Analysis of defective pathways in Multiple Sclerosis via machine learning approaches

Enrique J. deAndrés-Galiana[a,b,1,*], Juan L. Fernández-Martínez[b,2], Leo N. Saligan[c,3]

[a] *Department of Informatics and Computer Science, University of Oviedo. Calvo Sotelo s/n 33007, Oviedo, Asturias, Spain.*
[b] *Group of Inverse Problems, Optimization and Machine Learning, Department of Mathematics, University of Oviedo. Calvo Sotelo s/n 33007, Oviedo, Asturias, Spain.*
[c] *Symptom Management Branch. Division of Intramural Research. National Institute of Nursing Research. Building 3, Room 5E14 3 Center Drive Bethesda, MD 20892. USA.*

Abstract

Background: Although some studies show that there could be a genetic predisposition to develop Multiple Sclerosis (MS), attempts to find genetic signatures related to MS diagnosis and development are extremely rare. **Method**: In the present study, we carried out a retrospective analysis of two different microarray datasets, using machine learning techniques to understand the defective pathways involved in this disease. We have modeled two data sets publicly accessible. One was used to establish the list of most discriminatory genes and defective MS pathways; whereas, the second one was utilized for validation purposes. **Results**: Our analysis provides a list of high discriminatory genes with predictive Cross-validation accuracy higher than 95%, both in learning and blind validation. The most discriminatory genes seem to be related to the production

*Corresponding author
Email addresses: andresenrique@uniovi.es (Enrique J. deAndrés-Galiana), jlfm@uniovi.es (Juan L. Fernández-Martínez), saliganl@mail.nih.gov (Leo N. Saligan)

[1] **Enrique J. deAndrés-Galiana** is an assistant professor at Computer Science department, University of Oviedo, Asturias, Spain. His primary research interest is the application of bioinformatics to pharmacogenomics.

[2] **Juan L. Fernández-Martínez** is professor at Applied Mathematics department, University of Oviedo, Asturias, Spain. His research interests are inverse problems, optimization and machine learning applied to genomics.

[3] **Leo N. Saligan** in Symptom Management Branch investigator at Division of Intramural Research, National Institute of Nursing Research, USA. His current research focuses on understanding biobehavioral mechanisms of symptoms with the longer-term goal of developing novel interventions that can alleviate this symptom in a variety of clinical conditions.

of Hemoglobin. **Conclusions**: The biological processes involved were related to T-cell Receptor Signaling and co-stimulation, Interferon-Gamma Signaling and Antigen Processing and Presentation. This suggests a viral or bacterial infection as the plausible mechanism involved in MS development. The pathway analysis also confirmed coincidences with Epstein-Barr virus, and revealed them with Influenza A, Toxoplasmosis, Tuberculosis and Staphylococcus Aureus infections. Th17 Cell differentiation, and CD28 co-stimulation seemed to be crucial in the development of this disease. These results are confirmed via the holdout sampler. Additional knowledge provided by this analysis will help to identify new therapeutic targets.

Keywords: Multiple Sclerosis, Machine Learning, Pathway Analysis

---

1. Background

Multiple Sclerosis (MS) is a neurological disease characterized by the appearance of focal lesions in the white matter, in which the most striking aspect is demyelination with a relative preservation of the axons [1]. The said lesions are distributed along the Central Nervous System (CNS). It is accepted that MS is an autoimmune disorder, since acute injuries are detected in collaborative T-cells (CD4+) and in anomalous expression of major histocompatibility class II antigens in macrophages and atrocytes [2, 3].

MS affects around 2.5 million individuals worldwide and its development can be progressive or relapsing-remitting [4].According to the literature, there are several forms of multiple sclerosis: relapsing-remitting, secondary-progressive, primary-progressive and progressive-relapsing, being the relapsing-remitting the most recurrent type [5]. MS is also geographically dependent, being more common in regions further away from the equator, where about 1 in 2,000 people are affected, while in countries closer to the equator MS affects approximately 1 in 20,000 subjects [6]. Additionally, for unknown reasons MS affects women twice as often as men [7].

The two main causes of MS are genetic and environmental. Exposure to

Epstein-Barr virus [8], low levels of vitamin D [9, 10], and smoking [11] have been cited as plausible factors, which may increase the probabilities of developing MS. Regarding the genetic hypothesis, it has been shown that certain ethnic groups are more resistant than others to MS, and there is a higher prevalence in the case of Scandinavians [12]. Besides, it has been shown that the empiric recurrence between siblings increases by a factor of 10 to 50 and the concordance between homozygous twins is higher than in dizygotic ones [13]. An association with *HLA-DRB1* has been established [14]. Furthermore, changes in the *IL7R* gene could increase the risk of developing MS [15, 16].

Also, *NR1H3* seems to be associated with primary progressive MS [17]. The genetics of MS have been reviewed by [18]. Microarray gene expression profiling analysis in MS was also performed by [19]. These authors suggest the antigen presentation as the main cause for MS.

Although there could be a genetic predisposition of developing MS, attempts to find genetic signatures related to MS diagnosis and development are limited. In the present study, we identify a small-scale gene signature (22 genes) able to predict MS occurrence with a Leave-One-Out-Cross Validation (LOOCV) accuracy higher than 98% in a data-set with 113 samples (healthy controls and MS). The discriminatory power of this gene signature has been blind-validated with an independent cohort composed of 20 samples, obtaining a validation accuracy of 90%. Remarkably, expanding the gene signature to the 63 first most discriminatory genes, the validation accuracy increases to 95%. These results show that the genetic signatures used to establish the defective pathways are very robust. We, subsequently, unravel the biological pathways involved using an expanded list containing the most discriminatory genes with LOOCV accuracy greater than 90%, with the potential goal of identifying new therapeutic targets. Furthermore, we show a correlation tree to analyze the existing relationships among the most discriminatory genes to understand how they control gene expression. The main biological processes involved are: T-cell Receptor Signaling, Interferon-Gamma Signaling, T-cell co-stimulation, Antigen Processing and Presentation, Antigen Processing and Presentation of Exogenous Anti-

gen Via MHC Class II, suggesting viral/bacterial infections as possible causes for MS. Epstein-Barr Virus Infection and Influenza A pathways appear in the list of important defective pathways and could be related to its development. Toxoplasmosis, Tuberculosis and Staphylococcus Aureus infections also appear in the list of defective pathways. We believe that the most discriminatory genes (in this case HBB) should be considered as novel targets for drug repositioning to treat this disease.

This study introduces a novel methodology which introduces a revolutionary process that enables researchers to utilize the data to obtain pathways, and, consequently, the precise drug-repositioning through machine learning approaches. Furthermore, this methodology is extremely simple to apply using very few resources.

2. Method

2.1. Learning and Validation Data-sets

In this paper we have modeled two data sets (expression microarrays) that are public accessible. The first one (European Bioinformatic Institute E-GEOD-13732) was used to establish the list of most discriminatory genes and the defective MS pathways. The gene expression was originally measured from naive CD4+ T cells and it consists of 54675 probes and 113 samples: 73 MS patients and 40 healthy controls. The samples correspond to patients who were initially diagnosed as Clinical Isolated Syndrome (the earliest clinical manifestation of MS) and they finally progress to clinically definite MS. There were two samples per patient, one on the CIS detection and the other on the MS diagnosis. Characteristics of patients such as age, gender and ethnicity can be consulted in the original research of the data [20].

For the validation process we used a second independent microarray (European Bioinformatic Institute E-GEOD-43592). The expression was also measured from T cells and it was composed of 20 samples: 10 controls, 10 MS patients and 54675 probes. In this case, the samples correspond to relapsing

remitting subjects and they were taken after relapsing. More detailed information as specific time of sampling, age, sex or ethnicity can be consulted in the original research [21]. Both data-sets were measured with the Affymetrix Human Genome U133 Plus 2.0 platform.

2.2. Analysis

Phenotype prediction problems are highly under-determined since the number of genetic probes that are monitored is much higher than the number of samples (patients) used to analyze the genetic causes involved in the disease. Hypothetically, if the classifier was linear, its null-space (containing the genes that do not have any effect in the predictive accuracy) would have a dimension equal to the number of genetic probes minus the number of independent samples. Then, it can be easily understood that many different genetic signatures exist, all exhibiting a high predictive accuracy. These signatures form the uncertainty space of the classifier that is used to solve the phenotype prediction problems [22]. The use of feature selection methods to filter and to rank the genetic probes according to their discriminatory power is crucial to correctly address the analysis of the defective pathways in phenotype prediction problems. Multiple Sclerosis is not a special case in that sense.

Figure 1 shows the flow diagram of the whole methodology used in this paper. To establish the discriminatory power of the genes, we used a combination of Fold Change [23] and Fisher's Ratio [24], by finding the genes that are under and over expressed (high absolute fold change value) and at the same time obtain a very high Fisher's ratio.The Fisher's ratio tends to give higher importance to features with lower standard deviation (i.e. very stable values throughout all the feature), even though the difference between means in each class were low. In order to avoid this bias, we firstly applied Fold Change, removing thereby all the features with lower standard deviation. Subsequently, we calculate the Fisher's ratio and add it the absolute value of the Fold Change. This combined ratio is used to establish the "prior discriminatory power" of each individual gene in the discrimination of MS patients from healthy controls.

This methodology serves to perform the initial gene filtering needed to reduce the high degree of indeterminacy and to avoid the use of genes that only expand high frequency details in the phenotype prediction, which are irrelevant in the disease development.

Once the prior discriminatory genes are determined and ranked in decreasing order by their prior discriminatory power, the next step is to determine the shortest list of prognostic genes (small-scale signature) with the highest predictive accuracy. This predictive accuracy is what we call the "posterior discriminatory power", and it is determined via Leave-One-Out-Cross-Validation (LOOCV), to estimate how accurately the predictive model (classifier) would perform in practice for new incoming samples whose MS status is unknown. LOOCV is a well-established cross-validation method by which a single sample from the original data-set serves as sample test of the data, and the remaining samples act as training. The average LOOCV predictive accuracy is calculated by iterating over all the samples in the data-set. This predictive accuracy is theoretical and should be assessed using blind validation (if possible). The predicted class is obtained using in a distance-based classifier using the list of most discriminatory genes. In our case, we decided to preserve the genes with a prior discriminatory power higher than 1.52. In this classifier the class with the minimum distance in the reduced genetic signature to the sample test is assigned to be the estimated class of the new incoming sample.

This algorithm aids to find the small-scale signature with the highest predictive accuracy when it is combined with a feature elimination procedure that consists of generating increasingly shorter lists of high discriminatory genetic probes ranked by their prior discriminatory power. That is, we evaluate the accuracy of the entire set of genes ranked by their prior discriminatory power and, then, we remove the last gene, which has the least discriminatory power to finally evaluate the accuracy of this genetic signature. We keep removing genes and evaluating the accuracy until we eliminate all the genes of the ranked list. This operation is commonly called Backwards Feature Elimination and it allows us to obtain different subsets of genes with their associated posterior discrimina-

tory power. Lastly, we select the smallest subset of genes (small-scale signature) with the highest posterior discriminatory power or accuracy. This methodology has been successfully employed in different genetic studies [25, 26, 27].

The predictive power of the set of most discriminatory genes is validated with an independent cohort. We predict the class of the samples in this independent cohort via LOOCV using the distance-based classifier that was designed in the learning stage. To conclude and aiming to understand the defective genetic and biological pathways that are involved and identifying possible new druggable targets, we carried out a pathway analysis of the most prior discriminatory genes using GeneAnalytics© [28]. Furthermore, we depicted a correlation network between the 30 most discriminatory genes. This correlation network measures the mutual dependence between gene expressions and allows to analyze interrelationships among genes impacting both expression and function. For the set-up of the network, we employed the approach presented in [29], using the Pearson correlation (PC) coefficient [30] to measure the dependency between genes. Once the Pearson's coefficient is calculated, Kruskal's algorithm [31] is used to find the minimum-spanning-tree between the selected genes and to build the correlation network using as head the gene with the most discriminatory power. The first gene is the header (the most discriminatory one), and the branches contain the genes which are helpers.

3. Results and Discussion

Table 1 shows the small-scale gene signature with 98.2% accuracy in the first data-set. Working with the validation test, this list of genes provided 90% accuracy. Moreover, expanding the list to the first 63 most discriminatory genes we achieved 95% LOOCV accuracy in blind validation (please see supplementary material). This fact shows that the defective genes found in the first data-set are involved in the genesis of this disease.

Gene ID	μ_1	σ_1	μ_2	σ_2	Ratio	Acc
HBB	9.6178	1.7120	5.2536	1.0473	5.2054	93.8053
HBA1	9.7540	1.6073	5.3387	1.1988	4.8755	93.8053
HBB	10.1282	1.6107	6.3339	0.9792	4.3526	93.8053
HBB	9.4707	1.5321	6.3914	0.6838	3.1223	93.8053
214349_at	9.3558	0.8450	7.5038	1.0090	2.6888	93.8053
GAS7	7.9834	0.5455	6.8498	0.5561	2.3466	93.8053
AC079305.10	6.4999	0.5921	5.3935	0.6390	2.1945	94.6903
HBA1	8.8871	1.5835	6.1203	0.6607	2.1918	94.6903
CST3	8.6560	0.9675	6.8275	0.9471	2.1450	94.6903
BNIP1	8.1360	0.4495	7.1404	0.5862	2.1308	95.5752
JARID2	6.6288	0.4393	5.8171	0.5200	2.1112	95.5752
ITPRIP	7.9194	0.5279	6.7963	0.6319	2.0834	95.5752
SEC14L1	8.0327	0.7853	6.5745	0.8525	2.0178	96.4602
CAPRIN1	6.9504	0.3162	6.4497	0.2406	2.0047	95.5752
TP53INP2	7.5928	0.9166	6.0999	0.7304	1.9939	93.8053
NFKBIE	7.6954	0.4867	6.7592	0.4973	1.9916	93.8053
C14orf132	6.7044	0.5369	7.9096	0.7216	1.9786	95.5752
IRF8	9.5788	0.9937	8.0582	1.1007	1.9765	96.4602
241849_at	6.5765	0.5649	5.6072	0.5981	1.9090	96.4602
SYNGR1	7.5493	0.6663	6.2107	0.9396	1.8854	96.4602
MIR6883	8.2809	0.6185	7.0572	0.6545	1.8797	96.4602
TMEM45B	5.2715	0.7311	6.7097	0.7130	1.8608	**98.2301**

Table 1: List of the most discriminatory genes found by the algorithm. μ_1 and σ_1 refer, respectively, to the mean expression and standard deviation in the healthy control group whereas μ_2 and σ_2 indicate the same for the MS group. Ratio stands for the prior discriminatory power for each gene, which is a combination of Fisher's Ratio and Fold Change. These 22 genes correspond to the small-scale gene signature with the best LOOCV learning accuracy (98.2%). This list provided 90% of accuracy in the validation data-set. Expanding the list to the first 63 genes we achieved 95% accuracy in blind validation (please see supplementary material).

3.1. Ontological and pathway analysis

Tables 2, 3, 4, 5, and 6 shows the Genetic Pathways, the GO Biological processes, the GO Molecular Functions, the Phenotype and the Compounds obtained from the analysis of the most discriminatory genes via GeneAnalitics© tool. The score is calculated based on the binomial distribution to test the null hypothesis that input genes are not over-represented within any SuperPath, GO term or compound. The presented score for each match is a transformation (-log2) of the resulting p-value, where higher scores indicate better matches [28].

Interestingly, the main compound found was 1D09C3, which is an anti-MHC (major histocompatibility complex) class II monoclonal antibody. 1D09C3 binds to MHC class II molecules on the cell surface and selectively kills proliferating tumor cells. The second most important compound found was the Sebacic Acid.

Figure 2 shows the correlation tree of the first 31 most discriminatory genes. The numbers of the edges represent the Pearson coefficient between nodes. It can be observed two main branches related to the HBB gene, headed by GAS7 and HBA1/HBA2. The HBB gene provides instructions for making a protein called beta-globin. Beta-globin is a subunit of a the hemoglobin, that normally consists of four protein subunits: two subunits of beta-globin and two subunits of another protein called alpha-globin, which is produced from the HBA gene. Both genes are under-expressed in patients with Multiple Sclerosis. GAS7, plays a putative role in neuronal development and it is also under-expressed in MS subjects. This gene has also been associated to MS [32]. IRF8 (interferon regulatory factor 8) is a protein-coding gene: interferon consensus sequence-binding protein (ICSBP) is a transcription factor of the interferon (IFN) regulatory factor (IRF) family. IRF family proteins also control expression of IFN-alpha and IFN-beta-regulated genes that are induced by viral infection. Finally, JARID2 has also been associated to MS, since Jarid2-deficient NKT cells perturb Th17 differentiation, leading to reduced Th17-driven autoimmune pathology [33].

Score	SuperPath	Total Genes	Matched Genes
34.36	Rheumatoid Arthritis	90	12
24.27	Staphylococcus Aureus Infection	56	8
21.59	Interferon Gamma Signaling	202	12
19.91	Influenza A	315	14
19.54	G-protein Signaling N-RAS Regulation Pathway	60	7
18.78	Immune Response NFAT in Immune Response	162	10
18.37	Hematopoietic Cell Lineage	97	8
17.95	CTLA4 Signaling	71	7
17.91	Translocation of ZAP-70 to Immunological Synapse	46	6
16.86	Innate Immune System	2132	39
16.48	TCR Signaling (Qiagen)	239	11
16.18	CD28 Co-stimulation	86	7
15.97	Epstein-Barr Virus Infection	203	10
15.27	ICos-ICosL Pathway in T-Helper Cell	131	8
14.62	IL-2 Pathway	327	12
14.54	MHC Class II Antigen Presentation	103	7
14.29	Cytokine Signaling in Immune System	761	19
14.25	Cell Adhesion Molecules (CAMs)	145	8
13.78	Phagosome	152	8
13.54	Malaria	49	5
13.54	Interleukin-10 Signaling	49	5
13.15	Th17 Cell Differentiation	162	8
12.82	Class I MHC Mediated Antigen Processing & Presentation	823	19
12.46	NF-kappaB Signaling	327	11
12.14	Immunoregulatory Interactions b/w Lymphoid & Non-Lymphoid Cell	135	7
12.10	O2/CO2 Exchange in Erythrocytes	13	3
11.89	Toxoplasmosis	139	7
11.76	African Trypanosomiasis	35	4
11.24	Allograft Rejection	249	9
11.00	HTLV-I Infection	255	9
10.53	Tacrolimus/Cyclosporine Pathway, Pharmacodynamics	44	4
10.26	TCR Signaling (REACTOME)	122	6
10.07	Amb2 Integrin Signaling	48	4
9.86	IL12-mediated Signaling Events	86	5
9.76	Tuberculosis	179	7
9.23	NF-kappa B Signaling Pathway	95	5

Table 2: Pathway results obtained from the analysis of the most discriminatory genes via GeneAnalitics©.

Score	Name	Total Genes	Matched Genes
25.61	Interferon-gamma-mediated Signaling Pathway	71	9
22.22	Antigen Processing & Presentation of Peptide or Polysaccharide Antigen Via MHC Class II	14	5
17.69	T Cell Costimulation	73	7
14.99	Antigen Processing & Presentation of Exogenous Peptide Antigen Via MHC Class II	98	7
14.42	Antigen Processing & Presentation	43	5

Table 3: GO biological process analysis obtained from the analysis of the most discriminatory genes via GeneAnalitics©.

Score	Name	Total Genes	Matched Genes
18.36	Haptoglobin Binding	3	3
18.20	MHC Class II Receptor Activity	11	4
17.12	Peptide Antigen Binding	29	5
11.79	Oxygen Carrier Activity	14	3
11.23	MHC Class II Protein Complex Binding	16	3
10.64	Protein Binding	9398	105
9.75	Chemokine Activity	51	4
9.55	Peroxidase Activity	24	3

Table 4: GO molecular function analysis obtained from the analysis of the most discriminatory genes via GeneAnalitics©.

Score	Name	Total Genes	Matched Genes
17.88	Hemolytic Anemia	26	5
16.16	Increased Aerobic Running Capacity	5	3
16.16	Hyperoxia	5	3
12.42	Autoimmune Response	31	4
12.22	Increased IgM Level	94	6
12.10	Abnormal Hemoglobin	13	3
10.78	Anisocytosis	42	4
10.69	Abnormal Mean Corpuscular Volume	4	2
10.52	Abnormal Aerobic Respiration	19	3

Table 5: Phenotype analysis obtained from the analysis of the most discriminatory genes via GeneAnalitics©.

Score	Name	Total Genes	Matched Genes
18.36	1D09C3	3	3
12.67	Sebacic Acid	2	2

Table 6: Compounds analysis obtained from the analysis of the most discriminatory genes via GeneAnalitics©.

3.2. Pathway analysis via the holdout sampler

In this section we perform the pathway analysis of the defective pathways in MS via the holdout sampler [34] consists in performing different random holdout simulations and finding the minimum-size signature of high discriminatory genes for each holdout. The holdout sampler determines for each holdout the small-scale genetic signature in the training dataset (75% of the total data) and its predictive accuracy is established using the validation dataset (25% of the total dataset). Both datasets are randomly generated in each holdout. The pathways analysis taking into account all the small-scale genetic signatures with high predictive validation accuracy in all the holdouts, performing the frequency analysis, and considering the most frequently sampled genes to establish the defective pathways.

Table 7 shows the list of 30 most frequently-sampled genes (genes that appears in more than 17% of random simulations). It can be observed that most of these genes appear in the list of genes that initially provided a 90% of accuracy in blind validation (see table 1): CAPRIN1 appears in position 14, CST3 in position 9, GAS7 in position 6, IRF8 in position 18. We have also performed a pathways analysis using the genes with frequency more than 10% (372 genes) obtaining very similar results using the list of genes that provided a 90% of accuracy, being the most important Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell, Interleukin-10 signaling, Interferon Signaling, ATF4 activates genes and TRAF6 mediated NF-kB activation. All of them appear in table 2.

Frequency	Name
0.17	CAPRIN1
0.17	CST3
0.17	GAS7
0.17	IRF8
0.17	BNIP1
0.17	CCL3 /// CCL3L1 /// CCL3L3
0.17	MXD1
0.17	HBB
0.17	SYNGR1
0.17	PLAUR
0.17	HBB
0.17	HBA1 /// HBA2
0.17	214349_at
0.17	HBA1 /// HBA2
0.17	HBB
0.17	BFAR
0.17	ZNF703
0.17	TP53INP2
0.17	STARD4
0.17	CUL4A
0.17	NAPSB
0.17	SLC43A2
0.17	ERCC6L2
0.17	TMEM45B
0.17	AC079305.10
0.17	BRD7
0.17	MIR6883 /// PER1
0.17	ITPRIP
0.17	HLA-DRA

Table 7: List of 30 most frequently-sampled genes using the holdout sampler.

4. Conclusions

In this paper, we have performed the analysis of the defective genetic pathways in Multiple Sclerosis using machine learning technologies and two independent microarray data-sets. The first one is used to learn the most important discriminatory genes, while, the second one, is restricted to perform blind validation of the genetic signatures that were previously found. In both cases, we have obtained a LOOCV predictive accuracy higher than 95%. The most important defective genes found are related to hemoglobin production and neurogenesis. The pathway analysis suggested antigen presentation as the main possible cause for MS and Epstein-Barr virus and Influenza A as possible agents. Other infections such as Toxoplasmosis, Tuberculosis, and Staphylococcus Aureus, appeared in the list of defective pathways. Future studies should compare these results with the ones found in Fibromyalgia and Post-radiotherapy Chronic Fatigue [35, 36] since they have similarities. We understand that the knowledge obtained from this analysis would serve to find new and effective therapeutic targets for this important disease which affects 2.5 million people worldwide.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. We would like to acknowledge Ms. Carolina deAndrés and Mr. Antonio Gavala for the English revision and style corrections.

References

References

[1] Hauser SL and Oksenberg JR. The neurobiology of multiple sclerosis: genes, inflammation, and neurodegeneration. *Neuron*, 52(1):61–76, 2006. ISSN 0896-6273 (Print); 0896-6273 (Linking). doi:10.1016/j.neuron.2006.09.011.

[2] Trapp BD and Nave KA. Multiple sclerosis: an immune or neurodegenerative disorder? *Annu Rev Neurosci*, 31:247–269, 2008. ISSN 0147-006X (Print); 0147-006X (Linking). doi:10.1146/annurev.neuro.30.051606.094313.

[3] Weiner HL. Multiple sclerosis is an inflammatory t-cell–mediated autoimmune disease. *Archives of Neurology*, 61(10):1613–1615, 2004.

[4] Rosati G. The prevalence of multiple sclerosis in the world: an update. *Neurological sciences*, 22(2):117–139, 2001.

[5] Alter M, Yamoor M, and Harshe M. Multiple sclerosis and nutrition. *Archives of neurology*, 31(4):267–272, 1974.

[6] Hayes CE, Cantorna MT, and DeLuca HF. Vitamin d and multiple sclerosis. *Proceedings of the Society for Experimental Biology and Medicine*, 216(1):21–27, 1997.

[7] Kragt J, Van Amerongen B, Killestein J, et al. Higher levels of 25-hydroxyvitamin d are associated with a lower incidence of multiple sclerosis only in women. *Multiple Sclerosis Journal*, 15(1):9–15, 2009.

[8] Pender MP and Burrows SR. Epstein-barr virus and multiple sclerosis: potential opportunities for immunotherapy. *Clin Transl Immunology*, 3(10):e27, 2014. ISSN 2050-0068 (Linking). doi:10.1038/cti.2014.25.

[9] Garcion E, Wion-Barbot N, Montero-Menei CN, et al. New clues about vitamin d functions in the nervous system. *Trends Endocrinol Metab*, 13(3):100–105, 2002. ISSN 1043-2760 (Print); 1043-2760 (Linking).

[10] Hanwell HEC and Banwell B. Assessment of evidence for a protective role of vitamin d in multiple sclerosis. *Biochim Biophys Acta*, 1812(2):202–212, 2011. ISSN 0006-3002 (Print); 0006-3002 (Linking). doi:10.1016/j.bbadis.2010.07.017.

[11] Manouchehrinia A, Tench CR, Maxted J, et al. Tobacco smoking and disability progression in multiple sclerosis: United kingdom cohort study. *Brain*, 136(Pt 7):2298–2304, 2013. ISSN 1460-2156 (Electronic); 0006-8950 (Linking). doi:10.1093/brain/awt139.

[12] O'Gorman C, Lucas R, and Taylor B. Environmental risk factors for multiple sclerosis: a review with a focus on molecular mechanisms. *International journal of molecular sciences*, 13(9):11718–11752, 2012.

[13] Willer C, Dyment D, Risch N, et al. Twin concordance and sibling recurrence rates in multiple sclerosis. *Proceedings of the National Academy of Sciences*, 100(22):12877–12882, 2003.

[14] Alcina A, Abad-Grau MDM, Fedetz M, et al. Multiple sclerosis risk variant hla-drb1*1501 associates with high expression of drb1 gene in different human populations. *PLoS One*, 7(1):e29819, 2012. ISSN 1932-6203 (Electronic); 1932-6203 (Linking). doi:10.1371/journal.pone.0029819.

[15] Gregory SG, Schmidt S, Seth P, et al. Interleukin 7 receptor alpha chain (il7r) shows allelic and functional association with multiple sclerosis. *Nat Genet*, 39(9):1083–1091, 2007. ISSN 1061-4036 (Print); 1061-4036 (Linking). doi:10.1038/ng2103.

[16] Lundmark F, Duvefelt K, Iacobaeus E, et al. Variation in interleukin 7 receptor alpha chain (il7r) influences risk of multiple sclerosis. *Nat Genet*, 39(9):1108–1113, 2007. ISSN 1061-4036 (Print); 1061-4036 (Linking). doi:10.1038/ng2106.

[17] Wang Z, Sadovnick AD, Traboulsee AL, et al. Nuclear receptor nr1h3 in familial multiple sclerosis. *Neuron*, 92(2):555, 2016. ISSN 1097-4199 (Electronic); 0896-6273 (Linking). doi:10.1016/j.neuron.2016.09.028.

[18] Hoppenbrouwers IA and Hintzen RQ. Genetics of multiple sclerosis. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1812(2):194–201, 2011.

[19] Liu M, Hou X, Zhang P, et al. Microarray gene expression profiling analysis combined with bioinformatics in multiple sclerosis. *Molecular biology reports*, 40(5):3731–3737, 2013.

[20] Corvol JC, Pelletier D, Henry RG, et al. Abrogation of t cell quiescence characterizes patients at high risk for multiple sclerosis after the initial neurological event. *Proc Natl Acad Sci U S A*, 105(33):11839–11844, 2008. ISSN 1091-6490 (Electronic); 0027-8424 (Linking). doi:10.1073/pnas.0805065105.

[21] Jernas M, Malmestrom C, Axelsson M, et al. Microrna regulate immune pathways in t-cells in multiple sclerosis (ms). *BMC Immunol*, 14:32, 2013. ISSN 1471-2172 (Electronic); 1471-2172 (Linking). doi: 10.1186/1471-2172-14-32.

[22] Fernández-Martínez J, Fernández-Muñiz Z, Pallero J, et al. From bayes to tarantola: New insights to understand uncertainty in inverse problems. *Journal of Applied Geophysics*, 98:62–72, 2013.

[23] Tusher V, Tibshirani R, and Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci*, 98(9):5116–21, 2001.

[24] Fisher RA. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(7):179–188, 1936.

[25] deAndrés Galiana EJ, Fernandez-Martinez JL, and Sonis ST. Design of biomedical robots for phenotype prediction problems. *J Comput Biol*, 23(8):1–15, 2016. ISSN 1557-8666 (Electronic); 1066-5277 (Linking). doi: 10.1089/cmb.2016.0008.

[26] Fernandez-Martinez JL, deAndres Galiana EJ, and Sonis ST. Genomic data integration in chronic lymphocytic leukemia. *J Gene Med*, 2016. ISSN 1521-2254 (Electronic); 1099-498X (Linking). doi:10.1002/jgm.2936.

[27] Saligan LN, Fernandez-Martinez JL, deAndrés Galiana EJ, et al. Supervised classification by filter methods and recursive feature elimination pre-

dicts risk of radiotherapy-related fatigue in patients with prostate cancer. *Cancer Inform*, 13:141–152, 2014. ISSN 1176-9351 (Electronic); 1176-9351 (Linking). doi:10.4137/CIN.S19745.

[28] Stelzer G, Inger A, Olender T, et al. Genedecks: paralog hunting and gene-set distillation with genecards annotation. *OMICS*, 13(6):477–87, 2009.

[29] Lastra G, Luaces O, Quevedo JR, et al. Graphical feature selection for multilabel classification tasks. In *International Symposium on Intelligent Data Analysis*, pages 246–257. Springer, 2011.

[30] Pearson K. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

[31] Kruskal JB. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, 7(1):48–50, 1956.

[32] Serrano-Fernández P, Ibrahim S, Zettl U, et al. Intergenomic consensus in multifactorial inheritance loci: the case of multiple sclerosis. *Genes and Immunity*, 5(8):615, 2004.

[33] Pereira RM, Martinez GJ, Engel I, et al. Jarid2 is induced by tcr signalling and controls inkt cell maturation. *Nature communications*, 5:4540, 2014.

[34] Fernández-Martínez JL, Cernea A, Fernández-Ovies FJ, et al. Sampling defective pathways in phenotype prediction problems via the holdout sampler. In *International Conference on Bioinformatics and Biomedical Engineering*, pages 24–32. Springer, 2018.

[35] Feng LR, Fernández-Martínez JL, Zaal KJ, et al. mglur5 mediates post-radiotherapy fatigue development in cancer patients. *Translational psychiatry*, 8(1):110, 2018.

[36] Lukkahatai N, Walitt B, Deandrés-Galiana EJ, et al. a predictive algorithm to identify genes that discriminate individuals with fibromyalgia syndrome diagnosis from healthy controls. *Journal of pain research*, 11:2981, 2018.

Figures

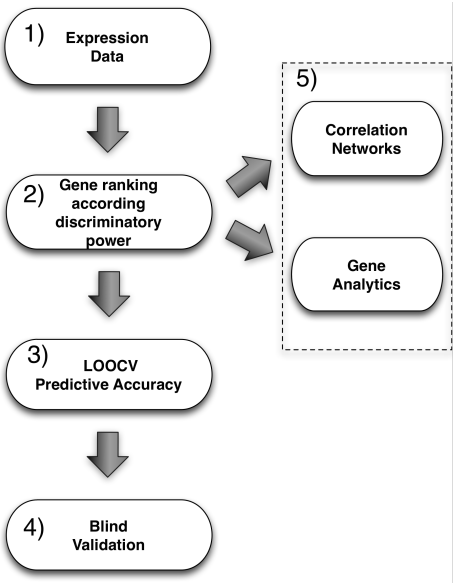

Figure 1: Flow chart of the methodology

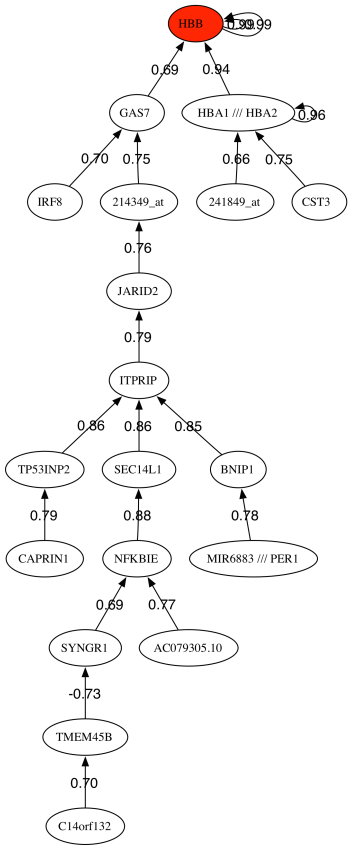

Figure 2: Correlation tree of the 31 most discriminatory genes.