Decoding genetic markers of multiple phenotypic layers through biologically constrained Genome-to-Phenome Bayesian Sparse Regression

Marie Deprez, Julien Moreira, Maxime Sermesant, Marco Lorenzi

frontiers

Decoding genetic markers of multiple phenotypic layers through biologically constrained Genome-to-Phenome Bayesian Sparse Regression

Marie Deprez [1*], **Julien Moreira** [1], **Maxime Sermesant** [1] **and Marco Lorenzi** [1]

[1] *University of Côte d'Azur, Inria, Epione Project-Team, Nice, France*

Correspondence*:
Corresponding Author
marie.deprez@inria.fr

ABSTRACT

The applicability of multivariate approaches for the joint analysis of genomics and phenomics information is currently limited by the lack of scalability, and by the difficulty of interpreting the related findings from a biological perspective. To tackle these limitations, we present Bayesian Genome-to-Phenome Sparse Regression (G2PSR), a novel multivariate regression method based on sparse SNP-gene constraints. The statistical framework of G2PSR is based on a Bayesian neural network, were constraints on SNPs-genes associations are integrated by incorporating *a priori* knowledge linking variants to their respective genes, to then reconstruct the phenotypic data in the output layer. Interpretability is promoted by inducing sparsity on the genes through variational dropout, allowing to estimate the uncertainty associated with each gene, and related SNPs, in the reconstruction task. Ultimately, G2PSR is conceived to prevent multiple testing correction and to assess the combined effect of SNPs, thus increasing the statistical power in detecting genome-to-phenome associations. The effectiveness of G2P was demonstrated on synthetic and real data, with respect to state-of-the-art methods based on group-wise sparsity constraints. The application on real data consisted in an imaging-genetics analysis on the Alzheimer's Disease Neuroimaging Initiative data, relating SNPS from more than 3500 genes to clinical and multi-variate brain volumetric information. The experimental results show that our method can provide accurate selection of relevant genes in dataset with large SNPs-to-samples ratio, thus overcoming the main limitations of current genome-to-phenome association methods.

Keywords: Bayesian, Variational Dropout, Genome, Phenome, Regression, Biological constraint

1 INTRODUCTION

Multi-omics data integration is an ever-growing field at the crossroad between biology and statistical learning. The goal of multi-omics analyses is to reveal novel insights on complex biological systems, through the combined analysis of multiple data types (4). Multi-omics data approaches are often designed to account for phenotypic information, providing quantitative features about the clinical or biological condition of an individual. Imaging-genetics is a typical application domain, in which genomic data under the form of Single-Nucleotide Polymorphisms (SNPs) are jointly analyzed with medical imaging information, composed of high dimensional imaging quantitative traits (26). The analysis of these complementary

data types has proven quite successful to improve the discovery of genetic risk factors of complex and rare diseases, for example when applied to age-related macular degeneration, obesity, schizophrenia and Alzheimer's Disease (29). Nevertheless, current approaches to jointly analyze genomic and phenotypic information face multiple limitations due to the inherent complexity and dimensionality of genetic and imaging data.

The basic form of analysis between genomics and phenotypic information is genome-wide association studies (GWAS) (28), based on mass-univariate testing of association between SNP and phenotypic features. Multiple comparison correction methods are generally used to mitigate the risk of false discovery, at the expense of potentially reduced detection power (17). To compensate for the large data dimensionality, the number of association tests can be reduced by aggregating either the genomic or phenotypic features (33). An intrinsic limitation of GWAS is the relatively small effect size of SNPs on the phenotypic features. Most SNPs account for under 1% of the variance in brain-imaging quantitative traits when considered individually. This aspect motivates the development of strategies combining the effect of multiple SNPs to increase the detection power. This combination can be inspired by known biological associations (Linkage Disequilibrium blocks) or structures (genes) and thus significantly reduce the dimensionality and improve the statistical power of the genomic data (9, 6). To this end, more recent approaches have been proposed based on different modeling rationale and complexity, from multivariate regression (30, 7) to deep learning methods (32).

For example, neural networks architectures allow the identification of compressed representation of the data in the bottleneck layers, potentially allowing more accurate integration of the complex interaction between high-dimensional features (24). For instance, Wang et al. (32) proposed an additive model via Feed-forward Neural networks with random weights to assess the role of each genetic feature independently in the prediction. Yet, neural network architectures require optimising a large number of parameters, which thus negatively affects the reliability of these approaches in studies with low sample size and/or when the number of genetic features analysed is (too) large. In an attempt to mitigate this limitation, several methods have been proposed to induce sparsity in the model parameters and their associated features (34, 7, 35, 27). The study (31) presented a Group-Sparse Multi-task Regression and Feature Selection method structured as a sparse model based on $l_{2,1}-$norm regularization, to perform feature selection at both the group/gene-level and the SNP level. This kind of strategy improves the interpretability of the model results, as non-relevant SNPs and their associated genes can be readily pruned by the associated parameters. However, this kind of approaches come with an increased computational burden, which typically limits the number of genes and SNPs that can be analysed. Scaling to a large number of genes without pre-selection remains a challenge in Genome-to-Phenome association studies, and most of current applications have been demonstrated on dataset composed of up to few hundreds of genes (31, 32, 34, 7, 8).

All in all, the applicability of current multivariate genome-to-phenome modeling approaches is limited by the lack of scalability, and by the difficulty of interpreting the related findings from a biological perspective. To tackle these limitations, in this paper we present Bayesian Genome-to-Phenome Sparse Regresion (G2PSR), a novel multivariate regression method based on sparse SNP-gene constraints. G2PSR is conceived to alleviate the need for multiple testing correction and considers SNPs combined effect, thus increasing the analysis detection power. The statistical framework of G2PSR is based on a Bayesian neural network, were constraints on SNPs-gene associations are integrated by encoding our knowledge on SNPs mapped in the same transcribed region, exons and introns, and promoter region to reduce genomic data dimensionality. This constraint maps the input SNPs into the corresponding genes represented in the intermediate layer of the network. We induce sparsity on the genes through variational dropout, to estimate

the uncertainty associated with each gene (and related SNPs) in reconstructing the phenotypic features (output layer). The effectiveness of G2PSR was demonstrated on synthetic and real data, with respect to state-of-the-art methods based on group-wise sparsity constraints. The application on real data consisted in an imaging-genetics analysis on the Alzheimer's Disease Neuroimaging Initiative (ADNI) data, relating SNPs from more than 3500 genes to clinical scores and multi-variate brain volumetric information. The experimental results show that our method can provide accurate selection of genes-of-interest in a dataset with large SNPs-to-samples ratio, thus overcoming the main limitations of current genome-to-phenome association methods.

2 MATERIALS AND EQUIPMENT

In this section we describe synthetic and real data used to develop, optimize and test the our Bayesian Genome-to-Phenome Sparse Regresion (G2PSR) framework presented in the Methods section. We analyzed genomic data, such as genetic variations under the form of single nucleotide polymorphisms (SNPs), and phenotypic data, such as brain volume measurements or clinical examination scores. To test and develop our proposed method, we first generated synthetic data mimicking real genomic and phenotypic data with controlled genome-to-phenome associations (section 2.1). We then applied our method in a imaging-genetics case study in Alzheimer's Disease, using data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (section 2.2).

2.1 Synthetic Data

Synthetic datasets were generated to mimic the properties of genomic and phenotypic data observed in real cases, with controlled genome-to-phenome associations. To evaluate G2PSR and benchmark it against state-of-the-art methods, we implemented a simulation system to generate pseudo-genomic data and associated phenotypic features. The pseudo-genomic data was generated to reproduce SNP information represented by multivariate arrays with entries 0, 1, or 2 corresponding to the number of alternative alleles found at the SNP location. The phenotypic data was subsequently defined to represent any discrete or continuous biological metric describing a phenotype, e.g. cognitive test results and volume measurements of different brain areas.

We denote by N the total number of SNPs, by G the total number of genes, by T the number of phenotype features, and by S the total number of samples to be generated. We define the $S \times N$ genotype matrix $\boldsymbol{X}$, where each row is partitioned in segments ($\boldsymbol{X_g}$) representing $g = 1, ..., G$ genes. The elements $X_{g,i}$ represent the SNP i mapped to the gene g generated to match the expected allele frequency observed in real data: $X_{g,i} \sim MD(p_0)$, where MD is a multinomial distribution parameterized by $p_0 = (p(x = 0), p(x = 1), p(x = 2))$. We estimated this distribution using $\sim 390 \times 10^6$ SNP values from the ADNI genetic dataset.

The $S \times T$ phenotype matrix $\boldsymbol{Y}$ is composed by T phenotypic features, obtained by concatenating genome-associated phenotype $\boldsymbol{Y_G}$, and uncorrelated phenotype $\boldsymbol{Y'}$: $\boldsymbol{Y} = [\boldsymbol{Y_g}, \boldsymbol{Y'}]$. We define the genome-to-phenome association through a linear transformation, $\boldsymbol{Y_g} = \boldsymbol{XV}$, where each column j of the association matrix $\boldsymbol{V}$ has non-zero entries at the positions of the n_g SNPs associated to the gene g related to the phenotype j. The non-zero entries are randomly sampled from a Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{Id})$, while the uncorrelated phenotype features are random $\boldsymbol{Y'} \sim \mathcal{N}(\boldsymbol{\mu_{Y'}}, \boldsymbol{\Sigma_{Y'}})$. The output phenotype is finally obtained as $\boldsymbol{Y''} = \boldsymbol{Y} + snr^{-1/2} \cdot \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is standard Gaussian noise modulated by the desired signal-to-noise ratio snr. To determine the number of SNPs associated to each generated gene we estimated

a distribution of the number of SNPs per gene rom real genomic data. We used the number of SNPs in the 23'952 genes found in the ADNI genetic dataset.

2.2 ADNI Dataset

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Up-to-date information is available at www.adni-info.org. We selected clinical, genotypic and phenotypic data available in the ADNI-1/GO/2 datasets for 808 subjects.

The genetic data was filtered and processed chromosome by chromosomes using multiple tools (vcftools (5), bedtools (22) and PLINK 2.1 (25, 23)). We only selected Single Nucleotide Polymorphisms annotated in dbSNP (the NCBI database of genetic variation). We mapped them to exonic gene regions using Homo Sapiens GFF annotation files (version GRCh.37, used by ADNI to map genetic variants). Variants were then filtered by their Minor Allele Frequency (MAF > 0.05). Missing data in the VCF files were imputed using the Sanger Imputation Server followed by post-imputation quality controls (imputation quality score, MAF and Hardy-Weinberg equilibrium). Lastly, SNPs with strong linkage disequilibrium (LD > 0.8) were filtered out, and all chromosomes were merged in a single file of annotated SNPs per sample. We obtained a genetic matrix of 485101 SNPs grouped into 23938 genes for 808 patients/samples. For our use case, we selected genes associated to the KEGG pathways (composed of 180 pathways including one focused on Alzheimer's Disease). We obtained a genomic data matrix composed of 104 854 SNPs grouped into 3953 genes.

The phenotypic data includes both clinical and brain volumes measurements. The clinical data is composed of six continuous variables generally recorded in memory clinics: the Alzheimer's Disease Assessment Scale (ADAS) Cognitive Subscale (COG), Clinical Dementia Ratings (CDR), the Mini-Mental State Examination (MMSE), the Functional Activities Questionnaires (FAQ) and the Rey Auditory Verbal Learning Tests (RAVL immediate and forgetting). Imaging data were processed from structural MRI (grey matter only) to provide volume data on selected areas: Hippocampus and Entorhinal cortex. We therefore obtained a phenotypic matrix composed of 8 features matching the genetic data for 491 samples. Samples clinical status and demographics are described in Table 1.

3 METHODS

In this section, we describe the theoretical framework of G2PSR (Section 3.1). In Section 3.2, we present the experimental design used to optimize G2PSR, followed by the description of our benchmark experiments including the synthetic scenarios tested, the state-of-the-art methods used for comparison with G2PSR, and the performance metric used to evaluate their respective accuracies.

3.1 Bayesian Genome-to-Phenome Sparse Regression

Our approach consists in a multivariate regression framework designed to predict multivariate phenotypic data from large arrays of SNP information. As the number of SNPs in GWAS is often an order of magnitude larger than the number of available samples, G2PSR is designed to account for biologically inspired

constraints, in which known functional relationships across SNPs are accounted for under the form of group-wise sparsity penalization. The group-wise sparsity relationship across SNPs is designed to associate each SNP to its related gene, either within its transcribed region or in its regulation range (Fig. 1.A). This constraint defines the G2PSR network architecture from the input layer to the biologically constrained intermediate layer, that is finally transformed to reconstruct the output phenotypic layer (Fig. 1.B).

Let $\boldsymbol{X}$ the $S \times N$ input matrix representing the N SNPs studied for the S subjects. Each row of the matrix is partitioned in segments $\boldsymbol{X^g}$, grouping the set of SNPs associated to the gene g. Let's denote by A^g the ensemble of SNPs indices associated to gene g, $\boldsymbol{X^g} = \{x_i, i \in A^g\}$, $\forall g \in (1, ..., G)$.

Based on the structure defined in Figure 2, for each gene we define a $N \times G$ linear transformation $\boldsymbol{W}$ mapping the input data to the gene representation, $\boldsymbol{L} = \boldsymbol{X} \cdot \boldsymbol{W}$. The elements of the matrix $\boldsymbol{W}_{i,g}$ are non-null if $i \in A^g$, and map the corresponding SNPs into the group-wise gene representation. In what follows, we denote by $\boldsymbol{W}^g$ the non-null elements of each of the G columns of $\boldsymbol{W}$. The representation $\boldsymbol{L} = (\boldsymbol{l^1}, ..., \boldsymbol{l^G})$ corresponds to the intermediate/gene layer of our network.

Finally, we assume that the phenotype is conditioned by the input layer through the likelihood :

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{L} \cdot \boldsymbol{V}, \Sigma), \tag{1}$$

where $\boldsymbol{V}$ is a $G \times T$ matrix allowing the reconstruction of the T phenotypic features from the gene layer $\boldsymbol{L}$, and Σ is the variance of the observational noise.

The parameters to be optimized in our model are the SNP-to-gene transformation $\boldsymbol{W}$, and the gene-to-phenotype transformation $\boldsymbol{V}$ or, in a more compact form, $\boldsymbol{\theta} = \{\boldsymbol{W}, \boldsymbol{V}\}$. The solution of the inference problem can be obtained by maximizing the marginal likelihood $p(\boldsymbol{Y}|\boldsymbol{X})$ with respect to the distribution of $\boldsymbol{\theta}$, which is usually an intractable problem. In this case, we can apply variational inference to learn a variational approximation $q(\boldsymbol{\theta}) = \{q(\boldsymbol{W}), q(\boldsymbol{V})\}$ to obtain an approximate posterior (1):

$$\begin{aligned} log(\boldsymbol{Y}|\boldsymbol{X}) &= log \int_{\boldsymbol{\theta}} p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{X}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &\geqslant \underbrace{\mathbb{E}_{q(\boldsymbol{\theta})} log\left(p(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{X})\right)}_{\text{A}} - \underbrace{KL(p(\boldsymbol{\theta}) \| q(\boldsymbol{\theta}))}_{\text{B}}. \end{aligned} \tag{2}$$

Equation (2) represents the evidence lower bound (ELBO) associated to the inference problem (2). Accordingly, the optimization of the marginal likelihood can be solved by optimizing jointly the reconstruction term A with respect to the variational distribution $q(\boldsymbol{\theta})$, and the regularization term B represented by the Kullback Leibler divergence between the variational distribution and the prior $p(\boldsymbol{\theta})$. This problem can be efficiently solved by using standard optimization techniques based on backpropagation and reparameterization trick (12).

To introduce biologically inspired constraints in the proposed G2PSR framework, the parameterization $\boldsymbol{\theta}$ is specified in the following section.

3.1.1 Regularization via variational dropout

To incorporate biological constraints in our framework, inspired by the seminal works (9, 19, 31, 8), we impose group-wise penalization to the weights $\boldsymbol{W^g}$ mapping the input SNPs to the common gene. The idea is that during optimization the model is forced to jointly discard all the SNPs mapping to genes which are not relevant to the predictive task, by setting to zero the associated parameters. To this aim, coherently with the variational approach detailed in section 3.1, we extend variational dropout as a group-wise regularization technique. Following (20), we parameterize the variational approximation $q(\boldsymbol{W^g})$ such that each element $W_i^g \sim \mathcal{N}(\mu_i^g; \alpha_g.\mu_i^{g2})$ (12), where the parameter α_g is optimized to quantify the common uncertainty associated with the ensemble of SNPs contributing to the gene g.

While the formulation is general, in what follows we restrict variational inference to the parameters $\mathbf{W}$ only, while optimizing the parameters $\mathbf{V}$ through maximum likelihood.

3.1.2 Variational dropout and sparsity

According to Molchanov et al. (20), the corresponding Kullback-Leibler divergence compatible with the proposed variational parameterization can be approximated by:

$$D_{KL}(q(W_i^g), q(W_i^g)) \approx -k_1\sigma(k_2 + k_3\ln(\alpha_g)) + 0.5\ln(1 + \alpha_g^{-1}) + k_1, \tag{3}$$

where $k_1 = 0.63576$, $k_2 = 1.87320$, $k_3 = 1.48695$ and $\sigma(.)$ is the sigmoid function.

During the optimization of D_{KL} the sparsity arises naturally as it keeps α_g large for the less relevant features associated to larger uncertainty, while minimizing its value for the most relevant ones. Therefore, α_g is inversely proportional to the relevance of the gene in reconstructing the phenotypic features.

3.2 Synthetic experiments

We used synthetic genome-to-phenome datasets to study the behaviour and evaluate the accuracy of G2PSR. We designed a 'reference' dataset with fixed parameters described in Table 2, and produced multiple testing scenarios based on variants of this reference scenario, obtained by modifying a single parameter of interest while keeping the others fixed, Table 3. Each scenario was generated with ten replicates to assess the variability of the results.

3.2.1 G2PSR optimization

As described in the previous section, G2PSR is a Bayesian neural network using a biologically inspired constraint to reduce the number of genomic features to consider compared to the dataset sample size. However, in most cases, genomic data's dimensionality still far exceeds the number of samples available and hence the number of associated parameters to estimate in the model. According to Li et al. (14), G2PSR can be considered an over-parametrized regime and has thus the capacity to potentially (over-)fit any set of phenotypic features. To prevent such behaviour of the model, we define an optimal strategy to identify the number of iterations needed by the model to achieve the minimum while preventing overfit. This strategy is based on the control of key optimization parameters, such as training loss and estimated noise variance, to identify an early stopping strategy and optimize its performance.

Inspired by the heuristics described by Li et al. (14), we control the evolution of the training loss and estimated noise variance during the optimization process (Figure 3.A.B). The idea is that over-optimization

of the loss is characterized by a consequent decrease of the estimated variance of the observational noise σ, thus pointing to over-fitting. These two metrics are thus complementary to identify model over-fitting.

To illustrate this strategy, Figure 3 provides an example of training on synthetically generated data. We applied G2PSR on the synthetic reference dataset described earlier (Table 2). Additionally to the loss and reconstruction error (Figure 3.A.B), we also studied our model's accuracy to identify the phenotype-relevant genes (Area Under the Precision-Recall Curve, AUPRC) and the corresponding α_g parameters which serve to quantify the genes uncertainty to reconstruct the phenotypic features (Figure 3.C.D). From this example we observe that, after 25000 epochs, over-fitting appears under the form of decrease in both training loss and noise variance. We observed that at 23 000 epochs G2PSR reaches its top accuracy, which later decreases when further optimizing beyond 35 000 epochs. From the analysis of the α_g parameters we determine that, before the decrease of the loss and σ values ($\sim$ 20000 epochs), the model correctly identifies the relevant features (i.e. $p(\alpha_g)$ of significant genes) essentially ignoring the noisy ones (Figure 3.D).

3.2.2 Benchmark experiments

We evaluated the performance and accuracy of G2PSR when compared to two state-of-the-art methods: Sparse Group Lasso (SGL) (27), and Bayesian Group Sparse Multi-Task Regression (BGSMTR) (8) (31). We selected these methods to obtain a comparison with respectively a general sparse regression approach (SGL), and with a specifically designed method to relate genetic information with phenotypic data using biologically inspired constraints (BGSMTR). Similarly to G2PSR, we optimized the parameter settings of each method to retrieve the best performance in each tested scenario.

SGL is an extension of the group Lasso regulariser inducing parameter-wise sparsity (sparsity both at the covariates and at the group level). In our setting we used the SNPs as covariates and their related genes as a group associated with the phenotypic data of interest. As SGL can only relate the genomic data to a single phenotypic feature at a time, we used the average accuracy metric obtained across all phenotypic features tested. We optimized SGL using cross-validation with different covariates and group-wise regularization/sparsity terms (l_1 and l_2 norms), using the optimal accuracy metric among all tested parameter combinations.

BGSMTR was designed to produce confidence estimates on the regression parameters in addition to the $l_{2,1}-norm$ penalty at both SNP and gene-level. Using a multivariate prior based on Gaussian scale mixture, the method is based on Markov Chain Monte Carlo (MCMC) to obtain the posterior distribution and interval estimates. To select potentially significant SNPs, the authors suggest evaluating the 95% equal-tail credible interval for each regression coefficient and selecting those SNPs and their related genes where at least one of the associated credible interval excludes 0. BGMSTR can relate multiple phenotypic features and produce a regression coefficient for each SNPs along with confidence intervals. As the optimization of so many parameters is costly in computation time and resources, we only tested BGMSTR with all phenotypic features being associated with the genome. We optimized BGSMTR using the cross-validation method proposed by the authors.

For each testing scenario, we set a limit of 72h of computation time per method, including the parameters optimization (SGL and BGSMTR $l_{1,2}-$norms).

3.2.3 Accuracy metrics

We evaluated and compared the accuracy of each method in identifying the phenotype-relevant genes using multiple classification metrics. Considering that the number of SNPs and genes relevant to the

phenotype is relatively small compared to all the SNPs evaluated, we relied mainly on the Area Under the Precision-Recall curve (AUPRC) and the F-measure. However, we used the F-measure as a reference accuracy metric for the BGSMTR method since only the binary classification of the relevant SNPs and their associated genes was available.

4 RESULTS

This section describes the experimental results of our benchmark of G2PSR on extensive synthetic experiments, and on the real data from ADNI.

4.1 G2PSR Benchmark

We evaluated the computational complexity of G2PSR against an increasing number of genes and SNPs. We then assessed the ability of G2PSR to associate genomic data with multiple phenotypic features, and finally we compared its accuracy in multiple synthetic experiments (Section 3.2).

4.1.1 G2PSR computational complexity

We evaluated the computational complexity of G2PSR by measuring the computation time required to run each tested scenario using a GPU node (DELL T630 GPU node, GeForce GTX 1080 Ti GPU), while using 50000 number of epochs. We observed the most significant variation in computation time with respect to the total number of processed genes (Figure 4.A), compared to other tested parameters (Supplementary Table S1). We noted a linear relationship between the computation time required and the total number of genes (and SNPs) analyzed. We estimated that, on average, the analysis of a thousand genes (~ 20000 SNPs) takes 12 hours using our GP unit.

4.1.2 Genome-to multiple phenotypic layers accuracies

We evaluated the performance of our algorithm to associate genotypic data with genetically-relevant and genetically-independent phenotypic features in mixed proportions respectively (Y_g and Y' in section 2.1), see Figure 4.B. We noted that with a single phenotypic feature, G2PSR reaches a mean AUPRC of 0.60. When increasing the number of phenotypic features studied, the AUPRC improves even in the most challenging case with only 20% genetically relevant target features. This result can be explained by the redundancy induced by the relevant phenotypic features, wich allow to better identify causal genes. This result is encouraging, as most phenotypic features available for multi-omics data integration study are well-curated datasets with strong biological *a priori* in the selection of phenotypic features.

4.1.3 Comparison with SGL and BGSMTR

We compared the performance of three group-wise genome-to-phenome methods applied to synthetic scenarios (Tables 2,3 and Figure 5). In the case of G2PSR, we evaluated its accuracy when considering 100% (G2PSR_{100}) or 25% (G2PSR_{25}) of gene-related phenotypic features, compared to SGL and BGSMTR, which were only tested by considering 100% relevant phenotypic features.

4.1.3.1 Noise level

We observe that BGSMTR is the most sensitive method when faced with increasingly noisy phenotypic data, compared to SGL, which uses only one phenotypic feature and still maintained a better F-score ($\sim +0.15 - 0.25$), see Figure 5.A and Supplementary Table S2. G2PSR_{25} has a lower performance than the other methods in low noise level scenarios (< 0.5) but performs slightly better than SGL when the

noise level is above 0.8 (~ 0.15). Overall, G2PSR$_{100}$ out-performs all the compared methods in most tested scenarios (noise level $< 150\%$).

4.1.3.2 Number of samples

We observe that G2PSR$_{100}$ out-performs all compared methods with generally less number of samples required (minimum ~ 100 in the tested scenarios), see Figure 5.B and Supplementary Table S3. SGL performs on average slightly better than G2PSR$_{25}$ ($\sim 0.05 - 0.20$) with the same number of samples. BGSMTR remains the most sensitive method and required at least 500 samples to achieve similar efficiency than other comparison methods. We note that the method was still too costly to perform the optimization with an increasing number of samples (> 1000 samples).

4.1.3.3 Number of target genes

We evaluated the number of relevant genes correctly identified among all the genes tested. All methods have an optimal performance when the number of genes to identify is small (Figure 5.C and Supplementary Table S4). However, as the number of significant genes increases, we observe a decrease in the overall performance across all tested methods. As previously observed, BGSMTR is the method for which performance decreases the most, while SGL still performs slightly better than G2PSR$_{25}$ and G2PSR$_{100}$, and is the method that maintains the highest performance > 0.75. We hypothesize that most genome-to-phenome methods are well suited for analysing on unbalanced data, with few relevant genes to identify.

4.1.3.4 Total number of genes

We observe that there is minimal variation in performance among all tested scenarios and also between G2PSR and SGL (BGSMTR was not evaluated here as it was too time-consuming to optimize), see Figure 5.D and Supplementary Table S5. All methods have an average performance over 0.9 with a small decrease as the number of genes increases.

4.2 Real data application: Alzheimer's Disease use case

We analyzed genomic, imaging and clinical data from 491 samples of the ADNI cohort. The analysed genomic data consisted of 104 854 SNPs grouped into 3 953 genes. Imaging and clinical data are used as phenotypic features, and are respectively composed of two-volume measurements (hippocampus and entorhinal cortex) and six cognitive scores (Materials section, 2.2). To ensure that the optimization process did not lead to local minima, we applied G2PSR 10 times with 50000 epochs to obtain average and standard deviation values for each parameter used in the analysis (loss value, σ and α_g parameters), see Supplementary Figure S1.

Following the early stopping strategy described above, we analyzed the sparsity parameters α_g at 36000 epochs, Figure 6.A. The distribution of the α_g values at the selected epoch shows a bimodal trend (Fig 6.B). One mode of this distribution corresponds to minimal α_g values (< 0.05, red dotted line), thus pointing to potentially relevant genes associated with the phenotype. From this analysis, we identified 177 genes with an α_g value lower than the nominal cut-off of 0.05 (Supplementary Table S6). Interestingly, the main known genetic risk-factor associated with AD, the gene APOE, was ranked 10^{th} by α_g value in this experiment. This set of genes was used to filter the gene list and SNP data to perform a second optimization G2PSR round. Similarly to the previous analysis, we identified our early stopping strategy at 20000 epochs and analyzed the corresponding sparsity α_g values of the 177 genes here considered (Figure 7.A). We also performed a correlation analysis between each gene, their associated SNPs, and the ensemble of phenotypic features (Figures 7.B.C. and Supplementary Table S7). We observe a negative correlation ($R^2 = 0.353$)

between the average gene-phenotype correlation and their respective α_g values obtained with G2PSR. In particular, top genes with a low α value have at least one associated SNPs with a notable correlation to the phenotypic features. The most relevant example is the APOE gene (red dots in Fig 7.B.C), which is the top relevant gene and has two SNPs (rs429358, rs769449) with a correlation over 0.2 with phenotypic features (Table 4). Among the other top genes identified, some have not yet been well characterized as associated with the Alzheimer's Disease but are described as involved in brain functions. For instance, the Methionine Adenosyltransferase 2B, MAT2B gene has been associated with reduces ischemic brain injuries in rat (13). The protein transporter THOC3,the NKD2-WNT signaling pathway inhibitor and the PI3-kinases , PIK3C2B, have been described in dysregulated pathways due to the APOE$-\epsilon 4$ toxicity in Human neurons (21, 15). We also identified genes such as PTPN11 (also known as protein tyrosine phosphatase SHP2), and MCM2, involved in cell maintenance and renewal, which have been characterized as risk factors for Alzheimer's Disease (10, 3).

As the correlation per SNP with the phenotypic features rarely exceeds 0.1, this analysis suggests that the grouping constraint of G2PSR mapping SNPs into genes increases the detection power of our method to identify phenome-relevant genes.

5 DISCUSSION

This study proposes G2PSR, a Bayesian neural network accounting for biologically inspired constraints to provide Genome-to-Phenome sparse regression. Through the extensive benchmarking of our method on synthetic and real experimental scenarios, we demonstrate that G2PSR overcomes critical limitations of classical genome-to-phenome analyses. Through the SNPs-genes constraint, our method improves detection power, and reduces the number of parameters to optimize (thus minimizing the risk of overfitting) while simultaneously providing an interpretable neural network architecture through group-wise sparsity constraint based on a sound Bayesian framework.

Scalability is one of the main features of G2PSR, as most of published application cases have been currently limited to less than a thousand genes. For example, Zhu et al. (34) proposed a Bayesian generalized low-rank regression (GLRR) applied to a dataset composed of 1071 SNPs from 40 AD genes, while Lu and al. (19) extended the above-mentioned GLRR model and applied it on the same dataset. In the study of Wang et al. (31), the proposed model was applied to study 3123 SNPs from 153 AD candidate genes. Our results show that sespite the limited number of samples (similar to previous works), our method is able to analyse a far larger number of SNPs, while still providing relevant genome-to-phenome associations. By simultaneously integrating multiple phenotypic traits under a sparsity constraint, our application of G2PSR on ADNI demonstrates that our method can effectively process an extensive number of genes and SNPs successfully identifying known genes associated to Alzheimer's Disease (e.g. APOE (18), PTPN11 (11), PIK3C2B (16)). Our method identified also genes for which the association with AD has not been demonstrated, but for which some SNPs have a notable correlation with phenotypic features in the studied dataset (LIST).

A limitation of our approach is that it does not allow to directly identify the actual SNPs within each gene that drive the selection through the sparsity metric. To provide this missing information, in the real analysis scenario we implemented a two-step workflow to first screen among a large number of genes for relevant genome-to-phenome association, and then adopt a second level analysis to identify the most relevant genes overall. Our results demonstrate that G2PSR achieves an optimal selection of relevant genes associated with the phenotype, which can be completed with more targeted GWAS to better characterize the relevant SNPs

among the selected genes. The use of G2PSR ultimately circumvents the multiple comparison problem typical of standard GWAS applications.

Our analysis reveals several directions of improvement that could be implemented in future work. While reducing the number of genetic covariates, the gene constraint does not eliminate the correlation between SNPs close to each other on the genome through linkage disequilibrium (LD), and thus the correlation between their related genes. An improved group-wise constraint could account for the correlation between SNPs to produce relatively independent groups, for example by considering a mix of genes and LD blocks. Regarding the multi-omics analysis rationale,it would be relevant to introduce another data level into the G2PSR framework. For instance, gene expression data, such as from RNA-sequencing, could be used as an additional constraint in the neural network architecture, or as a prior on the weights in our variational scheme. Such an approach could improve the selection of genes relevant to the phenotype, potentially reduce the burden of the optimization process, and improve the interpretability of the model from a more mechanistic point of view.

To conclude, our study shows that G2PSR is an effective tool to identify genetic correlates of phenotypic features in the high-dimension/low-sample size regime, and can thus be employed in future challenging applications, such as imaging-genetics and rare disease analysis.

6 NOMENCLATURE

6.1 Figures

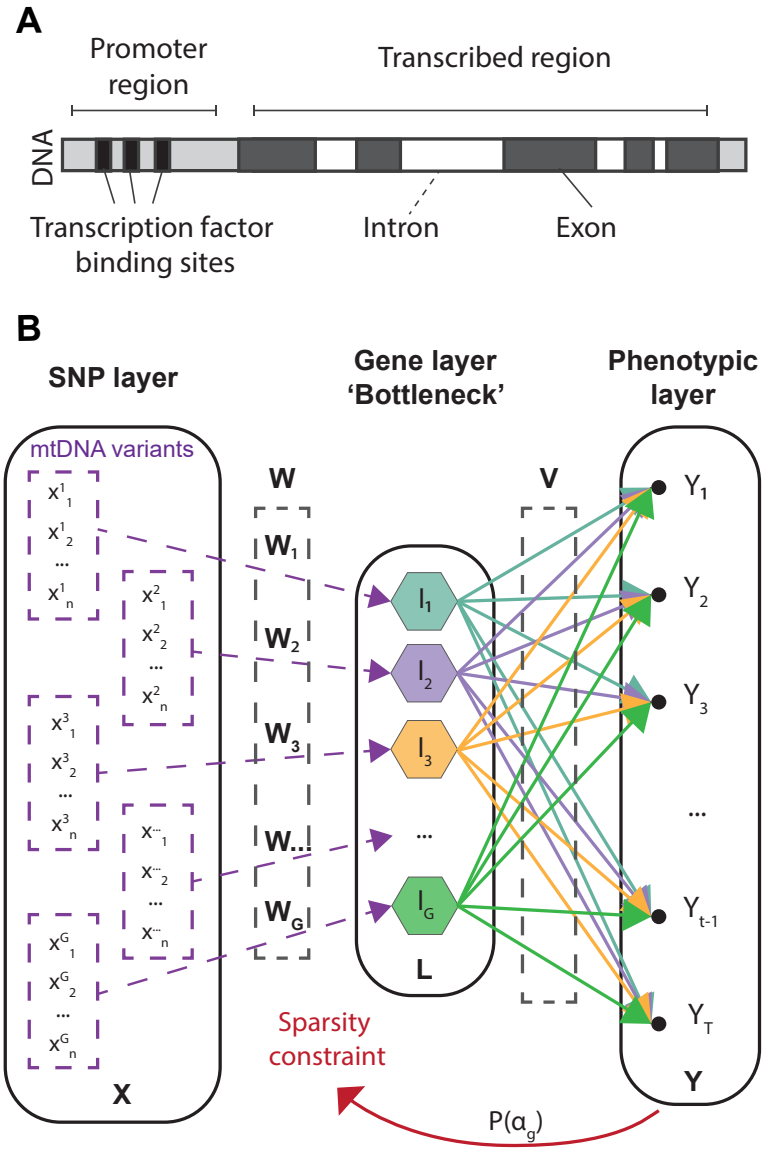

Figure 1. Genome-to-Phenome Sparse Regression architecture and description. **(A)** Gene structure describing the simplified constraint grouping SNPs into their associated genes, **(B)** G2PSR architecture from the input (SNP) layer to the constraint (Gene) layer and the output (Phenome) layer.

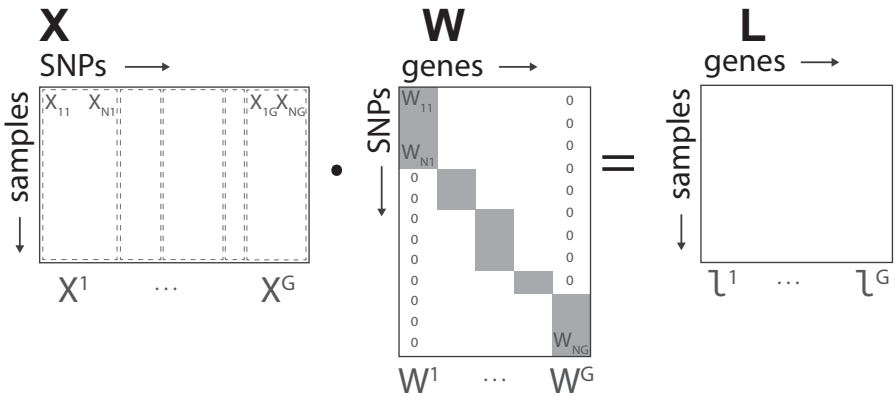

Figure 2. Genome-to-Phenome Sparse Regression architecture, detailed data structure for G2PSR generative model using the SNP-gene grouping constraint.

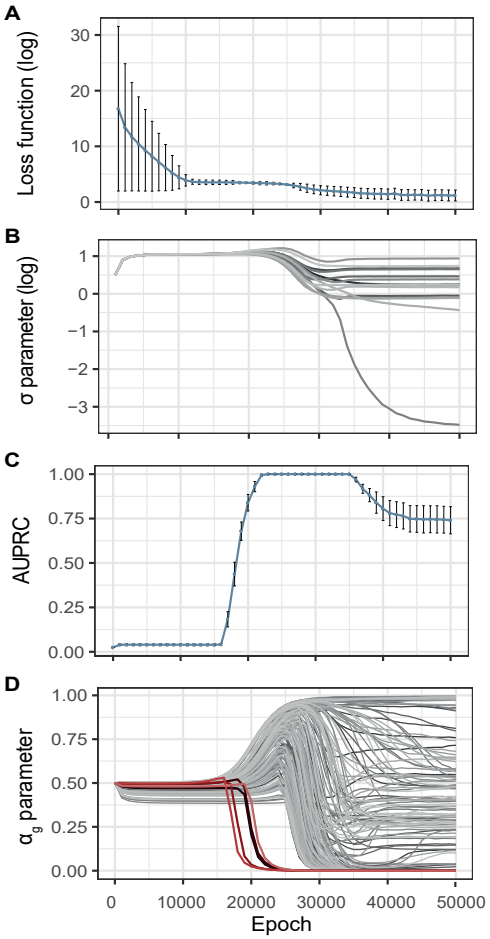

Figure 3. Evolution of the G2PSR parameters during the optimization process (Epoch/Iterations). **(A)** Loss value, **(B)** σ parameter, estimated noise variance, used to reconstruct the output ($\boldsymbol{Y}$), each line correspond to a phenotypic feature, **(C)** α_g parameter per gene (sparsity metric), red lines correspond to relevant genes, grey lines correspond to non-relevant genes, **(D)** Evolution of the Area under the Precision-Recall Curve (AUPRC) estimated at the gene level using the α_g parameter. Error-bars show the standard deviation at each measured epoch (10 replicates)

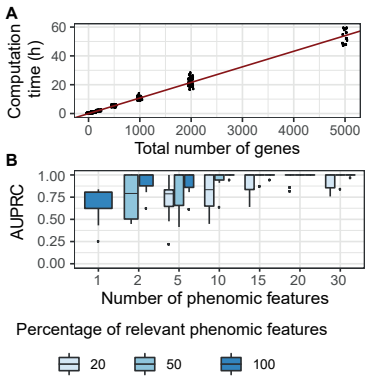

Figure 4. G2PSR model performance in specifically designed scenarios. **(A)** Computational time of G2PSR with respect to the total number of genes analysed. **(B)** Area Under the Precision-Recall Curve of our G2PSR model depending on the number of phenotypic features analysed and the percentage of genome-related phenotypic features among all features analysed ($\boldsymbol{Y} = [\boldsymbol{Y_g}, \boldsymbol{Y'}]$, see section 2.1)

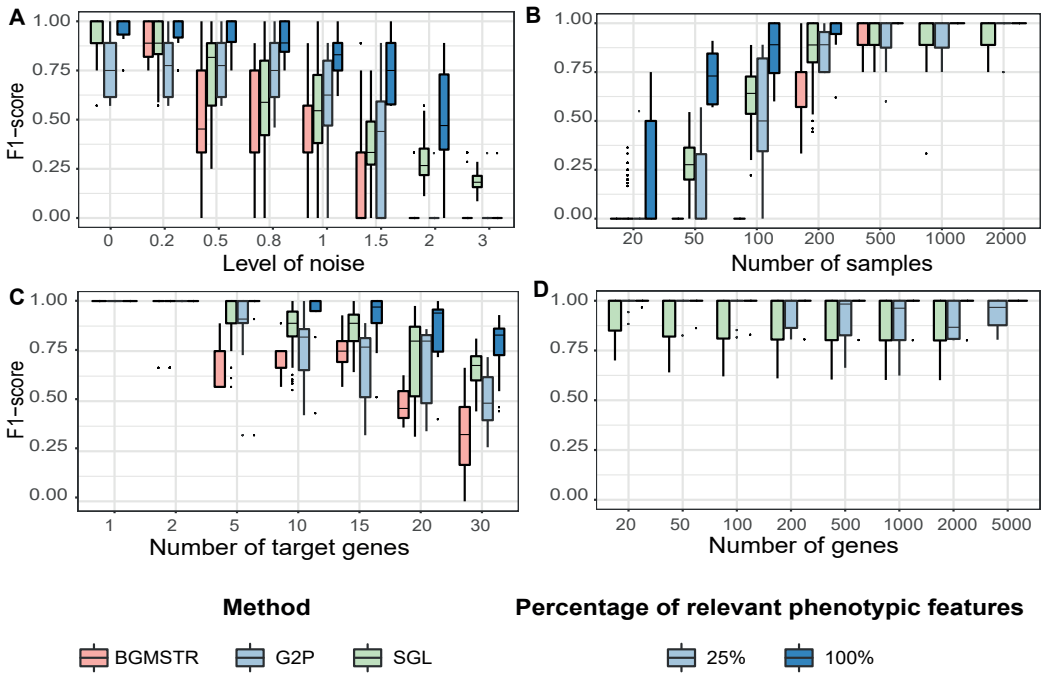

Figure 5. Testing benchmark of three group sparse methods applied to Genome-to-Phenome association according to varying generative parameters (Tab.3). **(A)** Noise level in the phenotypic data. **(B)** Number of samples available in the data. **(C)** Number of 'target'/relevant genes associated with the phenotype. **(D)** Total number of genes analysed (BGMSTR is not presented as its optimization in this scenario was computationally prohibitive). For our G2P model (blue boxplots), variations of the proposed scenarios are also shown with all 100% or only 25% of genome-related phenotypic features.

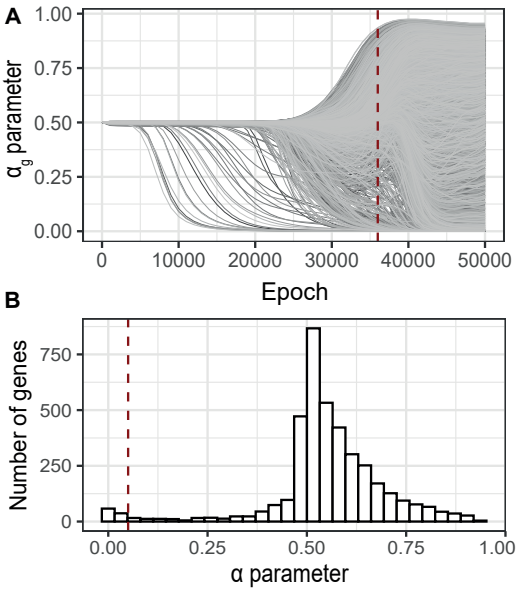

Figure 6. Identification of relevant genes using G2PSR on an ADNI dataset composed of 491 samples, 104854 SNPs grouped into 3953 genes and 8 phenotypic features. **(A)** α_g parameter per gene through the optimization process, **(B)** Distribution of the α_g parameter at the optimal epoch, 36000 (dotted red lines in A). Red dotted line correspond to the nominal 'relevance' threshold of 0.05 applied to the α_g parameter per genes.

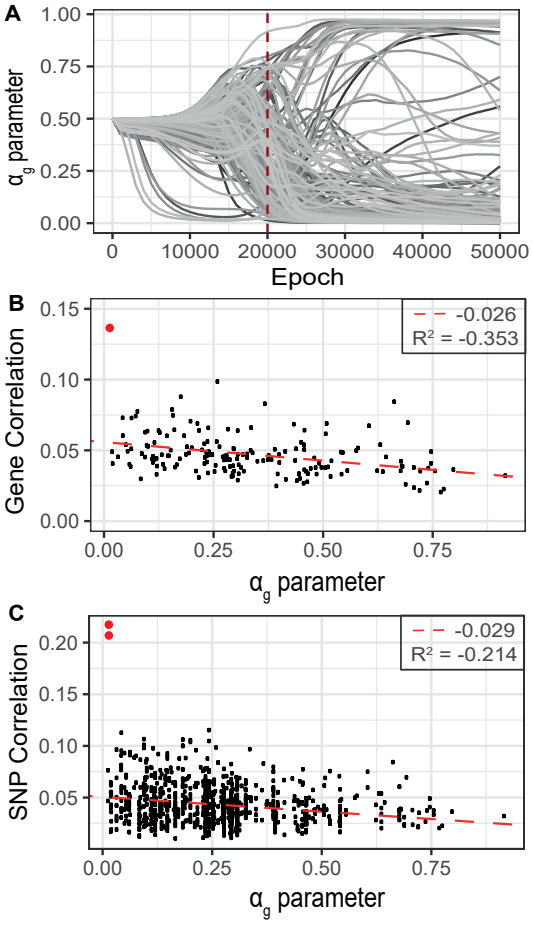

Figure 7. Refined analysis of relevant genes using G2PSR in the curated dataset identified in the primary analysis, 177 genes. **(A)** α_g parameter per gene through the optimization process, **(B)** Scatter plot of the gene-phenotypes correlation and the corresponding α_g value obtained with G2PSR (APOE gene in red), **(C)** Scatter plot of the SNP-phenotypes correlation and the corresponding α_g value obtained for their related genes (rs429358 and rs769449, APOE SNPs in red).

6.1.1 Permission to Reuse and Copyright

Figures, tables, and images will be published under a Creative Commons CC-BY licence and permission must be obtained for use of copyrighted material from other sources (including re-published/adapted/modified/partial figures and images from the internet). It is the responsibility of the authors to acquire the licenses, to follow any citation instructions requested by third-party rights holders, and cover any supplementary charges.

6.2 Tables

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Characteristics	Patients Cognitively Normal (CN)	Patients with Mild Cognitive Impairment (MCI)	Patients with Dementia (D)
Sample size	150	200	141
Age - *yr*	73.07 ± 6.97	73.73 ± 7.74	74.83 ± 7.52
Male sex - no. (%)	67 (45%)	116 (58%)	81 (57%)
APOE$-\epsilon 4$ variant - no. (%)	34 (23%)	67 (34%)	87 (62%)

Table 1. Summary of demographics and clinical information for the ADNI cohort analysed in our study.

Attribute description	Value
Number of genes	200
Number of phenotypic features	15
Number of relevant genes	5
Number of samples	500
Noise level	20

Table 2. Synthetic dataset fixed attributes used to generate 'reference' scenario.

Attribute description	Fixed Value	Iteration list
Number of genes	200	20 50 100 200 500 1000 2000 5000 10000
Number of phenotypic features	15	1 2 5 10 15 20 30
Number of relevant phenotypic features	15	4 15 as numbers OR 20, 50, 100 % as percentage
Number of relevant genes	5	1 2 5 10 15 20
Number of samples	500	20 50 100 200 500 1000 2000
Noise level	20	0 20 50 80 100 150 200 300

Table 3. Dataset attributes, varied one-at-a-time in the prescribed ranges, and used to generate scenarios according to section 2.1.

Gene Name	α_g value	Gene correlation	Relevant SNPs (correlation/p-value)
APOE	0.0138	0.1366	rs429358 (0.22 / $3.08e^{-08}$) rs769449 (0.21 / $3.17e^{-07}$)
MAT2B	0.0199	0.0490	
THOC3	0.0210	0.0406	
NKD2	0.0331	0.0454	
PTPN11	0.0438	0.0729	rs11614544 (0.08 / 0.94)
PIK3C2B	0.0441	0.0441	rs1008833 (0.11 / 0.34)
MCM2	0.0512	0.0540	rs11718485 (0.09 / 0.01)
GRK6	0.0551	0.0493	
RPL37	0.0594	0.0401	
EXOC3	0.0620	0.0386	

Table 4. G2PSR results applied to the ADNI dataset. Top genes ranked by α_g parameter. Gene correlation with phenotypic features and their associated significative SNPs (with corresponding SNP correlation and Bonferroni corrected p-value)

AUTHOR CONTRIBUTIONS

M.D., J.M. and M.L. conceived of the presented idea. M.D and M.L. developed theoretical formalism. M.D. performed the analytic calculations and performed the numerical simulations. M.S. and M.L. supervised the project. All authors discussed the results and contributed to the final manuscript.

FUNDING

This study is supported by the Neuromod Institute of the Université Côte d'Azur and by the grant ANR PARIS-15087. M.L. and M.S. are supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) (ANR-19-P3IA-0002). This publication was funded by the INRIA institution (Publication finance par une institution).

ACKNOWLEDGMENTS

The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support. INRIA sophia Antipolis - Mediterranee, "NEF" computation cluster.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.;Cogstate;Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research and Development, LLC.; Johnson and Johnson Pharmaceutical Research and Development LLC.; Lumosity; Lundbeck; Merck anf Co., Inc.; Meso Scale Diagnostics, LLC.;NeuroRx Research; Neurotrack Technologies;Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

DATA AVAILABILITY STATEMENT

The code used to generate the synthetic datasets analysed in this study can be found on the Gitlab (https://gitlab.inria.fr/mlorenzi/g2psr) as well as the G2PSR method.

REFERENCES

1 .Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* 112, 859–877. doi:10.1080/01621459.2017.1285773

2 .Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* 112, 859–877. doi:10.1080/01621459.2017.1285773

3 .Bonda, D. J., Evans, T. A., Santocanale, C., Llosá, J. C., Viña, J., Bajic, V. P., et al. (2009). Evidence for the progression through s-phase in the ectopic cell cycle re-entry of neurons in alzheimer disease. *Aging* 1, 382–388. 19946466[pmid]

4 .Civelek, M. and Lusis, A. J. (2014). Systems genetics approaches to understand complex traits. *Nature Reviews Genetics* 15, 34–48. doi:10.1038/nrg3575

5 .Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi:10.1093/bioinformatics/btr330

6 .Ge, T., Feng, J., Hibar, D. P., Thompson, P. M., and Nichols, T. E. (2012). Increasing power for voxel-wise genome-wide association studies: The random field theory, least square kernel machines and fast permutation procedures. *NeuroImage* 63, 858–873

7 .Greenlaw, K., Szefer, E., Graham, J., Lesperance, M., Nathoo, F. S., and Initiative, A. D. N. (2017). A Bayesian group sparse multi-task regression model for imaging genetics. *Bioinformatics* 33, 2513–2522. doi:10.1093/bioinformatics/btx215

8 .Greenlaw, K., Szefer, E., Graham, J., Lesperance, M., Nathoo, F. S., and Initiative, A. D. N. (2017). A Bayesian group sparse multi-task regression model for imaging genetics. *Bioinformatics* 33, 2513–2522. doi:10.1093/bioinformatics/btx215

9 .Hibar, D. P., Stein, J. L., Kohannim, O., Jahanshad, N., Saykin, A. J., Shen, L., et al. (2011). Voxelwise gene-wide association study (vgenewas): multivariate gene-based association testing in 731 elderly subjects. *NeuroImage* 56, 1875–1891. doi:10.1016/j.neuroimage.2011.03.077. 21497199[pmid]

10 .Kim, Y., Liu, G., Leugers, C. J., Mueller, J. D., Francis, M. B., Hefti, M. M., et al. (2019). Tau interacts with SHP2 in neuronal systems and in Alzheimer's disease brains. *Journal of Cell Science* 132. doi:10.1242/jcs.229054. Jcs229054

11 .Kim, Y., Liu, G., Leugers, C. J., Mueller, J. D., Francis, M. B., Hefti, M. M., et al. (2019). Tau interacts with SHP2 in neuronal systems and in Alzheimer's disease brains. *Journal of Cell Science* 132. doi:10.1242/jcs.229054. Jcs229054

12 .Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. *CoRR* abs/1312.6114

13 .Li, C., Fei, K., Tian, F., Gao, C., and Yang, S. (2019). Adipose-derived mesenchymal stem cells attenuate ischemic brain injuries in rats by modulating mir-21-3p/mat2b signaling transduction. *Croatian medical journal* 60, 439–448. doi:10.3325/cmj.2019.60.439. 31686458[pmid]

14 .Li, M., Soltanolkotabi, M., and Oymak, S. (2020). Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. *ArXiv* abs/1903.11680

15 .Li, Q. S. and De Muynck, L. (2021). Differentially expressed genes in alzheimer's disease highlighting the roles of microglia genes including olr1 and astrocyte gene cdk2ap1. *Brain, behavior, & immunity - health* 13, 100227–100227. doi:10.1016/j.bbih.2021.100227. 34589742[pmid]

16 .Liang, W. S., Dunckley, T., Beach, T. G., Grover, A., Mastroeni, D., Ramsey, K., et al. (2008). Altered neuronal gene expression in brain regions differentially affected by alzheimer's disease: a reference data set. *Physiological Genomics* 33, 240–256. doi:10.1152/physiolgenomics.00242.2007. PMID: 18270320

17 .Lindquist, M. A. and Mejia, A. (2015). Zen and the art of multiple comparisons. *Psychosomatic medicine* 77, 114–125. doi:10.1097/PSY.0000000000000148. 25647751[pmid]

18 .Liu, C.-C., Liu, C.-C., Kanekiyo, T., Xu, H., and Bu, G. (2013). Apolipoprotein e and alzheimer disease: risk, mechanisms and therapy. *Nature reviews. Neurology* 9, 106–118. doi:10.1038/nrneurol.2012.263. 23296339[pmid]

19 .Lu, Z., Khondker, Z., Ibrahim, J., Wang, Y., and Zhu, H. (2017). Bayesian longitudinal low-rank regression models for imaging genetic data from longitudinal studies. *NeuroImage* 149. doi:10.1016/j.neuroimage.2017.01.052

20 .Molchanov, D., Ashukha, A., and Vetrov, D. (2017). Variational dropout sparsifies deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (JMLR.org), ICML'17, 2498–2507

21 .Najm, R., Zalocusky, K. A., Zilberter, M., Yoon, S. Y., Hao, Y., Koutsodendris, N., et al. (2020). In vivo chimeric alzheimer's disease modeling of apolipoprotein e4 toxicity in human neurons. *Cell reports* 32, 107962–107962. doi:10.1016/j.celrep.2020.107962. 32726626[pmid]

22 .Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi:10.1093/bioinformatics/btq033

23 .S, P., B, N., K, T.-B., L, T., MAR, F., D, B., et al. (2007). Link: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics* 87

24 .Schmidt, W., Kraaijveld, M., and Duin, R. (1992). Feedforward neural networks with random weights. In *Proceedings., 11th IAPR International Conference on Pattern Recognition. Vol.II. Conference B: Pattern Recognition Methodology and Systems*. 1–4. doi:10.1109/ICPR.1992.201708

25 .[Dataset] Shaun, P. (2014). Plink (v1.9)

26 .Shen, L. and Thompson, P. M. (2020). Brain imaging genomics: Integrated analysis and machine learning. *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers* 108, 125–162. doi:10.1109/JPROC.2019.2947272. 31902950[pmid]

27 .Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics* 22, 231–245. doi:10.1080/10618600.2012.681250

28 .Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* 20, 467–484. doi:10.1038/s41576-019-0127-1

29 .Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., et al. (2017). 10 years of gwas discovery: Biology, function, and translation. *American journal of human genetics* 101, 5–22. doi:10.1016/j.ajhg.2017.06.005. 28686856[pmid]

30 .Vounou, M., Koritakova, E., Wolz, R., Stein, J., Thompson, P., Rueckert, D., et al. (2011). Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in alzheimer's disease. *NeuroImage* 60, 700–16. doi:10.1016/j.neuroimage.2011.12.029

31 .Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S. L., et al. (2012). Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the adni cohort. *Bioinformatics* 28, 229–237

32 .Wang, X., Chen, H., Yan, J., Nho, K., Risacher, S. L., Saykin, A. J., et al. (2018). Quantitative trait loci identification for brain endophenotypes via new additive model with random networks. *Bioinformatics* 34, i866–i874. doi:10.1093/bioinformatics/bty557

33 .Zhang, Y., Xu, Z., Shen, X., Pan, W., and Initiative, A. D. N. (2014). Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. *NeuroImage* 96, 309–325. doi:10.1016/j.neuroimage.2014.03.061. 24704269[pmid]

34 .Zhu, H., Khondker, Z., Lu, Z., Ibrahim, J. G., and Initiative, A. D. N. (2014). Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *Journal of the American Statistical Association* 109, 997–990. 25349462[pmid]

35 .Zhu, X., Li, X., Zhang, S., Ju, C., and Wu, X. (2017). Robust joint graph sparse coding for unsupervised spectral feature selection. *IEEE Transactions on Neural Networks and Learning Systems* 28, 1263–1275. doi:10.1109/TNNLS.2016.2521602

Attribute description	Mean (min)	Standard deviation	Minumum / Maximum
Number of genes	553	932	11 / 3913
Number of phenotypic features	128	15	101 / 161
Number of relevant phenotypic features			
Number of relevant genes	131	14	104 / 159
Number of samples	128	14	163
Noise level	104	14	104 / 165

Table 5. Average computation time and standard deviation (in minutes) per tested scenarios. Each line corresponds to the varying attribute in the considered scenarios.

6.3 Supplementary Table

Table S2. Benchmark results of three group sparse methods applied to Genome-to-Phenome association with varying noise level in the phenotypic data. Each line describes the method used, noise level tested, number of genome-related phenotypic features (15 - 100%, 4 - 25%), number of replicates and the mean F-score and standard deviation obtained.

Table S3. Benchmark results of three group sparse methods applied to Genome-to-Phenome association with varying sample size. Each line describes the method used, sample size, number of genome-related phenotypic features (15 - 100%, 4 - 25%), number of replicates and the mean F-score and standard deviation obtained.

Table S4. Benchmark results of three group sparse methods applied to Genome-to-Phenome association with varying significant genes. Each line describes the method used, number of target/significant genes, number of genome-related phenotypic features (15 - 100%, 4 - 25%), number of replicates and the mean F-score and standard deviation obtained.

Table S5. Benchmark results of three group sparse methods applied to Genome-to-Phenome association with varying number of genes processed. Each line describes the method used, total number of genes in the dataset, number of genome-related phenotypic features (15 - 100%, 4 - 25%), number of replicates and the mean F-score and standard deviation obtained.

Table S6. G2PSR results on the ADNI dataset. Table of the 3953 genes analysed with their corresponding α_g value at the optimal number of epochs (36000).

Table S7. G2PSR results on the refined ADNI dataset. Table of the 177 genes and their SNPs analysed. The table contains the genes α_g parameter at 20000 epochs, their average correlation with all phenotypic features and their corresponding SNPs correlation with each phenotypic feature and the associated correlated p-value.