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

Renato Umeton
Massachusetts Institute of Technology

Let us know how access to this document benefits you.
Follow this and additional works at: https://escholarship.umassmed.edu/faculty_pubs

Part of the Nervous System Diseases Commons, and the Neuroscience and Neurobiology Commons

Repository Citation
Retrieved from https://escholarship.umassmed.edu/faculty_pubs/1924

Creative Commons License
This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License. This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in University of Massachusetts Medical School Faculty Publications by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
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 Salvetti5,12,* Giovanni Ristori5,13,*

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

umeton@mit.edu

marco.salvetti@uniroma1.it

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 provides a freely available data resource at www.mscoloc.com for future research on transcriptional regulation in MS.
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,\textsuperscript{1-5} even though a recent study showed that low-frequency and rare coding variants may somewhat contribute to MS heritability.\textsuperscript{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;\textsuperscript{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;\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{13, 14} Such model is consistent with our previous results that showed heterogeneity in the MS etiology components in twin pairs studies.\textsuperscript{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.\textsuperscript{18} We also demonstrated that genomic variants of \textit{EBNA2} resulted to be MS-associated,\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{28} Along these lines, a recent study showed that T cells selectively filter oscillatory signals within the minute timescale,\textsuperscript{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),\textsuperscript{21,}
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 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. 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 mapping31, 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.\textsuperscript{18, 20} Therefore we further inquired whether DNA regions plausibly coding for tRNA 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.\textsuperscript{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.mscocoloc.com, detailing the genomic regions targeted by each of the four molecular transducers.

Building once again on the work by Michel et al.\textsuperscript{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 tRNA. 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 tRNA 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 tRNA-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](http://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,\textsuperscript{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,\textsuperscript{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,\textsuperscript{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.\textsuperscript{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. 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. 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, 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. 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. 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. 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](http://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;\textsuperscript{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\textsubscript{0003885}. All Transcription Factor Binding Site regions (TFBS) were obtained from the ENCODE portal.\textsuperscript{38} All data was organized in various databases and data pipelines as detailed below. For SNP overlaps and region colocalization, we used LOLA\textsuperscript{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 \(-\log(p\text{-value})\), support, and Odds Ratio (OR) were combined into a single score inspired by the harmonic mean\textsuperscript{40} and multi-objective optimization\textsuperscript{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 \url{www.mscoloc.com}.

\[
HarmonicScore = k_p \times \frac{\sum_i w_i}{\log P} + \frac{w_2}{\text{Supp}} + \frac{w_3}{\text{OR}}
\]
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

### Table 1. Enrichment of MS-associated genetic variants in lists of T-cell transient transcripts extracted from Michel et al.\textsuperscript{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.

<table>
<thead>
<tr>
<th>List</th>
<th>-log(p-value)</th>
<th>p-value</th>
<th>Odds Ratio</th>
<th>Support</th>
<th>List Size</th>
<th>Harmonic Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole transient transcriptome</td>
<td>8.55</td>
<td>2.80 x 10^-9</td>
<td>1.65</td>
<td>241</td>
<td>22126</td>
<td>41.3</td>
</tr>
<tr>
<td>Short half-life transcripts</td>
<td>7.68</td>
<td>2.06 x 10^-8</td>
<td>1.63</td>
<td>209</td>
<td>20143</td>
<td>40.1</td>
</tr>
<tr>
<td>Long half-life transcripts</td>
<td>1.05</td>
<td>0.09</td>
<td>1.29</td>
<td>35</td>
<td>1993</td>
<td>17.1</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Transducer</th>
<th>-log(pValue)</th>
<th>Odds Ratio</th>
<th>Support</th>
<th>Harmonic Score</th>
<th>-log(pValue)</th>
<th>Odds Ratio</th>
<th>Support</th>
<th>Harmonic Score</th>
<th>-log(pValue)</th>
<th>Odds Ratio</th>
<th>Support</th>
<th>Harmonic Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBNA2</td>
<td>10.658</td>
<td>1.790</td>
<td>158</td>
<td>45.544</td>
<td>8.616</td>
<td>1.509</td>
<td>239</td>
<td>38.327</td>
<td>15.444</td>
<td>1.542</td>
<td>421</td>
<td>41.913</td>
</tr>
<tr>
<td>(6880)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3335)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4823)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VDR</td>
<td>19.348</td>
<td>1.575</td>
<td>474</td>
<td>43.564</td>
<td>19.181</td>
<td>1.422</td>
<td>767</td>
<td>39.635</td>
<td>32.090</td>
<td>1.424</td>
<td>1329</td>
<td>40.872</td>
</tr>
<tr>
<td>(23409)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>± 50 KB</th>
<th>± 100 KB</th>
<th>± 200 KB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-log (pValue)</td>
<td>Odds Ratio</td>
<td>Support</td>
</tr>
<tr>
<td>EBNA2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long half-life</td>
<td>0.023</td>
<td>0.478</td>
<td>3</td>
</tr>
<tr>
<td>Short half-life</td>
<td>6.163</td>
<td>1.920</td>
<td>69</td>
</tr>
<tr>
<td>EBNA3C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long half-life</td>
<td>0.064</td>
<td>0.572</td>
<td>2</td>
</tr>
<tr>
<td>Short half-life</td>
<td>0.070</td>
<td>0.794</td>
<td>16</td>
</tr>
<tr>
<td>AID</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long half-life</td>
<td>0.089</td>
<td>0.682</td>
<td>3</td>
</tr>
<tr>
<td>VDR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long half-life</td>
<td>1.737</td>
<td>1.502</td>
<td>32</td>
</tr>
<tr>
<td>Short half-life</td>
<td>2.221</td>
<td>1.239</td>
<td>152</td>
</tr>
</tbody>
</table>
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 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).