Amor S, et al. Macrophages migrate in an activation-dependent manner to chemokines involved in neuroinflammation. J Neuroinflammation (2014) 11:23. 10.1186/1742-2094-11-23
- Weinshenker BG, Bass B, Rice GP, Noseworthy J, Carriere W, Baskerville J, et al. The natural history of multiple sclerosis: a geographically based study. I. Clinical course and disability. Brain. 1989;112(Pt 1):133–146. doi: 10.1093/brain/112.1.133
- Weinstock-Guttman B, Ransohoff RM, Kinkel RP. Rudick RA. The interferons: biological effects, mechanisms of action, and use in multiple sclerosis. Ann. Neurol. 1995;37:7–15. doi: 10.1002/ana.410370105
- Wellcome Trust Case Control Consortium. Genomewide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–678 (2007).
- Wesnes, K. *et al.* Body size and the risk of multiple sclerosis in Norway and Italy: the EnvIMS study. *Mult. Scler.* 21, 388–395 (2015)
- Wray N.R. Allele frequencies and the r2 measure of linkage disequilibrium: impact on design and interpretation of association studies. Twin Res. Hum. Genet. 2005;8:87–94. doi: 10.1375/1832427053738827.
- Wucherpfennig KW. Strominger JL. Molecular mimicry in T cell-mediated autoimmunity: viral peptides activate human T cell clones specific for myelin basic protein. Cell. 1995;80:695–705. doi: 10.1016/0092-8674(95)90348-8
- Yin, Y., Butler, C., & Zhang, Q. (2021). Challenges in the application of NGS in the clinical laboratory. Human Immunology, 82(11),812-819.
- Yong VW. Immunopathogenesis of multiple sclerosis. Contin. Lifelong Learn. Neurol. 2004;10:11–27
- Yousry TA, Major EO, Ryschkewitsch C, Fahle G, Fischer S, Hou J, et al. Evaluation of patients treated with natalizumab for progressive multifocal leukoencephalopathy. N. Engl. J. Med. 2006;354:924–933. doi: 10.1056/NEJMoa054693.
- Zhang ZG, Zhang L, Jiang Q, Zhang R, Davies K, Powers C, Bruggen Nv, Chopp M. VEGF enhances angiogenesis and promotes blood-brain barrier leakage in the ischemic brain. J Clin

Invest. 2000 Oct;106(7):829-38. doi: 10.1172/JCI9369. PMID: 11018070; PMCID: PMC517814.

- Zhou, Y. et al. Genetic loci for Epstein-Barr virus nuclear antigen-1 are associated with risk of multiple sclerosis. Mult. Scler. (2016).
- Zhu LH, Gao S, Jin R, Zhuang LL, Jiang L, Qiu LZ, Xu HG, Zhou GP. Repression of interferon regulatory factor 3 by the Epstein-Barr virus immediate-early protein Rta is mediated through E2F1 in HeLa cells. Mol Med Rep. 2014 Apr;9(4):1453-9. doi: 10.3892/mmr.2014.1957. Epub 2014 Feb 17. PMID: 24535579.
- Zuccalà M, Barizzone N, Boggio E, Gigliotti L, Sorosina M, Basagni C, Bordoni R, Clarelli F, Anand S, Mangano E, Vecchio D, Corsetti E, Martire S, Perga S, Ferrante D, Gajofatto A, Ivashynka A, Solaro C, Cantello R, Martinelli V, Comi G, Filippi M, Esposito F, Leone M, De Bellis G, Dianzani U, Martinelli-Boneschi F, D'Alfonso S. Genomic and functional evaluation of TNFSF14 in multiple sclerosis susceptibility. J Genet Genomics. 2021 Jun 20;48(6):497-507. doi: 10.1016/j.jgg.2021.03.017. Epub 2021 May 25. PMID: 34353742.

Publications

- **Higher prevalence of autoimmune comorbidities in multiple sclerosis from a population-based study with genetic linkage** Roberto Gnavi, Nadia Barizzone, Roberta Picariello, Paolo Emilio Alboini, Nicola Pomella, Martina Tosi, Endri Visha, Valentina Ciampana, Domizia Vecchio, Paola Cavalla, Maurizio Leone*, Sandra D'Alfonso*. Multiple Sclerosis Journal (Accepted)