GWAS-associated Variants, Non-genetic Factors and Transient Transcriptome in Multiple Sclerosis Etiopathogenesis: a Colocalization Analysis

Renato Umeton,[1,2,3,4,*] Gianmarco Bellucci,[5] Rachele Bigi,[5] Silvia Romano,[5] Maria Chiara Buscarinu,[5] Roberta Reniè,[5] Virginia Rinaldi,[5] Raffaella Pizzolato Umeton,[6,7,8,9] Emanuele Morena.[5] Carmela Romano,[5] Rosella Mechelli,[10,11] Marco Salvetti[5,12,*], Giovanni Ristori[5,13,*]

1. Department of Informatics and Analytics, Dana-Farber Cancer Institute, Boston, MA, USA
2. Massachusetts Institute of Technology, Cambridge, MA, USA
3. Harvard School of Public Health, Boston, MA, USA
4. Weill Cornell Medicine, New York City, NY, USA
5. Centre for Experimental Neurological Therapies (CENTERS), Department of Neurosciences, Mental Health and Sensory Organs, Sapienza University of Rome, Italy
6. Department of Neurology, UMass Memorial Health Care, Worcester, MA, USA
7. University of Massachusetts Medical School, Worcester, MA, USA
8. Department of Neurology, Massachusetts General Hospital, Boston, MA, USA
9. Harvard Medical School, Boston, MA, USA
10. IRCCS San Raffaele Pisana, Rome, Italy
11. San Raffaele Roma Open University, Rome, Italy
12. IRCCS Istituto Neurologico Mediterraneo Neuromed, Pozzilli, Italy
13. Neuroimmunology Unit, IRCCS Fondazione Santa Lucia, Rome, Italy

*Corresponding authors

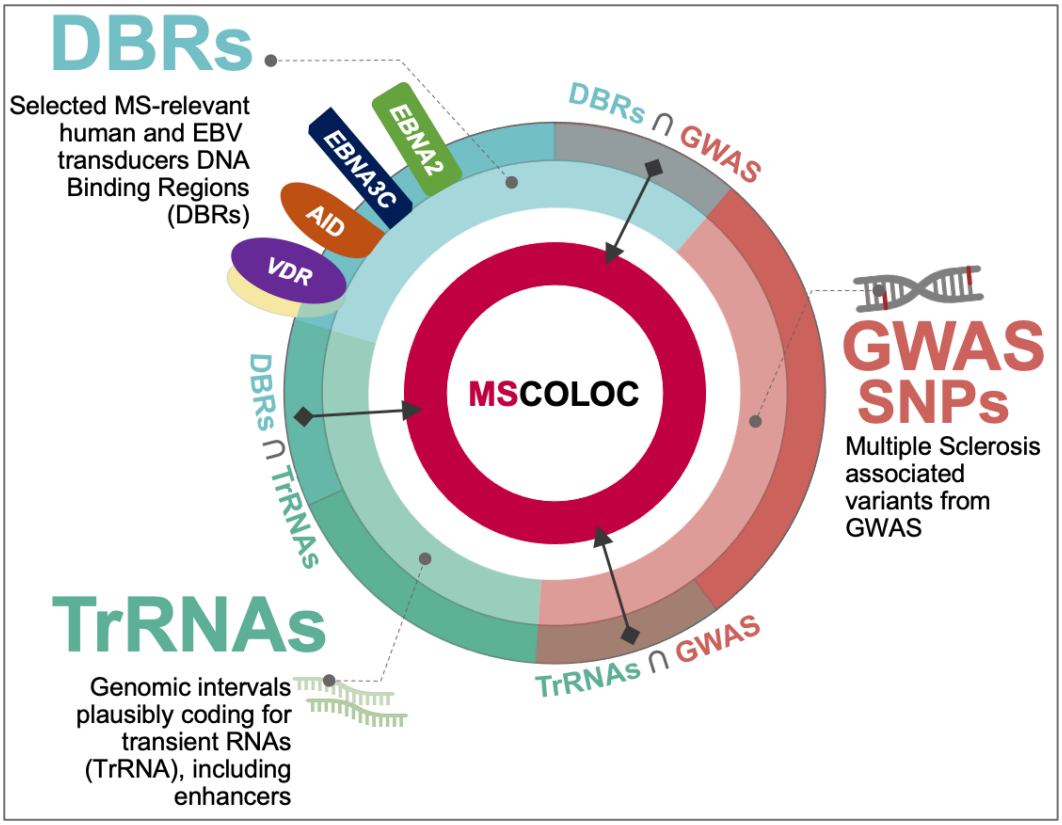
giovanni.ristori@uniroma1.it

ABSTRACT

A clinically actionable understanding of multiple sclerosis (MS) etiology goes through GWAS interpretation, prompting research on new gene regulatory models. Our previous works on these topics suggested a stochastic etiologic model where small-scale random perturbations could eventually reach a threshold for MS onset and progression. A new sequencing technology has mapped the transient transcriptome (TT), including intergenic RNAs, and antisense intronic RNAs. Through a rigorous colocalization analysis, here we show that genomic regions coding for the TT were significantly enriched for both MS-associated GWAS variants, and DNA binding sites for molecular 'transducers' mediating putative, non-genetic, etiopathogenetic factors for MS (e.g., vitamin D deficiency, Epstein Barr virus latent infection, B cell dysfunction).

These results suggest a model whereby TT-coding regions are "hotspots" of convergence between genetic ad non-genetic factors of risk/protection for MS (and plausibly for other complex disorders). Our colocalization analysis also provided a freely available data web interface at www.mscoloc.com for future research on transcriptional regulation in MS.

GRAPHICAL ABSTRACT



INTRODUCTION

A large body of literature agrees that regulatory genomic intervals, especially those encompassing enhancers, are enriched with disease-associated DNA elements. Most of this evidence comes from genome wide association studies (GWAS) based on single polymorphism nucleotides (SNPs) representing common variants,[1-5] even though a recent study showed that low-frequency and rare coding variants may somewhat contribute to MS heritability.[6] Several characteristics of regulatory disease-associated genetic variants complicates GWAS interpretation, prompting research on new gene regulatory models: (*i*) SNPs are chosen as haplotypes to spare the genotyping work needed for the large number of samples used in GWAS, therefore fine mapping and epigenetic studies are required to integrate GWAS data;[7-10] (*ii*) only a small fraction of supposedly causal disease-associated variants directly alters recognizable transcription factor binding motifs as it might be expected, according to their regulatory function;[4] *(iii*) the identified GWAS signals are likely to exert highly contextual (i.e., time- and position-dependent) regulatory effects, that may change according to the tissue and to the time when they receive an input from inside or outside the cell. In summary, current gene regulatory models help only in part to fully detail which disease-associated SNP signals are causal, and by which exact mechanisms they are causal. Recent studies on the biological spectrum of human DNase I hypersensitive sites (DHSs), that are disease-associated markers of regulatory DNA, may help to better rework GWAS data and particularly to contextualize the genomic variants according to tissue/cell states and to gene body colocalization of DHSs.[11] In this context, the latest version of the Genotype-Tissue Expression project may provide further insights into the tissue specificity of genetic effects, supporting the link between regulatory mechanisms and traits or complex diseases.[12]

Another layer of complexity is revealed by our recent studies suggesting an MS etiologic model where stochastic phenomena (i.e., random events not necessarily resulting in disease in all individuals) may contribute to the disease onset and progression. This model, embedded between physics of stochastic systems and cell biology, suggests how small-scale random perturbations would impact on large-scale phenomena, such as disease development. Such perturbations would allow to exceed the onset threshold that is believed to be set by genetic and non-genetic

susceptibility factors such as microbial or geographical characteristics.[13, 14] Such model is consistent with our previous results that showed heterogeneity in the MS etiology components in twin pairs studies.[15-17] This is also in line with prior bioinformatics analyses that determined a significant enrichment of binding motifs for Epstein-Barr virus (EBV) nuclear antigen 2 (EBNA2) and vitamin D receptor (VDR) in genomic regions containing MS-associated GWAS variants.[18] We also demonstrated that genomic variants of *EBNA2* resulted to be MS-associated,[19] and other groups have further developed our findings showing that enrichment of EBNA2-binding regions on GWAS DNA intervals is involved in the pathogenesis of autoimmune disorders, including MS.[20]

A recent sequencing innovation (called TT-seq) allowed to map the transient transcriptome that has a typical half-life within minutes, compared to stable RNA elements, such as protein-coding mRNAs, long-noncoding RNAs, and micro-RNAs, that persists at least a few hours.[21-23] The transient transcriptome (TT) includes mostly enhancer RNAs (eRNA), short intergenic non-coding RNAs (sincRNA; i.e. promoter-associated RNAs within 10 kb of a GENCODE mRNA transcription start site), and antisense RNAs (asRNA). Such TT elements present specific features, some similar to and some contrasting those of stable RNAs. More in detail, transient RNAs (trRNA) are relatively short in length, they generally lack a secondary structure, and would not present those chemical modifications that characterize unidirectional and polyadenylated stable RNAs.[21, 24, 25]

Other recent works based on time-resolved analysis, agree on the eRNAs very rapid functional dynamics model while interacting with the transcriptional co-activator acetyltransferase CBP/p300 complex.[26, 27] This confirms the highly contextual role of eRNAs through the control of transcription burst frequencies, which are known to influence cell-type-specific gene expression profiles.[28] Along these lines, a recent study showed that T cells selectively filter oscillatory signals within the minute timescale,[29] further supporting the aforementioned model.

In summary, on the basis of our previous research (i.e., the heterogeneity in the MS etiology components; the stochasticity in the interaction between genetic and non-genetic factors contributing to disease development; the enrichment of binding sites for environmental factors in MS-associated DNA intervals) and leveraging the recent sequencing innovations in the mapping of the transient transcriptome (i.e., the erratic time dynamics and the highly contextual expression),[21,

[22] we hypothesize that MS-associated GWAS signals prevalently fall within regulatory regions of DNA coding for transient RNAs (trRNA). In theory, the genomic intervals coding for the transient transcriptome may be the hotspots where temporo-spatial occurrences (stochastic in nature, as said) may coalesce and so contribute to physiological outcomes (that similarly present as developmental and/or adaptive), or possibly give rise to MS onset or progression. This study is aimed at verifying this working hypothesis through a colocalization analysis and its further dissection.

RESULTS

MS GWAS and trRNAs - MS-associated GWAS signals colocalize with regulatory regions of DNA plausibly coding for trRNAs

We set up our region-of-interest (ROI) inside GWAS catalogue[30] by considering all MS GWAS that were published, extracting all SNP positions, and creating a single set of genomic coordinates that therefore encompass all GWAS-derived or GWAS-verified signal for MS. We then refined the SNP list by pruning out about 1.5% of the SNPs as they did not contain intelligible genomic annotations or were duplicates. The final ROI list is reported in Supplementary Table 1 and consists of 603 unique single-nucleotide regions.

Next, we matched through colocalization analyses our ROI with lists of regions resulting from the work by Michel et al., which mapped the transient and stable transcriptome captured by TT-seq after T cell stimulation.[21] We found a significant enrichment of MS-associated genetic variants in the transient transcriptome (Tab. 1). Of note, when we split the transcriptome list in two subsets for long (≥ 60 minutes) and short (< 60 minutes) half-life, we found that only the short half-life subset significantly colocalized with the ROI. This finding was indicative of the relationship between MS-associated GWAS signals and the regulatory regions of DNA coding for trRNA.

When we further dissected the mapping of the ROI colocalization signals, we found a significant excess of intergenic and intron regions (as anticipated), as well as their prevalent distribution away from the transcription start site (TSS; Fig. 1A). Notably, when we extended this analysis to GWAS data coming from other multifactorial diseases or traits, dividing immune-mediated and other

complex conditions, we found highly comparable profiles (Fig. 1B, C; Supplementary Table 2). This suggests that the colocalization between MS-associated DNA intervals and intergenic or intronic sequences, plausibly referring to tr-RNA coding regions, is shared by the genetic architecture of most multifactorial disorders.

To consolidate this result and gain a deeper biological insight, we extended the colocalization analysis matching the ROI with a vast set of databases of regulatory DNA regions, including enhancers and super-enhancers, derived from experiments on diverse tissue types (references in Supplementary Table 3). To improve interpretability of the results through ranking, we implemented a harmonic score, based on the Odd Ratio, the -log(p-value), and the support of each match. Statistically significant results came from sets included in SEA, seDB, dbSuper and other single lists of enhancers and non-coding RNAs (Fig. 2). Interestingly, we found a strong enrichment of MS-associated genetic variants in cell lines of hematopoietic lineage, including CD19+ and CD20+ B lymphocytes, CD4+ T helper cells, and CD14+ monocytes. Moreover, among the top-scoring hits, we found microglial-specific enhancers, which is in line with recent reports on brain cell type-specific enhancer–promoter interactome activities and the latest GWAS on MS genomic mapping[31, 32]. On the other hand, non-relevant tissues serving as controls (such as kidney, muscle, glands, etc.) scored low in the ranking, crowding the bottom-left corner of Figure 2. Given the cell-specificity of the action of regions encoding for eRNA and other non-coding RNAs, these results are consistent with a functional relevance of GWAS variants in contributing to MS disease mechanisms, by modulating transcription regulation. The whole list of results derived from ROI and databases matches is freely available as a data resource at www.mscoloc.com to support future research on the actors of transcription regulation in MS.

trRNAs and DBRs - Genetic and non-genetic factors for MS etiology converge in genomic regions plausibly coding for the transient transcriptome

Independent studies support the fact that MS GWAS intervals are enriched with DNA binding regions (DBRs) for protein ‘transducers’ mediating non-genetic factors of putative etiologic

relevance in MS, such as vitamin D deficiency or EBV latent infection.[18, 20] Therefore we further inquired whether DNA regions plausibly coding for trRNA share these features (i.e., they colocalize with such DRBs). We set up the DBRs for vitamin D receptor (VDR), activation-induced cytidine deaminase (AID), Epstein Barr nuclear antigen 2 (EBNA2), Epstein Barr nuclear antigen 3 (EBNA3C), chosen among viral or host's nuclear factors potentially associated to MS etiopathogenesis.[33-36] The DBRs for each nuclear factor were derived from recent literature (Supplementary Table 4) and matched with the GWAS-derived MS signals to confirm and expand previous results. We found statistically significant results for VDR, EBNA2, and AID for all the SNP position extensions (±50, 100, 200 kb), while for EBNA3C significant results came out at extension of ±100 and 200 kilobases. This finding suggests that several DBRs are capable of impacting on the MS-associated DNA intervals through colocalization (Tab. 2). Notably, when we searched for MS-associated regions shared by the DBRs analyzed, we were able to prioritize 275 genomic regions capable of binding at least 2 molecular transducers which represent almost half of the MS-associated GWAS SNPs. These regions are 'hotspots' of interactions between genetic and non-genetic modifier of MS risk/protection. Moreover, all four proteins (VDR, AID, EBNA2, EBNA3C) proved to target 24 regions, 3 of them 115 regions, and 2 of them 136 regions (Fig. 3). This data is available at www.mscoloc.com, detailing the genomic regions targeted by each of the four molecular transducers.

Building once again on the work by Michel et al.[22], we inquired whether there was a colocalization between genomic regions containing MS-associated variants, DBRs for VDR, EBNA2, EBNA3C, AID, and DNA intervals plausibly coding for trRNA. To this end, we considered the transient transcriptome that proved to be enriched with MS-associated variants (Tab. 1), and we then matched the corresponding coding regions with the DBRs for the four molecular transducers (for this analysis DBRs for EBNA2, EBNA3C, AID and VDR represented the ROI, while the ENCODE database of Transcription Factors Binding Sites served as control; Fig. 4). We report the results of this analysis in table 3, which shows the colocalization between DNA regions plausibly coding for trRNA of both MS-relevant GWAS signals, and DBR of 3 out of 4 factors active at nuclear level,

and potentially associated to the disease. The fourth DBR, EBNA3C, did not reach statistical significance, though it showed higher values of support for short half-life transcripts.

DBRs and MS GWAS - Genomic regions plausibly coding for the transient transcriptome colocalize with MS-associated GWAS signals

To review and confirm previous colocalizations, we considered the genomic regions resulting from the above reported match between the MS-associated GWAS intervals and the databases of regulatory DNA regions, containing enhancers and super-enhancers, plausibly enriched in trRNA-coding sequences (see results in Fig. 2 and the web resource). We therefore matched these DNA regions with the DBR for VDR, EBNA2, EBNA3C and AID, finding significant enrichments that allow to contextualize and prioritize genomic positions, cell/tissue identity or cell status associated to MS. Considering the harmonic score obtained from these colocalization analyses, the top hits in EBNA2, EBNA3C, and AID involved lymphoid (CD19+ B cell lines and lymphomas, T regulatory cells, tonsils) and monocyte-macrophage lineages (peripheral macrophages, dendritic cells) from experiments included in the ENCODE, dbsuper, roadmapEpigenomics databases (see also Supplementary Table 5). Even though immune cells prevailed also in VDR top hits, a less stringent polarization was seen, somehow reflecting the wide-spreading actions of this transducer in human biology. However, with a more stringent cutoff of harmonic score >40 that selects the most significant hits (Supplementary Fig.1), a core subset of MS-relevant cell lineages, shared across all four examined transducers, became evident (Supplementary Table 6). These data and results are available in full at www.mscoloc.com.

DISCUSSION

Our study supports the hypothesis that investigations on the transient transcriptome may contribute to clarify how the GWAS signals affect the etiopathogenesis of MS and possibly of other complex disorders. Specifically, we show that genomic regions coding for the transient transcriptome recently described in T cells,[22] are significantly enriched for both MS-associated GWAS variants, as well as for DNA binding sites for protein 'transducers' of non-genetic signals, chosen among those plausibly associated to MS. The colocalization of GWAS intervals and some DNA-binding factors involved in MS etiology has already been reported,[18, 20] and here we reinforce this premise further suggesting a model in which trRNA-coding regions are hotspots of convergence between genetic ad non-genetic factors of risk/protection for MS. Our analysis showed that these hotspots are shared by two or more of the four nuclear factors, that we arbitrarily chose, indicating possible additive pathogenic effects for MS development (see Fig. 3 and Supplementary Table 6).

In homeostatic conditions, it can be hypothesized that DNA sequences coding for trRNA are composed of regulatory regions where genetic variability and non-genetic signals interact to finely regulate the gene expression according to cell identity, developmental or adaptive states, and time-dependent stimuli. As a matter of fact, the high sequence variability of these regions and the strict time-dependence of their transcription could be instrumental to adaptative features; however, these same features make these regions susceptible to become dysfunctional or to be the targets of pathogenic interaction. In some instances, these detrimental interactions come from outside the cell, such as in the case of EBV interference with host transcription,[35, 42] and the pathogenic consequences of vitamin D deficiency; in other cases, the dysfunction develops within the cell, such as the tumorigenic activity of AID in B cells.[43, 44]

The mapping of transient transcripts by TT-seq approach fits very well with our results obtained from GWAS data for MS and other multifactorial conditions, showing a significant excess of intergenic and intronic regions (coding for eRNA, sincRNA, and asRNA), and having a distribution in DNA intervals mostly far off from transcription start sites (TSS; see Fig. 1). This is in agreement with recent evidence of regulatory DNA region markers which contains genetic variants for

complex disease or traits; indeed, a systematic framework of common coordinates for these markers showed that about half of them lie within introns and most localize away from the TSS.[11] To further support the relationship between trRNA and transcription of regulatory DNA regions, we matched a large dataset of enhancers and super-enhancers with MS-GWAS signals and DBR for VDR, EBNA2, EBNA3C and AID. The significant enrichment in cell lines and cell status coming from the hematopoietic lineages and the CNS-specific cell subsets corroborates data coming from recent reports such as the relevance of contextualizing and prioritizing the role of MS-associated GWAS signals.[31, 32, 45, 46] Our analysis supports the pivotal regulatory role of enhancer transcription (i.e., a main component of transient transcriptome) was recently reported as not dispensable for gene expression at the immunoglobulin locus and for antibody class switch recombination,[47] though more research is needed to unravel such topic at a finer grain.

Reports on dynamics of time-course data are a recent area of focus within the analysis of gene expression, specifically in immune cells. Although current studies use methods that investigate time points related to the stable transcriptome (RNA-seq performed with time spans of hours), they clearly show that gene expression dynamics may influence allele specificity, regulatory programs that seem to depend on autoimmune disease-associated loci, and different transcriptional profiles based on cell status after stimulation.[48] A recent work showed that an *IL2ra* enhancer, which harbors autoimmunity risk variants and was one of the first MS-associated loci from GWAS, has no impact on the gene level expression. In times of gene activation, the same *IL2ra* enhancer would be delaying transcription in response to extracellular stimuli.[7] The importance of the timing in the gene expression control emerges also from several studies implicating enhancers and super-enhancers in the process of phase separation and formation of condensates. In this context, the transcriptional apparatus steps-up to drive robust genic responses.[49-52] The overall process seems to be highly dynamic, with time spans of seconds or minutes, and hence compatible with the temporal features of the transient transcriptome, which could somehow act upstream for the formation of these phase-separated condensates.

We suggest that studies on transient transcriptomes may integrate previous RNA-seq data in accounting for the interplay between genetic variability and non-genetic etiologic factors leading to

MS development. Components of a more-complex-than-anticipated regulation of gene expression could include transcriptional noise, transitory time-courses, erratic dynamics, and highly flexibility of some DNA regions, possibly oscillating between bistable states of enhancer and silencer.[53] The availability of tools to map trRNA could further contribute to the development of studies on immune cells isolated from patients and matched controls, aimed at dissecting key aspects of the complex transcriptional response in MS. Our analysis provides a platform for future studies on transient transcriptome, which we support by making our data resource available at www.mscoloc.com. Moreover, new gene regulatory models may emerge from this approach in order to better evaluate the meaning of GWAS in complex traits and the impact of the enhancer transcription,[47] which was recently reported as an ancient and conserved, yet flexible, genomic regulatory syntax.[54]

DATA ANALYSIS

Analyses were performed in Python and R. A data freeze was applied on 3/1/2020. All GWAS data was gathered from the GWAS Catalog through its REST API;[30, 37] about 1.5% of this data was filtered out as part of a QC process aimed at homogenizing legacy and more recent data. The MS GWAS regions were extracted from the overall GWAS Catalog data filtering by trait EFO_0003885. All Transcription Factor Binding Site regions (TFBS) were obtained from the ENCODE portal.[38] All data was organized in various databases and data pipelines as detailed below. For SNP overlaps and region colocalization, we used LOLA[39] and Fisher's exact test with False Discovery Rate (Benjamini-Hochberg) to control for multiple testing; this was containerized and modularized to work on a parallel cluster environment. Resulting -log(p-value), support, and Odds Ratio (OR) were combined into a single score inspired by the harmonic mean[40] and multi-objective optimization[41] with the formula below, where the spacing parameter k_p was set to 10.0 and we consider all three contributors equally, setting therefore weights w_i to 1.0. Statistical significance was taken at $p<0.05$. A modular and parallel data pipeline was created to: (i) readily generate and evaluate all experiments in the paper, (ii) manage and organize all data coming from various region databases (10,000+ regions), multiple ROIs (MS GWAS, EBNA2, EBNA3C, VDR, AID, etc.), databases of vast background regions as they were populated with the data obtained from GWAS Catalog, ENCODE, and other data resources, (iii) provide overlaps and intersection among various data elements, annotate them with the original MS GWAS loci that generated the signal, and (iv) generate the overarching data resource available at www.mscoloc.com.

$$Harmonic_{Score} = k_p * \frac{\sum_i w_i}{\frac{w_1}{logP} + \frac{w_2}{Supp} + \frac{w_3}{OR}}$$

Funding

GR is supported by “Progetti Grandi Ateneo” 2020, Sapienza University of Rome. MS and GR are supported by CENTERS, a Special Project of, and financed by, FISM - Fondazione Italiana Sclerosi Multipla.

References

1. Ernst J, Kheradpour P, Mikkelsen TS, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*. May 2011;473(7345):43-9. doi:10.1038/nature09906
2. Maurano MT, Humbert R, Rynes E, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. Sep 2012;337(6099):1190-5. doi:10.1126/science.1222794
3. Gusev A, Lee SH, Trynka G, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am J Hum Genet*. Nov 2014;95(5):535-52. doi:10.1016/j.ajhg.2014.10.004
4. Farh KK, Marson A, Zhu J, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*. Feb 2015;518(7539):337-43. doi:10.1038/nature13835
5. Vahedi G, Kanno Y, Furumoto Y, et al. Super-enhancers delineate disease-associated regulatory nodes in T cells. *Nature*. Apr 2015;520(7548):558-62. doi:10.1038/nature14154
6. chris.cotsapas@yale.edu IMSGCEa, Consortium IMSG. Low-Frequency and Rare-Coding Variation Contributes to Multiple Sclerosis Risk. *Cell*. Jan 2020;180(2):403. doi:10.1016/j.cell.2020.01.002
7. Mumbach MR, Satpathy AT, Boyle EA, et al. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat Genet*. Nov 2017;49(11):1602-1612. doi:10.1038/ng.3963
8. van Arensbergen J, Pagie L, FitzPatrick VD, et al. High-throughput identification of human SNPs affecting regulatory element activity. *Nat Genet*. 07 2019;51(7):1160-1169. doi:10.1038/s41588-019-0455-2
9. Calderon D, Nguyen MLT, Mezger A, et al. Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nat Genet*. 10 2019;51(10):1494-1505. doi:10.1038/s41588-019-0505-9
10. Ohkura N, Yasumizu Y, Kitagawa Y, et al. Regulatory T Cell-Specific Epigenomic Region Variants Are a Key Determinant of Susceptibility to Common Autoimmune Diseases. *Immunity*. Jun 2020;52(6):1119-1132.e4. doi:10.1016/j.immuni.2020.04.006
11. Meuleman W, Muratov A, Rynes E, et al. Index and biological spectrum of human DNase I hypersensitive sites. *Nature*. 08 2020;584(7820):244-251. doi:10.1038/s41586-020-2559-3
12. Consortium G. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. 09 2020;369(6509):1318-1330. doi:10.1126/science.aaz1776
13. Bordi I, Umeton R, Ricigliano VA, et al. A mechanistic, stochastic model helps understand multiple sclerosis course and pathogenesis. *Int J Genomics*. 2013;2013:910321. doi:10.1155/2013/910321
14. Bordi I, Ricigliano VA, Umeton R, et al. Noise in multiple sclerosis: unwanted and necessary. *Ann Clin Transl Neurol*. Jul 2014;1(7):502-11. doi:10.1002/acn3.72
15. Ristori G, Cannoni S, Stazi MA, et al. Multiple sclerosis in twins from continental Italy and Sardinia: a nationwide study. *Ann Neurol*. Jan 2006;59(1):27-34. doi:10.1002/ana.20683
16. Fagnani C, Ricigliano VA, Buscarinu MC, et al. Shared environmental effects on multiple sclerosis susceptibility: conflicting evidence from twin studies. *Brain*. Jul 2014;137(Pt 7):e287. doi:10.1093/brain/awu098
17. Fagnani C, Neale MC, Nisticò L, et al. Twin studies in multiple sclerosis: A meta-estimation of heritability and environmentality. *Mult Scler*. Oct 2015;21(11):1404-13. doi:10.1177/1352458514564492
18. Ricigliano VA, Handel AE, Sandve GK, et al. EBNA2 binds to genomic intervals associated with multiple sclerosis and overlaps with vitamin D receptor occupancy. *PLoS One*. 2015;10(4):e0119605. doi:10.1371/journal.pone.0119605

19. Mechelli R, Manzari C, Policano C, et al. Epstein-Barr virus genetic variants are associated with multiple sclerosis. *Neurology*. Mar 2015;84(13):1362-8. doi:10.1212/WNL.0000000000001420
20. Harley JB, Chen X, Pujato M, et al. Transcription factors operate across disease loci, with EBNA2 implicated in autoimmunity. *Nat Genet*. 05 2018;50(5):699-707. doi:10.1038/s41588-018-0102-3
21. Schwalb B, Michel M, Zacher B, et al. TT-seq maps the human transient transcriptome. *Science*. Jun 2016;352(6290):1225-8. doi:10.1126/science.aad9841
22. Michel M, Demel C, Zacher B, et al. TT-seq captures enhancer landscapes immediately after T-cell stimulation. *Mol Syst Biol*. 03 2017;13(3):920. doi:10.15252/msb.20167507
23. Villamil G, Wachutka L, Cramer P, Gagneur J, Schwalb B. Transient transcriptome sequencing: computational pipeline to quantify genome-wide RNA kinetic parameters and transcriptional enhancer activity. *bioRxiv*. 2019:659912. doi:10.1101/659912
24. Roundtree IA, Evans ME, Pan T, He C. Dynamic RNA Modifications in Gene Expression Regulation. *Cell*. Jun 2017;169(7):1187-1200. doi:10.1016/j.cell.2017.05.045
25. Natoli G, Andrau JC. Noncoding transcription at enhancers: general principles and functional models. *Annu Rev Genet*. 2012;46:1-19. doi:10.1146/annurev-genet-110711-155459
26. Bose DA, Donahue G, Reinberg D, Shiekhattar R, Bonasio R, Berger SL. RNA Binding to CBP Stimulates Histone Acetylation and Transcription. *Cell*. Jan 2017;168(1-2):135-149.e22. doi:10.1016/j.cell.2016.12.020
27. Weinert BT, Narita T, Satpathy S, et al. Time-Resolved Analysis Reveals Rapid Dynamics and Broad Scope of the CBP/p300 Acetylome. *Cell*. 06 2018;174(1):231-244.e12. doi:10.1016/j.cell.2018.04.033
28. Larsson AJM, Johnsson P, Hagemann-Jensen M, et al. Genomic encoding of transcriptional burst kinetics. *Nature*. 01 2019;565(7738):251-254. doi:10.1038/s41586-018-0836-1
29. O'Donoghue GP, Bugaj LJ, Anderson W, Daniels KG, Rawlings DJ, Lim WA. T cells selectively filter oscillatory signals on the minutes timescale. *Proc Natl Acad Sci U S A*. Mar 2021;118(9)doi:10.1073/pnas.2019285118
30. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 01 2019;47(D1):D1005-D1012. doi:10.1093/nar/gky1120
31. Nott A, Holtman IR, Coufal NG, et al. Brain cell type-specific enhancer-promoter interactome maps and disease. *Science*. 11 2019;366(6469):1134-1139. doi:10.1126/science.aay0793
32. Consortium IMSG. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science*. 09 2019;365(6460)doi:10.1126/science.aav7188
33. Marcucci SB, Obeidat AZ. EBNA1, EBNA2, and EBNA3 link Epstein-Barr virus and hypovitaminosis D in multiple sclerosis pathogenesis. *J Neuroimmunol*. 02 2020;339:577116. doi:10.1016/j.jneuroim.2019.577116
34. Sun Y, Peng I, Senger K, et al. Critical role of activation induced cytidine deaminase in experimental autoimmune encephalomyelitis. *Autoimmunity*. Mar 2013;46(2):157-67. doi:10.3109/08916934.2012.750301
35. Mechelli R. Romano C., Reniè R., Manfrè G., Buscarinu M.C., Romano S., Marrone A., Bigi R., Bellucci G., Ballerini C., Angeloni B., Rinaldi V., Salvetti M., Ristori G. Viruses and neuroinflammation in multiple sclerosis. Neuroimmunology and Neuroinflammation; 2021, Accepted.
36. Bäcker-Koduah P, Bellmann-Strobl J, Scheel M, et al. Vitamin D and Disease Severity in Multiple Sclerosis-Baseline Data From the Randomized Controlled Trial (EVIDIMS). *Front Neurol*. 2020;11:129. doi:10.3389/fneur.2020.00129
37. Lee K, Famiglietti ML, McMahon A, et al. Scaling up data curation using deep learning: An application to literature triage in genomic variation resources. *PLoS Comput Biol*. 08 2018;14(8):e1006390. doi:10.1371/journal.pcbi.1006390
38. Sloan CA, Chan ET, Davidson JM, et al. ENCODE data at the ENCODE portal. *Nucleic Acids Res*. Jan 2016;44(D1):D726-32. doi:10.1093/nar/gkv1160
39. Sheffield NC, Bock C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics*. Feb 2016;32(4):587-9. doi:10.1093/bioinformatics/btv612

40. Wilson DJ. The harmonic mean. *Proc Natl Acad Sci U S A*. 01 2019;116(4):1195-1200. doi:10.1073/pnas.1814092116
41. Umeton R. SG, Sorathiya A., Liò P., Papini A., and Nicosia G. Design of robust metabolic pathways. In Proceedings of the 48th Design Automation Conference (DAC '11). ACM, New York, NY, USA, 747-752; 2011.
42. Park A, Oh S, Jung KL, et al. Global epigenomic analysis of KSHV-infected primary effusion lymphoma identifies functional MYC superenhancers and enhancer RNAs. *Proc Natl Acad Sci U S A*. 09 2020;117(35):21618-21627. doi:10.1073/pnas.1922216117
43. Meng FL, Du Z, Federation A, et al. Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability. *Cell*. Dec 2014;159(7):1538-48. doi:10.1016/j.cell.2014.11.014
44. Qian J, Wang Q, Dose M, et al. B cell super-enhancers and regulatory clusters recruit AID tumorigenic activity. *Cell*. Dec 2014;159(7):1524-37. doi:10.1016/j.cell.2014.11.013
45. Orrù V, Steri M, Sidore C, et al. Complex genetic signatures in immune cells underlie autoimmunity and inform therapy. *Nat Genet*. 10 2020;52(10):1036-1045. doi:10.1038/s41588-020-0684-4
46. Factor DC, Barbeau AM, Allan KC, et al. Cell Type-Specific Intralocus Interactions Reveal Oligodendrocyte Mechanisms in MS. *Cell*. 04 2020;181(2):382-395.e21. doi:10.1016/j.cell.2020.03.002
47. Fitz J, Neumann T, Steininger M, et al. Spt5-mediated enhancer transcription directly couples enhancer activation with physical promoter interaction. *Nat Genet*. 05 2020;52(5):505-515. doi:10.1038/s41588-020-0605-6
48. Gutierrez-Arcelus M, Baglaenko Y, Arora J, et al. Allele-specific expression changes dynamically during T cell activation in HLA and other autoimmune loci. *Nat Genet*. 03 2020;52(3):247-253. doi:10.1038/s41588-020-0579-4
49. Isoda T, Moore AJ, He Z, et al. Non-coding Transcription Instructs Chromatin Folding and Compartmentalization to Dictate Enhancer-Promoter Communication and T Cell Fate. *Cell*. Sep 2017;171(1):103-119.e18. doi:10.1016/j.cell.2017.09.001
50. Chong S, Dugast-Darzacq C, Liu Z, et al. Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science*. 07 2018;361(6400)doi:10.1126/science.aar2555
51. Sabari BR, Dall'Agnese A, Boija A, et al. Coactivator condensation at super-enhancers links phase separation and gene control. *Science*. 07 2018;361(6400)doi:10.1126/science.aar3958
52. Cho WK, Spille JH, Hecht M, et al. Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science*. 07 2018;361(6400):412-415. doi:10.1126/science.aar4199
53. Halfon MS. Silencers, Enhancers, and the Multifunctional Regulatory Genome. *Trends Genet*. 03 2020;36(3):149-151. doi:10.1016/j.tig.2019.12.005
54. Wong ES, Zheng D, Tan SZ, et al. Deep conservation of the enhancer regulatory code in animals. *Science*. 11 2020;370(6517)doi:10.1126/science.aax8137

TABLES

List	-log(p-value)	p-value	Odds Ratio	Support	List Size	Harmonic Score
Whole transient transcriptome	8.55	**2.80×10^{-9}**	1.65	241	22126	41.3
Short half-life transcripts	7.68	**2.06×10^{-8}**	1.63	209	20143	40.1
Long half-life transcripts	1.05	0.09	1.29	35	1993	17.1

Table 1. Enrichment of MS-associated genetic variants in lists of T-cell transient transcripts extracted from Michel et al.[22] The whole transcriptome list was split in two sub-lists depending on the transcripts' half-life: short (<60') and long (≥60'), respectively. Results are considered significant at $p<0.05$ and are highlighted in bold.

	± 50 KB				± 100 KB				± 200 KB			
	-log (pValue)	Odds Ratio	Support	Harmonic Score	-log (pValue)	Odds Ratio	Support	Harmonic Score	-log (pValue)	Odds Ratio	Support	Harmonic Score
EBNA2 ***(6880)***	**10.658**	**1.790**	**158**	**45.544**	**8.616**	**1.509**	**239**	**38.327**	**15.444**	**1.542**	**421**	**41.913**
EBNA3C ***(3335)***	0.614	1.108	55	11.765	**1.647**	**1.227**	**109**	**20.956**	**3.448**	**1.294**	**199**	**28.098**
AID ***(4823)***	**4.963**	**1.596**	**99**	**35.793**	**3.890**	**1.374**	**153**	**30.259**	**13.924**	**1.619**	**309**	**43.308**
VDR ***(23409)***	**19.348**	**1.575**	**474**	**43.564**	**19.181**	**1.422**	**767**	**39.635**	**32.090**	**1.424**	**1329**	**40.872**

Table 2. Enrichment of MS-GWAS regions in lists of DNA binding sites of human and viral molecular transducers. For each molecular transducer, the number of DBRs is shown in brackets in the right-most column. Columns display the statistical analysis results derived from the colocalization analysis of the MS-associated genomic positions at different ranges of extension (±50,100,200 kb). Results are considered significant at $p<0.05$, corresponding to a $-\log(p)>1.301$. Significant results are highlighted in bold.

		± 50 KB				± 100 KB				± 200 KB			
		-log (pValue)	Odds Ratio	Support	Harmonic Score	-log (pValue)	Odds Ratio	Support	Harmonic Score	-log (pValue)	Odds Ratio	Support	Harmonic Score
EBNA2	*Long half-life*	0.023	0.478	3	0.644	0.062	0.717	8	1.708	**1.879**	**1.531**	**33**	**24.679**
	Short half-life	**6.163**	**1.920**	**69**	**43.011**	**3.241**	**1.433**	**95**	**29.496**	**8.945**	**1.610**	**189**	**40.642**
EBNA3C	*Long half-life*	0.064	0.572	2	1.669	0.006	0.321	2	0.185	0.182	0.914	11	4.500
	Short half-life	0.070	0.794	16	1.923	0.023	0.752	28	0.661	0.066	0.875	58	1.841
AID	*Long half-life*	0.089	0.682	3	2.303	0.283	1.024	8	6.477	0.051	0.726	11	1.432
	Short half-life	**1.769**	**1.465**	**37**	**23.531**	**1.346**	**1.267**	**59**	**19.367**	**3.954**	**1.442**	**119**	**31.416**
VDR	*Long half-life*	**1.737**	**1.502**	**32**	**23.571**	0.845	1.187	45	14.646	**2.315**	**1.322**	**97**	**25.031**
	Short half-life	**2.221**	**1.239**	**152**	**23.734**	**2.336**	**1.181**	**267**	**23.460**	**11.478**	**1.367**	**548**	**36.561**

Table 3. Colocalization of human and viral transducer DBRs and MS-GWAS positions in transient transcripts. MS-associated genomic positions were analyzed at different ranges of extension (±50,100, 200 kb). The transcript half-life is considered short if <60' and long if ≥60', respectively. Results are considered significant at $p<0.05$, corresponding to a $-\log(p)>1.301$. Significant results are highlighted in bold.

FIGURES

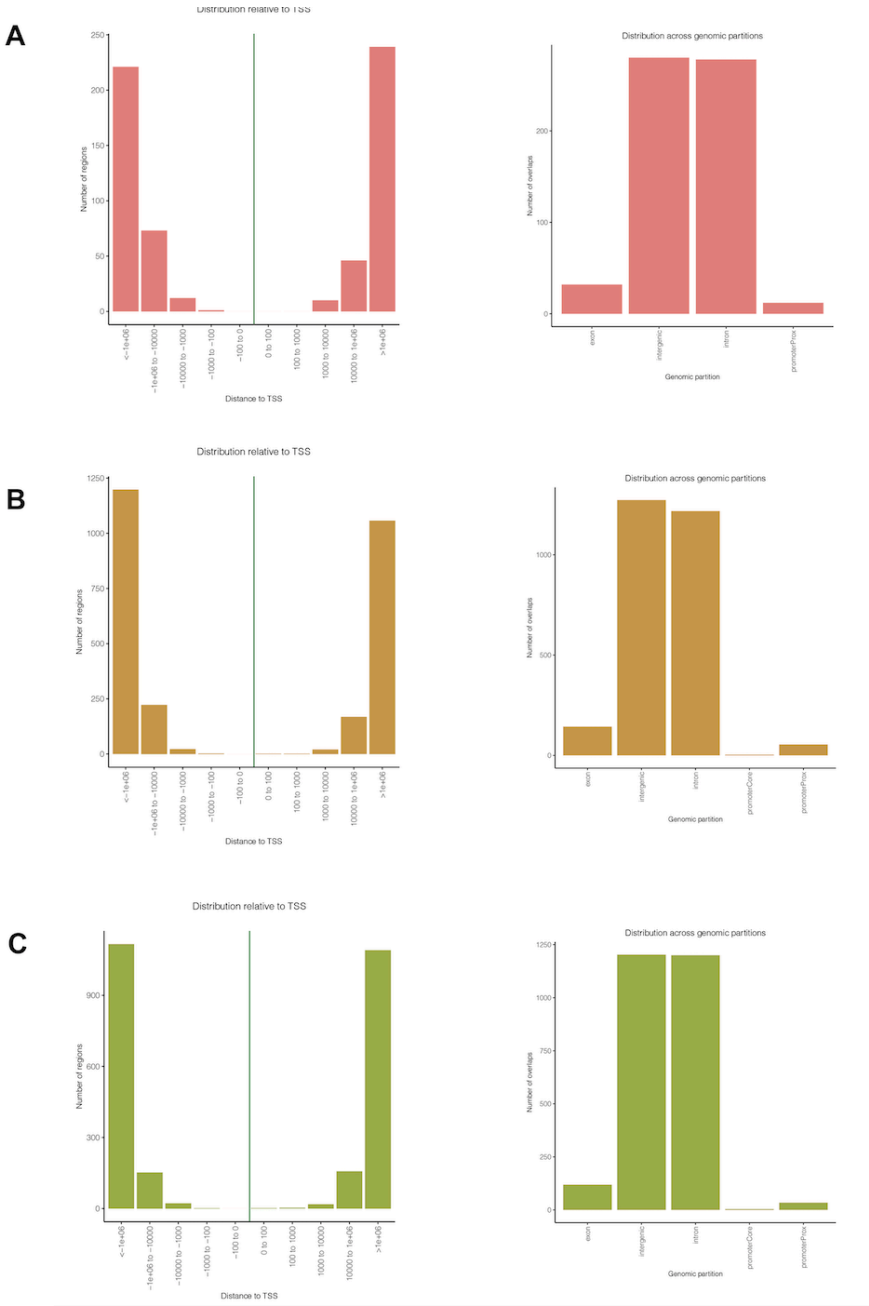


Figure 1. GWAS-associated SNP distribution across genomic partitions and their distance relative to the transcription starting site (TSS). Panel A) Multiple Sclerosis; B) Immune-mediated conditions: Multiple Sclerosis, Rheumatoid Arthritis, Systemic Lupus Erythematosus, Crohn's Disease, Ulcerative Colitis, Inflammatory Bowel Disease, Celiac Disease, Asthma, Type I Diabetes Mellitus; C) Non-immunological complex conditions: Type II Diabetes Mellitus, Aging, Obesity, Hypertension, Coronary Artery Disease, Bipolar Disorder. Supplementary table 2 include links to these traits in the GWAS catalog.

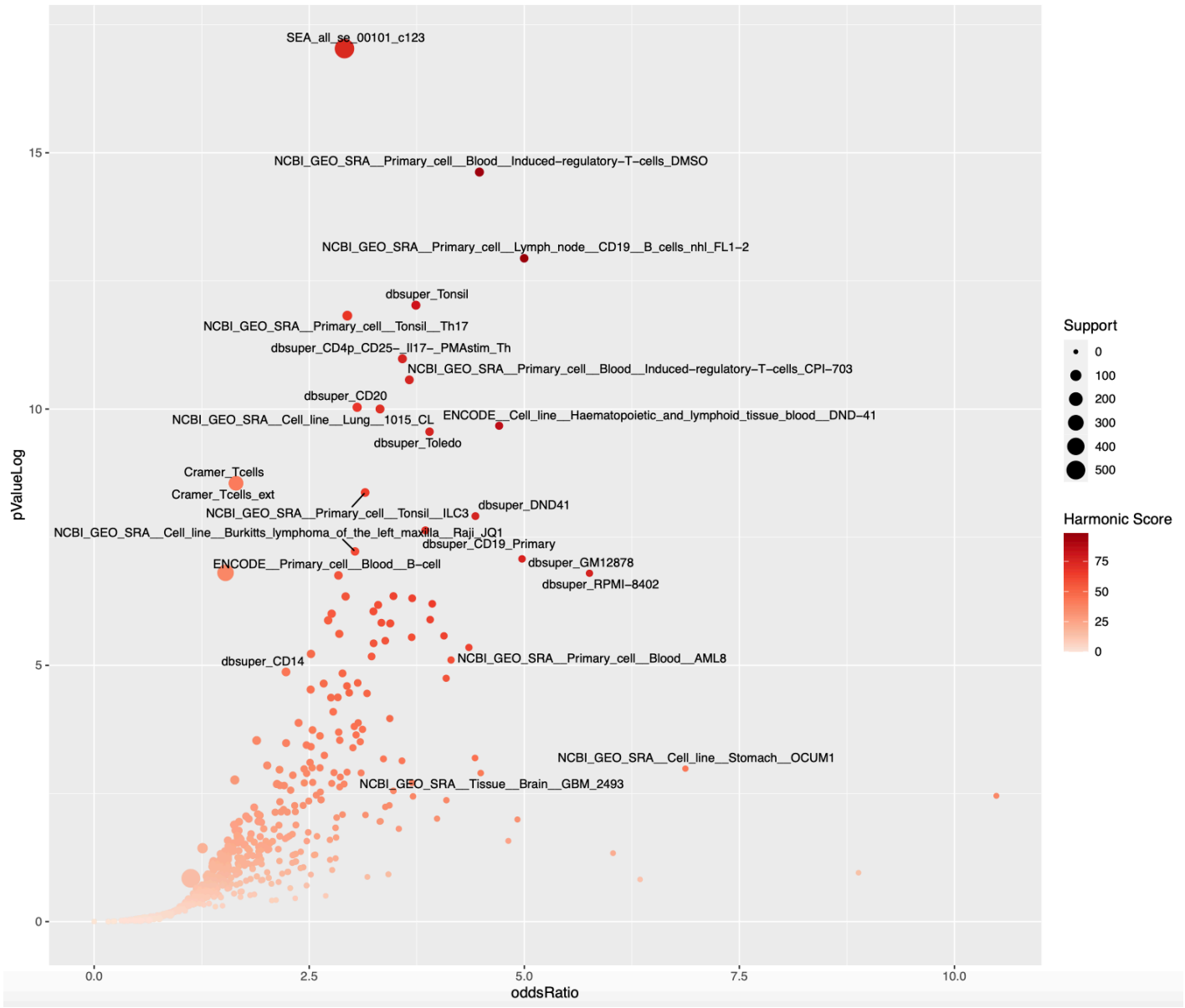


Figure 2. Enrichment of MS-associated SNPs in databases of regulatory elements, sorted by experiment/cell lines. X-axis shows the Odd Ratio, y-axis shows the -log(pValue); dot size is proportional to the support of each match, i.e., the number of hits resulting from the colocalization analysis. Color of each point is related to the Harmonic Score (HS), a comprehensive estimation of the relevance of hits, as derived by merging and balancing the OR, pValue and Support of each match. Thus, prioritized hits are represented by the darker dots that occupy the upper-right area of the chart. Labeled points have HS>40.

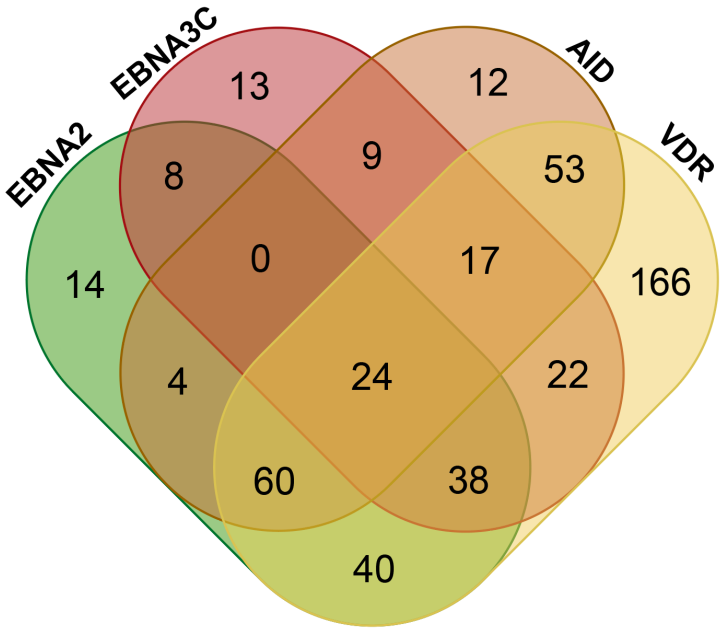


Figure 3. MS-GWAS regions in DNA binding sites of human and viral molecular transducers. This Venn diagram shows the number of non-redundant MS-associated SNPs (from the ROI, supplementary table 1) and their extensions at ±50, 100, 200 kb that colocalized within DNA binding regions of the molecular transducers AID, VDR, EBNA2, EBNA3C. For each transducer, SNPs are considered only once if present in more than one match. Intersections show the numbers of regions colocalizing with DBRs of multiple transducers. For instance: 8 regions colocalize with both EBNA2 and EBNA3C DBRs, but not with AID nor VDR DBRs; 24 regions colocalize with all four DBRs

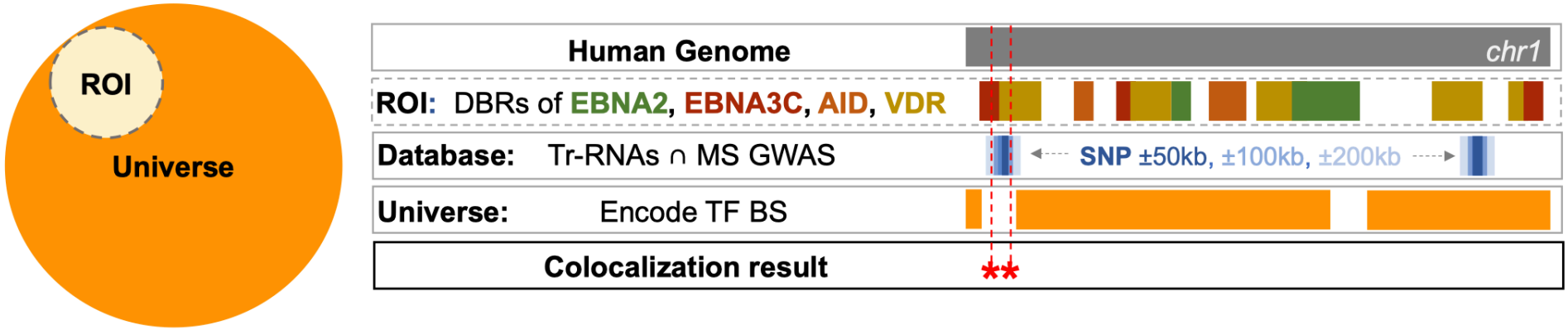


Figure 4. Schematic representation of the colocalization analysis. (ROI: Region of interest; DBR: DNA Binding Regions; ENCODE TFBS: Transcription Factor Binding Site). The figure shows the tracks we considered for the colocalization analyses. In brief, the ROI included the DBRs of MS-related viral and human transducers and was matched with MS-associated SNPs extended by 50, 100, and 200 kilobases that colocalize with regions plausibly coding for trRNAs (Database). As a control (Universe), we took from ENCODE the entire list of transcription factors binding sites. Results were considered significant if a colocalization was found across ROI and a Databases element without occurring in the Universe as a statistically significant match.

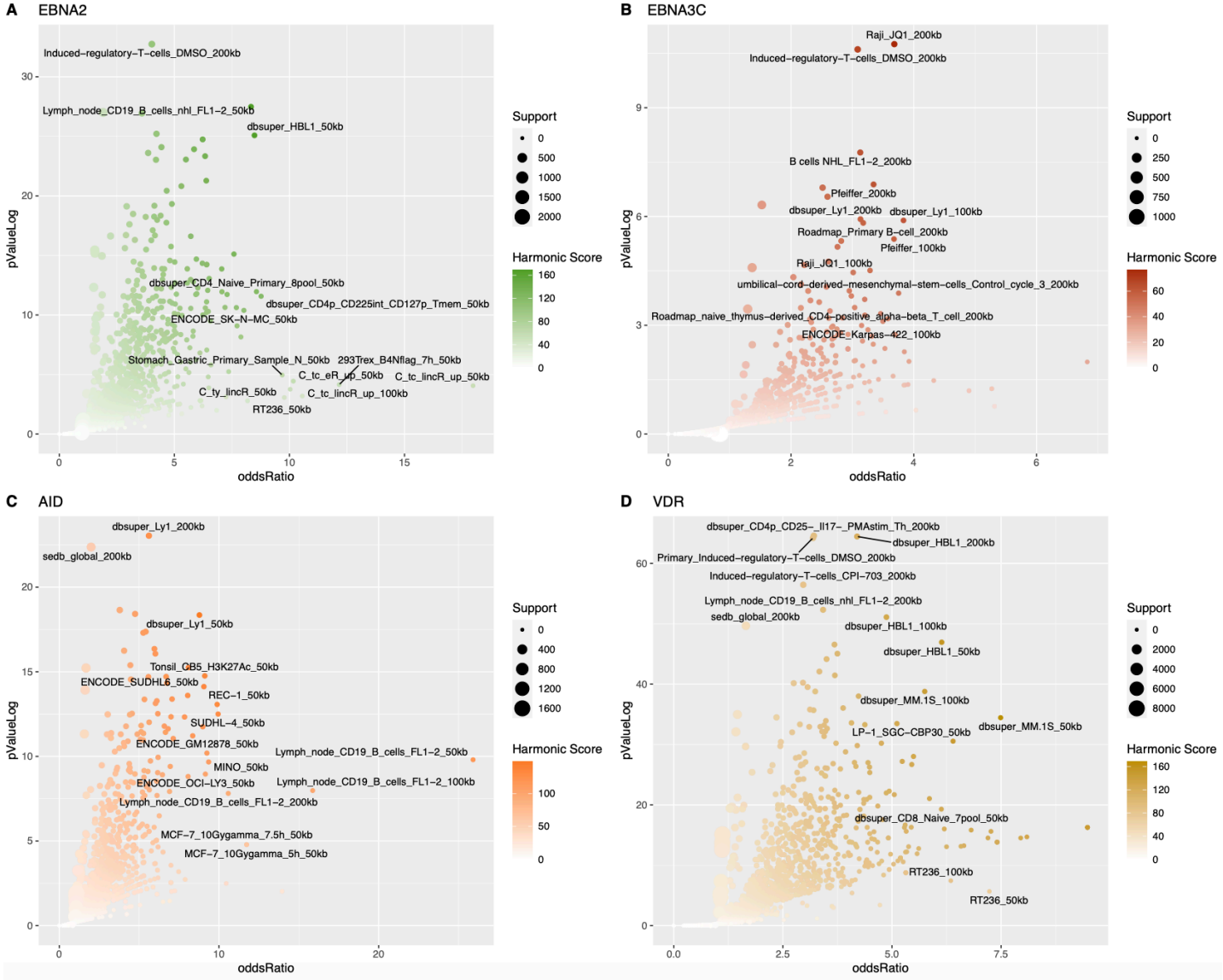
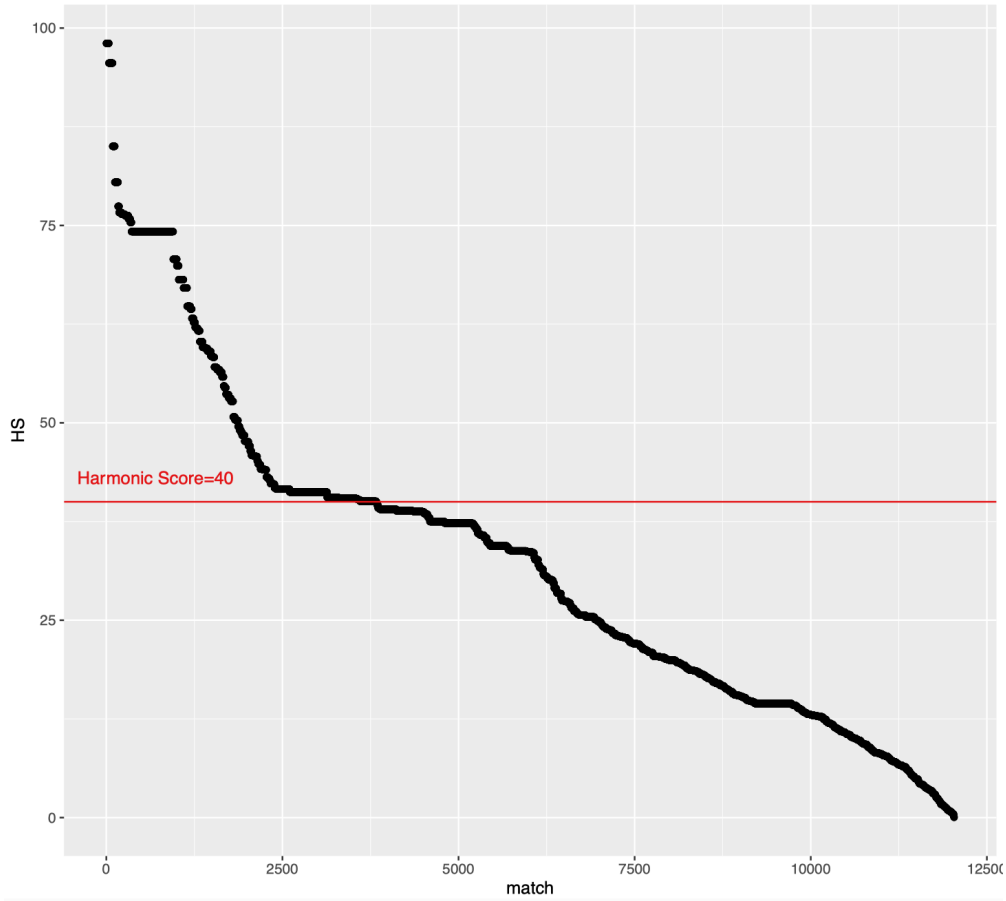


Figure 5. Colocalization analysis of DBRs for human and viral molecular transducers, MS-associated SNPs and DNA regulatory regions derived from databases. (A) EBNA2; (B) EBNA3C; (C) AID; (D) VDR. The charts display results of all matches, i.e. with MS-associated SNPs and their extension at ±50, 100, 200 kb. X-axis shows the Odd Ratio, y-axis shows the log(pValue). Dot size is proportional to the support of each match, i.e. the number of hits resulting from each colocalization analysis. The color of each dot is related to the Harmonic Score (HS); Labels are shown starting at HS>40.

SUPPLEMENTARY FIGURES



Supplementary Figure 1. Harmonic Score threshold defining the top colocalization hits. The plot shows all of colocalization matches, ranked by Harmonic Score (HS). The curve inflection point, highlighted with a red line, suggests a threshold for selecting the most relevant hits (i.e., those scoring at HS>40).