NeuroImage 63 (2012) 11–24

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares

Édith Le Floch [a,c,j,*], Vincent Guillemot [a,c], Vincent Frouin [a,c], Philippe Pinel [b,c], Christophe Lalanne [f,g], Laura Trinchera [h], Arthur Tenenhaus [i], Antonio Moreno [b,c], Monica Zilbovicius [c,k], Thomas Bourgeron [e], Stanislas Dehaene [b,c,l], Bertrand Thirion [a,c,d], Jean-Baptiste Poline [a,c,d,j], Édouard Duchesnay [a,c,j]

[a] Laboratoire de Neuroimagerie Assistée par Ordinateur, Neurospin Center, I2BM, DSV, CEA, Gif-sur-Yvette, France
[b] Cognitive Neuroimaging Unit U992, INSERM-CEA, Neurospin Center, Gif-sur-Yvette, France
[c] Université Paris-Sud, IFR49, Institut d'Imagerie Neurofonctionnelle, Paris, France
[d] Parietal Project Team, INRIA Saclay-Ile de France, Neurospin Center, I2BM, DSV, CEA, Gif-sur-Yvette, France
[e] Institut Pasteur, Laboratoire de Génétique Humaine et Fonctions Cognitives, Paris, France
[f] INSERM U669, PSIGIAM, Paris, France
[g] AP-HP, Department of Clinical Research, Saint-Louis Hospital, Paris
[h] AgroParisTech, UMR518 MIA, Paris, France
[i] Supélec, Department of Signal Processing and Electronic Systems, Gif-sur-Yvette, France
[j] INSERM-CEA U1000, Neuroimaging & Psychiatry Unit, SHFJ, Orsay, France
[k] INSERM-CEA U1000, Neuroimaging & Psychiatry Unit, Necker Hospital, Paris, France
[l] Collège de France, F-75005 Paris, France

ARTICLE INFO

Article history:
Accepted 27 June 2012
Available online 8 July 2012

Keywords:
Multivariate genetic analysis
Partial Least Squares regression
Canonical Correlation Analysis
Feature selection
Regularisation

ABSTRACT

Brain imaging is increasingly recognised as an intermediate phenotype to understand the complex path between genetics and behavioural or clinical phenotypes. In this context, a first goal is to propose methods to identify the part of genetic variability that explains some neuroimaging variability. Classical univariate approaches often ignore the potential joint effects that may exist between genes or the potential covariations between brain regions. In this paper, we propose instead to investigate an exploratory multivariate method in order to identify a set of Single Nucleotide Polymorphisms (SNPs) covarying with a set of neuroimaging phenotypes derived from functional Magnetic Resonance Imaging (fMRI). Recently, Partial Least Squares (PLS) regression or Canonical Correlation Analysis (CCA) have been proposed to analyse DNA and transcriptomics. Here, we propose to transpose this idea to the DNA vs. imaging context. However, in very high-dimensional settings like in imaging genetics studies, such multivariate methods may encounter overfitting issues. Thus we investigate the use of different strategies of regularisation and dimension reduction techniques combined with PLS or CCA to face the very high dimensionality of imaging genetics studies. We propose a comparison study of the different strategies on a simulated dataset first and then on a real dataset composed of 94 subjects, around 600,000 SNPs and 34 functional MRI lateralisation indexes computed from reading and speech comprehension contrast maps. We estimate the generalisability of the multivariate association with a cross-validation scheme and demonstrate the significance of this link, using a permutation procedure. Univariate selection appears to be necessary to reduce the dimensionality. However, the significant association uncovered by this two-step approach combining univariate filtering and L1-regularised PLS suggests that discovering meaningful genetic associations calls for a multivariate approach.
© 2012 Elsevier Inc. All rights reserved.

1. Introduction

Imaging genetics studies that include a large amount of data in both the imaging and the genetic components are facing challenges for which the neuroimaging community has no definitive answer so far. Current imaging genetics studies are often either limiting the brain imaging endophenotype studied to a few candidate variables but testing their relationship with a large number of Single Nucleotide Polymorphisms (SNPs) as one usually proceeds during gene screening (e.g., Furlanello et al., 2003), or limiting the number of candidate SNPs or genes to be tested on the whole brain or some large portion of it (e.g., Glahn et al., 2007; McAllister et al., 2006; Roffman et al., 2006). When faced with both a large number of SNPs and a large number of voxels, one has to design an appropriate analysis

* Corresponding author at: CEA, Neurospin, LNAO, Bâtiment 145, F-91191 Gif-sur-Yvette, France.
E-mail address: edith.lefloch@gmail.com (É. Le Floch).

1053-8119/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2012.06.061

strategy that should be as sensitive and specific as possible. Without any priors on genetic or brain regions involved, exploratory methods can be used. The simplest approach to exploratory imaging genetics studies is clearly to apply a massive univariate analysis on both genetic and imaging data (Stein et al., 2010), which may be called Mass-Univariate Linear Modelling (MULM). However, while univariate techniques are simpler, they encounter a multiple comparison problem in the order of 10^{11}. Moreover, the link between genetic and imaging data is likely to be in part multivariate, as for instance epistasis or pleiotropy are likely phenomena in common traits or diseases. Indeed, brain imaging endophenotypes are probably influenced by the combined effects of several SNPs and different brain regions may also be influenced by the same SNP(s). A way to partially take into account epistasis may be to use a gene-based method to test for the joint effect of the different SNPs within each gene across the voxels of the whole brain (Hibar et al., 2011).

In this work, we try to go further and to identify a functional brain network covarying with a set of genetic polymorphisms, using some multivariate methods that take into account potential joint effects or covariations within each block of variables. Partial Least Squares (PLS) regression (Wold et al., 1983) and Canonical Correlation Analysis (CCA) (Hotelling, 1936) appear to be good candidates in order to look for associations between two blocks of data, as they extract pairs of covarying/correlated latent variables (one linear combination of the variables for each block). Another approach has also been proposed by Calhoun et al. (2009) based on parallel Independent Component Analysis in order to combine functional MRI data and SNPs from candidate regions. Nevertheless, all these multivariate methods encounter critical overfitting issues due to the very high dimensionality of the data.

To face these issues, methods based on dimension reduction or regularisation can be used.

Dimension reduction is essentially based on two paradigms: feature extraction and feature selection. Feature extraction looks for a low-dimensional representation of the data that explains most of its variability, the transformation being either linear such as Principal Components Analysis (PCA) or non-linear such as manifold learning methods. Feature selection methods may be divided into two categories: some univariate methods (filters), which select relevant features independently from each other, and some multivariate methods, which consider feature inter-relations to select a subset of variables (Guyon et al., 2006).

As for regularisation, multivariate methods based on L1 and/or L2 penalisations, like sparse Partial Least Squares (Chun and Keleş, 2010; Lê Cao et al., 2008, 2009; Parkhomenko et al., 2007, 2009; Waaijenborg et al., 2008; Witten and Tibshirani, 2009) or regularised CCA (Soneson et al., 2010), have recently been shown to provide good results in correlating two blocks of data such as transcriptomic and metabolomic data, gene expression levels and gene copy numbers, or gene expression levels and SNP data. One may note that such sparse multivariate methods based on L1 penalisation actually perform variable selection. Vounou et al. (2010) also introduced a promising similar method, called sparse Reduced-Rank Regression (sRRR) and based on L1 penalisation, that they applied to a simulated dataset made of 111 brain imaging features and 10s of 1000s of SNPs. The implementation of the method becomes equivalent to sparse PLS in high dimensional settings, since they make the classical approximation that in this case the covariance matrix of each block may be replaced by its diagonal elements (see Appendix). However, whether these multivariate techniques can resist even higher dimensions remains an open question. In this paper we investigate this question by adding a first step of dimension reduction on SNPs, either by PCA or univariate filtering, before applying (sparse) PLS or (regularised) CCA. We first use a simulated dataset mimicking fMRI and genome-wide SNP data and compare the performances of the different methods, by assessing their positive predictive value, as well as their capacity to generalise the link found between the two blocks with a cross-validation procedure. Indeed, we first compared PLS and CCA, then we investigated the influence of L2 regularisation on CCA and L1 regularisation on PLS, and finally we tried to add a first step of dimension reduction such as PCA or filtering.

Finally, we apply these different methods with the same cross-validation procedure on a real dataset made of fMRI and genome-wide SNP data and the statistical significance of the link obtained on "test" subjects is assessed with randomisation techniques.

In the next sections we first detail the datasets, then introduce the multivariate methods and the performance evaluation techniques that we used, and illustrate the results we obtained. Last we discuss the potential pitfalls and extensions of this work.

2. Data

2.1. Experimental dataset

This study is based on N = 94 subjects who were genotyped for 1,054,068 SNPs and participated in a general cognitive assessment fMRI task described in Pinel et al. (2007). The study (both imaging and genetics components) was approved by the local ethics committee and all subjects gave their informed consent. The task consisted of a short 5 min BOLD acquisition during which subjects were reading or listening to sentences, asked to perform a motor response (button click), subtract numbers, or were shown visual checkerboard. The functional images were acquired either on a 3 T Bruker scanner or a 3 T Siemens trio scanner using an EPI sequence (TR = 2400 ms, TE = 60 ms, matrix size = 64 × 64, FOV = 19,2 cm × 19,2 cm). T1 anatomical images were acquired during the same acquisition session with a resolution of $(1.1 \times 1.1 \times 1.2)$ mm^3. Pre-processing classically comprised slice-timing correction, motion estimation, spatial normalisation (with a resampling of the functional images at 3 mm resolution) and smoothing (FWHM = 10 mm). The preprocessings and first level model analyses were performed with SPM5 (www.fil.ucl.ac.uk/spm).

In our study, we focused only on two activation contrasts: *reading minus checkerboard viewing* and *speech comprehension minus rest*. We used a first level, subject-specific, General Linear Model (GLM), to obtain parametric estimates of the BOLD activity at each voxel in each subject; the analysis was performed using SPM5, with standard parameters (frequency cut = 128 s, AR(1) temporal noise model). For each subject s in $\{1, \ldots, n\}$ and each voxel v of the normalised volume, we obtained a map $\hat{\beta}_s(v)$ that represents the amount of BOLD signal associated with the contrast, normalised by the average signal. We defined a global brain mask for the group by considering all the voxels that belong to at least half of the individual brain masks (the individual masks were estimated using the standard SPM5 procedure). Then we selected thirty-four brain locations of interest (19 from the "reading" contrast and 15 from the "speech comprehension" contrast): most of them were the peaks of maximal activation, while the others had been reported to be atypically activated during reading in dyslexia (Paulesu et al., 2001). Each contrast map was locally averaged within 4 voxel-radius spheres centred on these peaks, keeping only active clusters of voxels ($T \geq 1$ and cluster size ≥ 10 voxels) (Pinel and Dehaene, 2009). This yielded 34 average values corresponding to 34 regions of interest (ROI) and we computed the average values for the 34 mirror ROIs by symmetry with respect to the inter-hemispheric plane. Finally, lateralisation indexes were derived from those regions. For each pair of ROIs in the normalised volume and in each subject, an index was computed as follows:

$$\mathrm{Index}_s = \frac{\hat{\beta}_s^{\mathrm{right}} - \hat{\beta}_s^{\mathrm{left}}}{\sqrt{\left(\hat{\beta}_s^{\mathrm{right}}\right)^2 + \left(\hat{\beta}_s^{\mathrm{left}}\right)^2}}. \quad (1)$$

The distribution of these indexes spanned the range of $[-1.5; 1.5]$, and variances were homogeneous across regions of interest. The term "phenotypes" will now refer to the lateralisation indexes thus obtained in the different regions.

For each subject, an Illumina platform was used to genotype 1,054,068 SNPs and processed with the standard platform software. Considering all genotyped data available, we successively applied the following filters on all SNPs: (1) Minor Allele Frequency (MAF) at least 10%, (2) call rate at least 95%, and (3) Hardy-Weinberg test not significant at the 0.005 level. Assuming an additive genetic model, genetic data were recoded as the number of minor alleles (denoted as A), {0, 1, 2}, hence a value of 0 means homozygous wild-type individuals (BB). The frequency of homozygous individuals for the minor allele (AA) was 0.03–0.13 in 75% of the cases. Missing SNP data were imputed with their corresponding median value across subjects.[1] These analyses were carried out using the open-source R software (R Development Core Team, 2009) and the storage facilities for genetic data provided in the package snpMatrix (Clayton and Cheung, 2007).

After these preprocessing steps, our analyses were performed on two blocks of data **Y** (fMRI) and **X** (genetics) of size 94×34 and $94\times622{,}534$ respectively.

2.2. Simulated dataset

A simulated dataset mimicking the real dataset was also simulated in order to study the behaviour of the different methods of interest, while knowing ground truth. 500 samples of 34 imaging phenotypes were simulated from a multivariate normal distribution with parameters estimated from the experimental data.

In order to simulate genotyping data with a genetic structure similar to that of our real data, we considered a simulation method that uses the HapMap CEU panel. We used the gs algorithm proposed by Li and Chen (2008) with the phased (phase III) data for CEU unrelated individuals for chromosome 1; we only consider the genotype simulation capability of this software that may also generate linked phenotypes. We generated a dataset consisting in 85,772 SNPs and 500 samples, using the extension method of the algorithm. We randomly selected 10 SNPs (out of 85,772) having a MAF = 0.2 and 8 imaging phenotypes (out of 34). We induced two independent causal patterns: for the first pattern we associated the first 5 SNPs with the first 4 imaging phenotypes; the second pattern was created associating the 5 remaining SNPs with the 4 last phenotypes. For each causal pattern $i\in\{1,2\}$, we induced a genetic effect using an additive genetic model involving the average of the causative SNPs (x_{ik}): $\bar{x}_i = \sum_{k=1}^{5} \frac{1}{5} x_{ik}$. Then each imaging phenotype y_{ij} ($j\in[1,\ldots,4]$) of the pattern i was affected using a linear model:

$$y_{ij}^{\star} = y_{ij} + \beta_{ij}\bar{x}_i \tag{2}$$

The parameter β_{ij} was setted by controlling for the correlation (at a value of 0.5) between the j^{th} affected imaging phenotype ($y_{ij}^{\star}$) and the causal SNPs ($\bar{x}_i$) i.e.: $\mathrm{corr}\left(y_{ij}^{\star}, x_i\right) = 0.5$. Such control of the correlation (or the explained variance) is equivalent to the control of the effect size while controlling for the variances of SNPs ($\mathrm{var}(x_i)$) and (unaffected) imaging phenotypes ($\mathrm{var}(y_{ij})$, as well as any spurious covariance between them $\left(\mathrm{cov}\left(y_{ij}, x_i\right)\right)$. We favour such control over a simple control for the effect size since the later may result in arbitrary huge or weak associations depending on the genetic/imaging variances ratios.

[1] Other imputation methods were tested, e.g. the Markov Chain based haplotyper proposed by Abecasis and coworkers (Sanna et al., 2008; Willer et al., 2008). All yield similar profiles of allele frequencies for our data set.

SNP whose r^2 coefficient with any of the causal SNPs is at least 0.8 is also considered as causal. Such LD threshold, commonly used in the literature (de Bakker et al., 2005), led to 56 causal SNPs: 32 in "pattern 1" and 24 in "pattern 2". We will use those SNPs as "ground truth" of truly causal SNPs to compute the true positive rates of the learning methods. Finally, we striped off 10 blocks of SNPs around the 10 causal SNPs, from the whole genetic dataset, considering that neighbouring SNPs were in LD with the marker if their r^2 were at least 0.2. The 5 first (resp. last) blocks, of pattern 1 (resp. 2), are made of 127 (resp. 71) SNPs and contain all the 32 (resp. 24) SNPs that were declared as causal. The striped blocks were concatenated and moved at the beginning of the dataset leading to 198 $(127+71)$ informative features followed by 85,574 $(85{,}772-198)$ non-informative (noise) features. Such a dataset organisation provides a simple way to study the methods' performances while the dimensionality of the input dataset increases from 200 (mostly made of informative features) to 85,772 mostly made of noise.

Next, we will present the different strategies we investigated in order to analyse such data.

3. Methods

3.1. Partial Least Squares regression

Partial Least Squares regression is used to model the associations between two blocks of variables hypothesising that they are linked through unobserved latent variables. A latent variable (or component) corresponding to one block is a linear combination of the observed variables of this block.

More precisely, PLS regression builds successive and orthogonal latent variables for each block such that at each step the covariance between the pair of latent variables is maximal. For each step h in $1..H$, where H is the maximal number of pairs of components, it optimises the following criterion:

$$\begin{aligned} &max_{\|\mathbf{u}_h\|_2=\|\mathbf{v}_h\|_2=1}\ \mathrm{cov}(\mathbf{X}_{h-1}\mathbf{u}_h, \mathbf{Y}_{h-1}\mathbf{v}_h) \\ &= max_{\|\mathbf{u}_h\|_2=\|\mathbf{v}_h\|_2=1}\ \mathrm{corr}(\mathbf{X}_{h-1}\mathbf{u}_h, \mathbf{Y}_{h-1}\mathbf{v}_h)\sqrt{\mathrm{var}(\mathbf{X}_{h-1}\mathbf{u}_h)}\sqrt{\mathrm{var}(\mathbf{Y}_{h-1}\mathbf{v}_h)} \end{aligned} \tag{3}$$

where $\mathbf{u}_h$ and $\mathbf{v}_h$ are the weight vectors for the linear combinations of the variables of blocks **X** and **Y** respectively. $\mathbf{X}_{h-1}$ and $\mathbf{Y}_{h-1}$ are the residual (deflated) **X** and **Y** matrices after their regression on the *h-1* previous pairs of latent variables , starting with $\mathbf{X}_0 = \mathbf{X}$ and $\mathbf{Y}_0 = \mathbf{Y}$ (whose columns have been standardised). There exist two ways of deflation: an asymmetric way (the original PLS regression) and a symmetric way (canonical-mode PLS). The difference is that in the first case both blocks are deflated on the latent variables of block **X** (which becomes the predictor block), while in the second case each bock is deflated on its own latent variables. In our case, we are more interested in canonical-mode PLS as we investigate exploratory methods trying to extract covarying networks among a huge amount of neuroimaging and SNP data, many of which are very likely to be irrelevant. Note that, on the first pair of components, the original PLS regression and canonical-mode PLS give exactly the same results. In the rest of the paper, we have dropped the h index that stands for the number of pairs of components to make the notations simpler.

Once the variables are standardised, the previous criterion for each new pair of components is equivalent to optimising:

$$\underset{\|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1}{max}\ \mathbf{u}'\mathbf{X}'\mathbf{Y}\mathbf{v} \tag{4}$$

This optimisation problem is solved using the iterative algorithm called NIPALS (Wold, 1966) and more precisely the NIPALS inner loop, the NIPALS outer loop being the iteration over the number of pairs of components. The optimal vectors **u** and **v** are in fact the

first pair of singular vectors of the matrix $\mathbf{X'Y}$. Please note that the criterion tends to maximise the relative value of the covariance, which implies that the covariance is forced to be null or positive. In the case of a negative association between a variable from block $\mathbf{X}$ and a variable from block $\mathbf{Y}$, a negative weight will thus be assigned to one of them in order to obtain a positive covariance.

However, multivariate methods such as PLS regression encounter overfitting issues in high-dimensional settings. For instance, Chun and Keleş (2010) recently showed that asymptotic consistency of the PLS regression estimator does not hold when $p = \mathcal{O}(n)$, where p is the number of variables for blocks $\mathbf{X}$ and n the number of observations or individuals.

3.2. PLS–SVD

A variant of PLS regression is called Tucker Inter-battery Analysis (Tucker, 1958) or PLS–SVD (McIntosh et al., 1996). This variant is symmetric and consists in computing all pairs of left and right singular vectors of $\mathbf{X}'\mathbf{Y}$ at once, which form the weight vectors $\mathbf{u}_h$ and $\mathbf{v}_h$ for $\mathbf{X}$ and $\mathbf{Y}$ blocks respectively. It gives the same results as PLS regression on the first pair of latent variables, but differs on further pairs due to a different orthogonality constraint. While PLS regression forces successive latent variables of each block to be orthogonal, PLS–SVD forces successive weight vectors of each block to be orthogonal, which leads to the orthogonality between each latent variable $\mathbf{Xu_h}$ of block $\mathbf{X}$ and each latent variable $\mathbf{Yv_j}$ of block $\mathbf{Y}$, as long as they are of different order ($h \neq j$).

3.3. Canonical Correlation Analysis

A similar method is Canonical Correlation Analysis (CCA), which differs in that the correlation between the two latent variables, instead of the covariance, is maximised at each step. CCA builds successive and orthogonal latent variables for each block such as, at each step h in 1..H, they optimise the following criterion:

$$\underset{\|\mathbf{u}_h\|_2=\|\mathbf{v}_h\|_2=1}{max} \mathrm{corr}(\mathbf{X}\mathbf{u}_h, \mathbf{Y}\mathbf{v}_h)$$

where $\mathbf{u}_h$ and $\mathbf{v}_h$ are weight vectors.

Once the variables are standardised, it becomes equivalent to optimising:

$$\underset{\|\mathbf{u}_h\|_2=\|\mathbf{v}_h\|_2=1}{max} \frac{\mathbf{u}_h\mathbf{X}'\mathbf{Y}\mathbf{v}_h}{\sqrt{\mathbf{u}_h\mathbf{X}'\mathbf{X}\mathbf{u}_h}\sqrt{\mathbf{v}_h\mathbf{Y}'\mathbf{Y}\mathbf{v}_h}}$$

The solution may be obtained by computing the SVD of $\mathbf{X}'\mathbf{X}^{-1/2}\ \mathbf{X}'\mathbf{Y}\ \mathbf{Y}'\mathbf{Y}^{-1/2}$. The successive pairs of weight vectors $\mathbf{u}_h$ and $\mathbf{v}_h$ are obtained by:

$\mathbf{u}_h = \mathbf{X}'\mathbf{X}^{-1/2}\mathbf{e}$ and $\mathbf{v}_h = \mathbf{Y}'\mathbf{Y}^{-1/2}\mathbf{f}$, where the columns of $\mathbf{e}$ and $\mathbf{f}$ are the left and right singular vectors respectively.

Like PLS regression, CCA has to face overfitting issues in high-dimensional settings. Moreover, CCA requires the inversion of the scatter matrices $\mathbf{X}'\mathbf{X}$ and $\mathbf{Y}'\mathbf{Y}$, which are ill-conditioned in our high-dimensional settings with very large p and q (numbers of variables for blocks $\mathbf{X}$ and $\mathbf{Y}$ respectively) and a small N (number of observations or individuals).

For numerical issues, we used the dual formulation of CCA based on a linear kernel: Kernel CCA (KCCA).

3.4. Regularisation techniques

3.4.1. L2 regularisation

In order to first solve the overfitting and the non-invertibility issues of CCA, regularisation based on L2 penalisation may be used, by replacing the matrices $\mathbf{X}'\mathbf{X}$ and $\mathbf{Y}'\mathbf{Y}$ by $\mathbf{X}'\mathbf{X} + \lambda_2 I$ and $\mathbf{Y}'\mathbf{Y} + \lambda_2 I$ respectively. However, in such high-dimensional settings the approximation is often made that the scatter matrices $\mathbf{X}'\mathbf{X}$ and $\mathbf{Y}'\mathbf{Y}$ may be replaced by identity matrices, which is an extreme case of shrinkage of the scatter matrices and makes CCA equivalent to PLS–SVD and thus to PLS regression as well on the first component. Shrinkage of the scatter matrices is similar to L2-regularisation, leading to proportional solutions for weight vectors (with a $1 + \lambda_2$ factor).

3.4.2. L1 regularisation

Another solution to the overfitting issue may be to use regularisation techniques based on L1 penalisation. Recently Lê Cao et al. (2008) proposed an approach that includes variable selection in PLS regression, based on L1 penalisation (Tibshirani, 1996) and leading to a sparse solution. By contrast, it should be noted that L1 penalisation may not be easily implemented on PLS–SVD without loosing the orthogonality constraint on weight vectors (Zou et al., 2006). In sparse PLS regression (sPLS), the PLS regression criterion for each new pair of components is modified by adding a L1 penalisation on weight vectors $\mathbf{u}$ and $\mathbf{v}$:

$$\underset{\|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1}{min} -\mathbf{u}'\mathbf{X}'\mathbf{Y}\mathbf{v} + \lambda_{1X}\|\mathbf{u}\|_1 + \lambda_{1Y}\|\mathbf{v}\|_1 \quad (5)$$

where λ_{1X} and λ_{1Y} are L1-penalisation parameters for the weight vectors of blocks $\mathbf{X}$ and $\mathbf{Y}$ respectively. The sPLS criterion is bi-convex in $\mathbf{u}$ and $\mathbf{v}$ and may be solved iteratively for $\mathbf{u}$ fixed or $\mathbf{v}$ fixed, using soft-thresholding of variable weights at each iteration of the NIPALS inner loop. Weight vectors $\mathbf{u}$ and $\mathbf{v}$ are computed using the following algorithm:

1. Initialise $\mathbf{u}$ and $\mathbf{v}$ using for instance the first pair of singular vectors of the matrix $\mathbf{X}'\mathbf{Y}$ and normalise them.
2. Until convergence of $\mathbf{u}$ and $\mathbf{v}$:
   (a) For fixed $\mathbf{v}$:

$$\hat{\mathbf{u}} = \underset{\|\mathbf{u}\|_2=1}{arg\ min} -\mathbf{u}'\mathbf{X}'\mathbf{Y}\mathbf{v} + \lambda_{1X}\|\mathbf{u}\|_1 = g_{\lambda_{1X}}\left(\mathbf{X}'\mathbf{Y}\mathbf{v}\right) \quad (6)$$

   where $g_\lambda(y) = \mathrm{sign}(y)(|y| - \lambda)_+$ is the soft-thresholding function.
   (b) Normalise $\mathbf{u}$: $\mathbf{u} \leftarrow \frac{\mathbf{u}}{\|\mathbf{u}\|_2}$.
   (c) For fixed $\mathbf{u}$:

$$\hat{\mathbf{v}} = \underset{\|\mathbf{v}\|_2=1}{arg\ min} -\mathbf{u}'\mathbf{X}'\mathbf{Y}\mathbf{v} + \lambda_{1Y}\|\mathbf{v}\|_1 = g_{\lambda_{1Y}}\left(\mathbf{Y}'\mathbf{X}\mathbf{u}\right) \quad (7)$$

   (d) Normalise $\mathbf{v}$: $\mathbf{v} \leftarrow \frac{\mathbf{v}}{\|\mathbf{v}\|_2}$.

In the version of sparse PLS that we used, L1 penalisation is performed by soft-thresholding of variable weights and instead of setting λ_{1X} and λ_{1Y} directly, the corresponding numbers of $\mathbf{X}$ and $\mathbf{Y}$ variables to be kept in the model are chosen. We then defined the sPLS selection rates, $s_{\lambda_{1X}}$ and $s_{\lambda_{1Y}}$, as the number of selected variables from each block out of the total number of variables of that block. In our case, we chose to apply sparsity on SNPs only and to set $s_{\lambda_{1Y}}$ to 1 for imaging phenotypes, as we had a very large number of SNPs and only a few imaging phenotypes.

Sparse versions of CCA have also been proposed by Parkhomenko et al. (2007, 2009), Waaijenborg et al. (2008), Witten and Tibshirani (2009). However, in order to solve the non-invertibility issue, they make the approximation that the covariance matrices $\frac{1}{n-1}\mathbf{X}'\mathbf{X}$ and $\frac{1}{n-1}\mathbf{Y}'\mathbf{Y}$ may be replaced by their diagonal elements, which makes sparse CCA equivalent to sparse PLS regression.

However, whether sparse PLS can face overfitting issues by itself in the case of such high-dimensional data remains an open question. This is the reason why we decided to combine it with a first step of dimension reduction on SNPs.

3.5. Dimension reduction methods

3.5.1. PC-based dimension reduction

A first way to perform dimension reduction might be to add a first step of Principal Component Analysis on each block of data before applying PLS or CCA. Regularisation is not necessary anymore in that case, as the dimension has been dramatically reduced. For each block of data, we kept as many components as necessary to explain 99% of the variance of that block. We also investigated the performance of Principal Component Regression (PCR) of the two first imaging principal components onto the genetic components explaining 99% of the genetic variance.

3.5.2. Univariate SNP filtering

Another way to perform dimension reduction is to add to sparse PLS or regularised CCA a first step of massive univariate filtering. This step consisted of 1 — $p \times q$ pair-wise linear regressions based on an additive genetic model, 2 — ranking the SNPs according to the minimal p-value each SNP gets across all phenotypes, and 3 — keeping the set of SNPs with the lowest "minimal" p-values. Indeed, even though univariate filtering may seem to contradict the very nature of multivariate methods such as PLS or CCA, it still allows them to extract multivariate patterns among the remaining variables and may even be necessary to overcome the overfitting issue in very high dimensional settings. We may note at this point that the univariate approach alone did not yield any significant SNP/phenotype associations at the 5% level after Bonferroni or FDR correction.

3.6. Comparison study

We compared the performances of the different methods on both simulated and real datasets. Indeed we first compared PLS and CCA, then we investigated how their performance is improved by regularisation with sparse PLS and L2-regularised CCA, and we finally assessed the influence of a first dimension reduction step by PCA or filtering. Note that computations were always limited to the two first pairs of latent variables for computational time purposes. Moreover we were also interested in comparing the different methods with MULM. Table 1 summarises the different methods we tested and the acronyms we used.

In this paper we investigated in particular the performance of fsPLS on both simulated and real data and we tried to assess how much the performance of fsPLS is influenced by the fact of varying the sparse PLS penalisation parameter $s_{\lambda_{1X}}$ and the number k of SNPs kept by the filter.

3.7. Performance evaluation

We decided to evaluate the performances of the different methods by assessing the generalisability of the link they find between the blocks, on both simulated and real data, using a 5-fold and a 10-fold cross-validation (CV) scheme respectively. On the real dataset, we used "training" sets of 84 or 85 subjects and "test" sets of 9 or 10 subjects. In order to have "training" sets of about the same size on the simulated dataset, we used "training" sets of 100 subjects and "test" sets of 400 subjects.

For each method, at each fold of the CV, the estimation of the model (weight vectors) was done on the training sample and tested on the hold-out sample (Fig. 1 for filter-based methods and Fig. 2 for PC-based methods). Indeed, at each fold, the weights thus obtained were used to build the factorial scores of the "test" samples (the set of left out subjects) and the correlation coefficient between those factorial scores was computed. This yielded an average "test" correlation coefficient over folds, called the out-of-sample correlation coefficient, which reflects the link between the two blocks estimated on unseen subjects. Please note that at each fold, while the correlation coefficient obtained on the training samples is forced to be positive, the out-of-sample correlation coefficient may happen to be negative.

We performed a CV for MULM as well, where at each fold the two most significantly associated SNP/phenotype pairs on the training sample were extracted and tested by computing their correlation coefficient on the hold-out sample.

Finally, in the case of simulated data, ground truth was known and we could also compare the performances of the different methods by computing the Positive Predictive Value (PPV) when 50 SNPs are selected by each method. This is almost equivalent in our case to the specificity of each method when 50 SNPs are selected, since there are 56 causal SNPs in our simulated dataset. PPV curves were separately computed on 5 non-overlapping subsamples of 100

Table 1
Summary of the different strategies investigated.

Method	Acronym
Mass Univariate Linear Modelling	MULM
Partial Least Squares	PLS
Kernel Canonical Correlation Analysis	KCCA
sparse PLS	sPLS
regularised KCCA	rKCCA
Principal Component Analysis + PLS	PCPLS
Principal Component Analysis + KCCA	PCKCCA
Filtering + (sparse) PLS	f(s)PLS
Filtering + (regularised) KCCA	f(r)KCCA

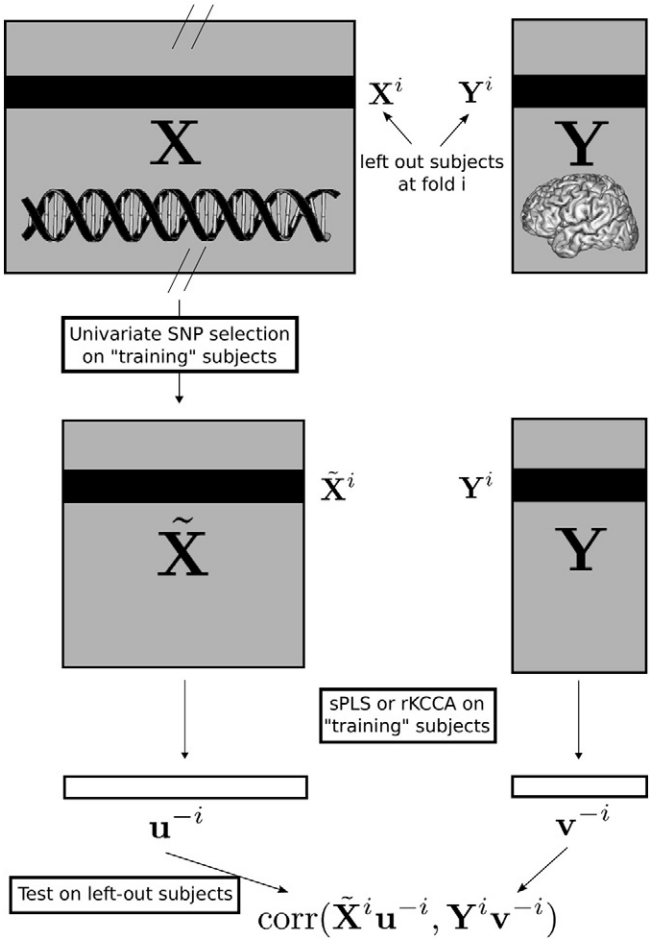

Fig. 1. Illustration of the cross-validation scheme for filter-based methods. At each fold i, a univariate selection of k SNPs is performed on the data of "training" subjects $\mathbf{X}^{-i}$ and $\mathbf{Y}^{-i}$; the weight vectors, u^{-i} and v^{-i}, are then estimated by sPLS or rKCCA on the "training" subjects and finally the scores of the left out subjects corresponding to this ith fold are computed using their observed responses $\tilde{\mathbf{X}}^i$ and $\mathbf{Y}^i$ and these weight vectors.

observations and averaged over these 5 subsamples. It should be noted that the informative SNPs that are not considered as causal are only slightly correlated to causal SNPs. Therefore they were removed to compute the PPV, since they could not really be identified as true or false effects.

4. Results

4.1. Performance assessment on simulated data

4.1.1. Influence of regularisation

We were first interested in comparing the performances of PLS and CCA when the number of SNPs p increases, from 200 (mostly made of the 198 informative features) up to 85,772 SNPs (mostly made of noise), and investigating the influence of L1 regularisation on PLS and of L2 regularisation on CCA.

Fig. 3, on the left panel, shows the out-of-sample correlation coefficients obtained with the different methods for the two first component pairs, and it shows that in the lower dimensional space ($p=200$) mostly made of informative features, the pure CCA, rKCCA without regularisation ($\lambda_2=0$), has overfitted the "training" data on the first component pair ("training" corr. ≈ 1 and "test" corr. ≈ 0.2). Such a result highlights the limits of pure CCA to deal with situations where the number of training samples (100) is smaller than the dimension ($p=200$). However, with a suitable regularisation in such a low-dimensional setting, rKCCA($\lambda_2=100$) performed better than all other methods, notably all (sparse) PLS. These results were expected since the evaluation criterion (correlation between factorial scores) is exactly the one which is maximised by CCA.

Nevertheless, the increase of space dimensionality (with irrelevant features) clearly highlights the superiority of PLS and more notably sPLS over rKCCA in high-dimensional settings: the performance of rKCCA rapidly decreases while sPLS ($s_{\lambda_{1X}}=0.1$) tolerates an increase of the dimensionality up to 1000 features before its performance starts to decrease. One may note that as expected theoretically, along with the increase of penalisation (λ_2), rKCCA curves smoothly converge toward PLS.

On the second component pair, the results are less clearly interpretable. However (s)PLS curves are above the rKCCA ones.

The four graphs on the right panel of Fig. 3 demonstrate the superiority of sPLS methods to identify causal SNPs on the two first genetic components. Indeed, for each method and for different values of p, we computed the PPV for the two first genetic components and for each simulated pattern. PPV curves show a smooth increase of the performance, when moving from unregularised CCA (rKCCA($\lambda_2=0$)) to strongly regularised PLS (sPLS($s_{\lambda_{1X}}=0.1$)). Moreover, while the out-of-sample correlation coefficient was not an appropriate measure to distinguish between the two causal patterns, PPV curves were computed for each pattern separately. One may note that the PPV on the first genetic component appears to be much higher for the first pattern than for the second pattern, especially in low dimensions, while the opposite trend is observed on the second genetic component. This observation tends to show that the first causal pattern is captured by the first component pair, while the second pattern is captured by the second pair. It should be noted that the PPV even reaches one when $p=200$ for the first pattern on the first component, meaning that only true positives from the first pattern are detected on this component. Similarly, the PPV reaches one when $p=200$ for the second pattern on the second component.

4.1.2. Influence of the dimension reduction step

Then we investigated the influence of a first step of dimension reduction. Fig. 4 presents different dimension reduction strategies: Principal Component (PC), filter (f), sparse (s) and combined filter + sparse (fs) methods. Here the parameter setting, 50 selected SNPs, was derived from the known ground truth (56 true causal SNPs). The 50 SNPs were either the 50 best ranked SNPs for (f) methods, the 50 non-null weights for sparse PLS or a combination of both: either 10% of the 500 best ranked SNPs or 50% of 100 for fsPLS.

Fig. 4, on the left panel, shows that all PC-based methods (green curves) failed to identify generalisable covariations when the number of irrelevant features increases.

Dimension reduction based on filtering slightly improved the performance of CCA and greatly improved the performance of PLS: fPLS($k=50$) is the second best approach in our comparative study.

Moreover, as previously observed in Fig. 3, L1 regularisation limits the overfitting phenomenon (see sPLS($s_{\lambda_{1X}}*p=50$) in Fig. 4) and delays the decrease of PLS performance when the dimensionality increases. Finally the best performance is obtained by combining filtering and L1 regularisation: fsPLS($k=100$, $s_{\lambda_{1X}}=0.5$), which keeps 100 SNPs after filtering and selects 50% of those SNPs by sPLS. Please note that the performance of fsPLS ($k=500$, $s_{\mathrm{da}_{1X}}=0.1$) is lower and similar to that of sPLS(50) in low dimensions, but becomes more robust than sPLS and equivalent to fsPLS($k=100$, $s_{\lambda_{1X}}=0.5$) in higher dimensions. However, the purely univariate strategy based on MULM shows poor generalisability, which suggests that even though filtering appears necessary to remove irrelevant features, it is not able to capture the imaging/genetics link by itself and needs to be combined with a multivariate step which will take advantage of the cumulative effects of several SNPs. Nevertheless, it should be noted that the way we assessed the generalisability of MULM was arbitrary, since we only looked at the two best SNP/phenotype associations.

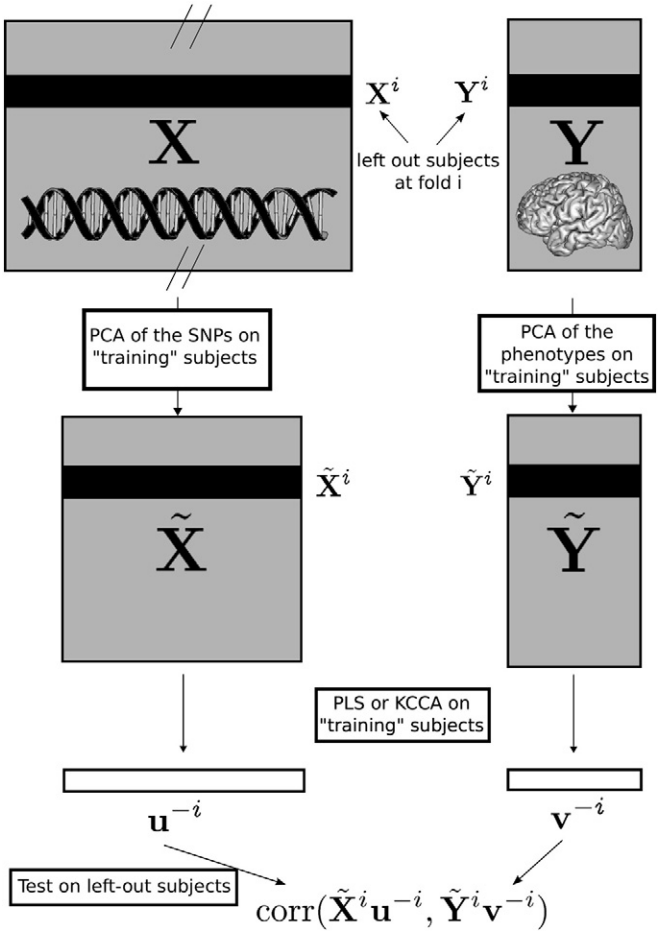

Fig. 2. Illustration of the cross-validation scheme for PC-based methods. At each fold i, two PCAs are performed on SNPs and on phenotypes of "training" subjects $\mathbf{X}^{-i}$ and $\mathbf{Y}^{-i}$; the weight vectors, u^{-i} and v^{-i}, are then estimated by PLS or KCCA on the "training" subjects, and finally the scores of the left out subjects corresponding to this ith fold are computed using the projection of their observed responses on the principal components, $\tilde{}\mathbf{X}^i$ and $\tilde{}\mathbf{Y}^i$, and these weight vectors.

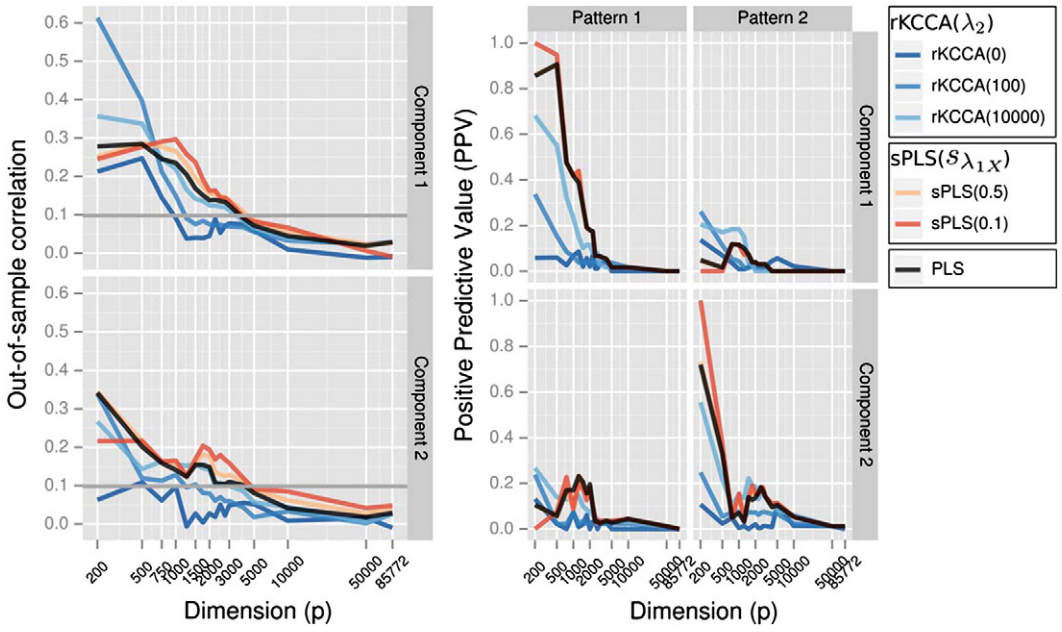

Fig. 3. Comparison of regularisation methods to deal with genetic datasets containing an increasing number of irrelevant features. The total number of features varies along the *x*-axis between 200 and 85,772 SNPs. We compared: (i) in blue, regularised kernel CCA (rKCCA) with various L2 regularisation values (λ_2) ranging from 0 (pure CCA) to 10,000; (ii) in black, PLS; (iii) in red, sparse PLS (sPLS) with various L1 regularisation values ($s_{\lambda_{1X}}$) ranging from 0.75 (75% of input features have a non null weight) to 0.1. The *y*-axis of the two left panels shows the (5-fold CV) average out-of-sample correlation coefficients between the two first component pairs. The four right panels present the power of the methods to identify causal SNPs implied in the two causal patterns. The *y*-axis depicts the Positive Predictive Values when 50 SNPs are selected, for each of the two first genetic components: (u_1, u_2).

Again, on the second component pair, the results are less clearly interpretable. However the curves of the strategies that combine filtering and sparsity are above the other ones.

The four graphs on the right panel of Fig. 4 show that the results in terms of PPV performance are similar to cross-validation results. However, it should be noted that the PPV does not take into account the weights/ranks assigned by the different methods to the selected SNPs. Therefore, the PPV curves of fKCCA($k=50$) and fPLS($k=50$) are superimposed on the MULM curve in our case, since the 50 SNPs selected by the filter are the 50 best SNPs obtained with MULM.

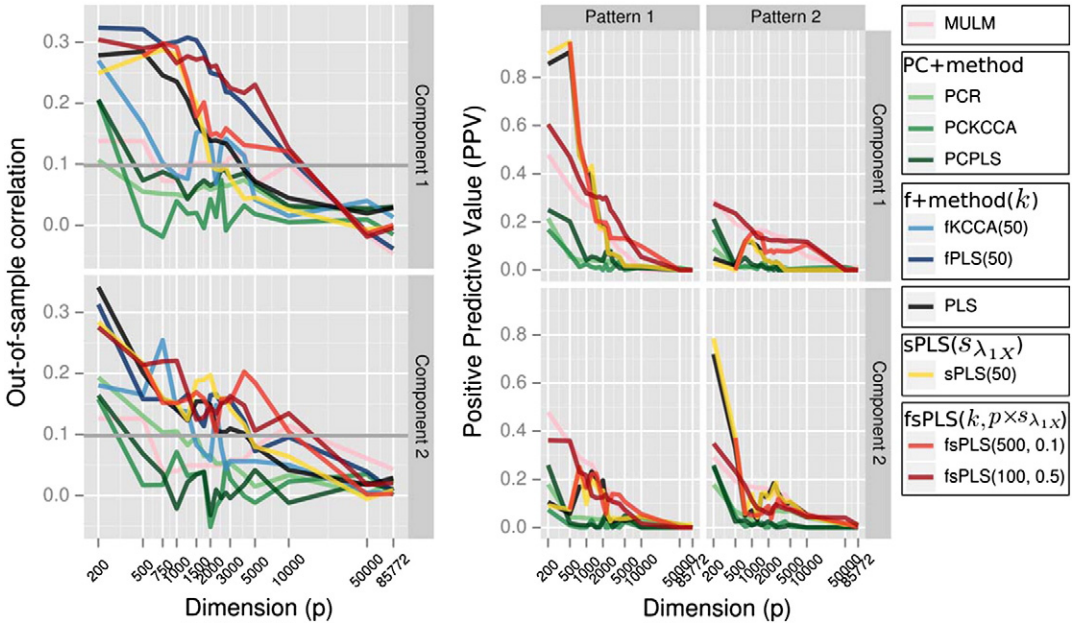

Fig. 4. Comparison of dimension reduction methods to deal with genetic datasets containing an increasing number of irrelevant features. The total number of features varies along the x-axis between 200 and 85,772 SNPs. We compared: (i) in green, Principal Component (PC) based methods: PC regression (PCR), PCA + KCCA (PCKCCA), PCA + PLS (PCPLS). (ii) in blue, filter (f) based methods: f + KCCA (fKCCA), f + PLS (fPLS). We selected only the 50 best SNPs, while according to ground truth 56 SNPs were identified as causal. (iii) in black, PLS. (iv) in yellow, sparse PLS (sPLS) where $s_{\lambda_{1X}}$ is such that 50 features have a non null weight. (v) in red, filter + sparse PLS (fsPLS) with settings both leading to 50 selected features: fsPLS($k=500$, $s_{\lambda_{1X}}=0.1$) (resp. fsPLS($k=100$, $s_{\lambda_{1X}}=0.5$) keeps the 500 (resp. 100) best ranked features and then 10% (50%) get a non null weight. (vi) finally in pink, we add MULM. The y-axis of the two left panels shows the (5-fold CV) average out-of-sample correlation coefficients between the two first component pairs. The four right panels present the power of the methods to identify causal SNPs implied in the two causal patterns. The y-axis depicts the Positive Predictive Values when 50 SNPs are selected, for each of the two first genetic components: (u_1, u_2).

Table 2
The two first average correlation coefficients found on left-out "test" samples and on "training" samples.

	ρ^1_{test}	ρ^2_{test}	$\rho^1_{training}$	$\rho^2_{training}$
MULM	0.036	−0.104	−0.458	−0.451
PLS	−0.092	0.218	0.990	0.984
sPLS ($s_{\lambda_{1X}}=0.1\%$)	0.008	0.201	0.938	0.922
PCKCCA	0.010	0.008	1.000	1.000
PCPLS	−0.088	0.217	0.990	0.984
frKCCA ($k=1000$, $\lambda_2=1{,}000{,}000$)	0.245	0.324	0.963	0.954
fPLS ($k=1000$)	0.236	0.268	0.962	0.953
fsPLS ($k=1000$, $s_{\lambda_{1X}}=5\%$)	0.432	0.210	0.772	0.788

4.2. Performance assessment on real data

4.2.1. Comparative analysis

Table 2 summarises the two first average correlation coefficients obtained on "test" samples (ρ^1_{test} and ρ^2_{test}) for the different methods tested, as well as the two first average correlation coefficients obtained on "training" samples ($\rho^1_{training}$ and $\rho^2_{training}$). The "optimal" parameters for regularisation and filtering chosen here are those that gave the best average cross-validated correlation coefficients, among all parameters tested.

Table 2 shows that, for the first pair of components, L1 regularisation of PLS cannot solve the overfitting issue by itself. Indeed, like PLS, sparse PLS ($s_{\lambda_{1X}}=0.1\%$) completely failed to extract a generalisable link in such high dimensions and captured only noise. In such high dimensions, KCCA requires such an extreme L2 regularisation that it is equivalent to PLS in terms of correlation between latent variables (with a proportionality factor of $\frac{1}{1+\lambda_2}$ on weight vectors).

Therefore a first step of dimension reduction appears to be necessary in order to overcome the overfitting issue. Indeed, even though PC-based methods do not succeed either, filtering-based methods perform much better. Among filtering-based methods, fsPLS yields the highest out-of-sample correlation coefficient of 0.43 when 1000 SNPs are left after the univariate filter and respectively 5% of the remaining SNPs are kept by sparse PLS. The second best performance on the first pair of components is obtained with frKCCA with an out-of-sample correlation coefficient of 0.24 ($k=1000$ and $\lambda_2=1{,}000{,}000$). However, with such an extreme L2 regularisation, it is almost equivalent to fPLS (with a proportionality factor of $\frac{1}{1+\lambda_2}$ on weight vectors), as can be seen in Table 2.

As for the second component pair, the out-of-sample correlation coefficient obtained by fsPLS is lower than on the first component pair. However for all the other PLS-based methods, the correlation appears to be slightly higher on the second component pair than on the first one. This may be explained by the fact that once the noise leading to overfitting on the first component pair has been removed, some real effects may be observed on further components, while on the opposite, fsPLS prevents from overfitting and can capture some effects on both pairs of components. Finally, MULM and PCA + KCCA do not seem able to capture any generalisable effects on any of the component pairs.

Table 3
Out-of-sample correlation coefficient on the first component pair as a function of k and $s_{\lambda_{1X}}$. Empirical p-values still significant ($p<.05$) after correction are shown here as: *.

		$s_{\lambda_{1X}}$						
		1%	5%	10%	25%	50%	75%	100%
k	10	0.041	0.041	0.041	0.041	0.144	0.112	0.112
	100	0.182	0.074	0.085	0.057	0.069	0.188	0.243
	1000	0.151	**0.432** *	**0.414** *	0.400	0.317	0.285	0.236
	10000	0.004	0.120	0.130	0.027	−0.006	−0.031	−0.061

4.2.2. Sensitivity analysis of fsPLS and significance assessment

We now detail the sensitivity analysis we performed in order to assess how much the performance of fsPLS is influenced by the sparse PLS penalisation parameter $s_{\lambda_{1X}}$ and by the number k of SNPs kept by the filter, and to select the best pair of parameters. Indeed, we tested different values for the number k of SNPs to be kept by the univariate filter: the 10, 100, 1000 and 10000 "best" ranked SNPs. Seven different sPLS selection rates $s_{\lambda_{1X}}$ were also tested on SNPs ($\mathbf{X}$): 1, 5, 10, 25, 50, 75 and 100%. For instance, when considering 1000 SNPs kept after univariate filtering, the 75% condition means that only 750 SNPs will have non-zero PLS weights. The 10-fold cross-validation procedure presented in 3.7 was repeated for each pair of parameters (k, $s_{\lambda_{1X}}$).

Moreover, we calibrated the degree of significance of the out-of-sample correlation coefficients thus obtained using a randomisation procedure where, at each permutation, the rows of $\mathbf{Y}$ were permuted and the cross-validation procedure was repeated on the permuted dataset for each pair of parameters. We performed 1000 permutations in order to get a good estimation of the empirical p-values. We then corrected our empirical p-values for multiple comparisons, because of the different pairs of parameters tested, using a maxT procedure which derives corrected p-values from the empirical distribution of the maximal statistic over tests (Westfall and Young, 1993). Table 3 summarises the out-of-sample correlation coefficient obtained for the first pair of components using fsPLS, together with its statistical significance, as a function of k and $s_{\lambda_{1X}}$. One can see in Table 3 that the best out-of-sample correlation coefficient of 0.43, obtained with $k=1000$ and $s_{\lambda_{1X}}=5\%$, happens to be significant after correction ($p=0.034$). The second best out-of-sample correlation coefficient of 0.41 with $k=1000$ and $s_{\lambda_{1X}}=10\%$ is significant as well ($p=0.043$). Out-of-sample correlation coefficients were not significant for the second component pair.

4.3. Imaging genetics findings

In order to obtain the SNPs and the brain phenotypes involved in the link between the two blocks, we then applied fsPLS on all the subjects simultaneously for the pair of parameters giving the most significant results on the first component pair: 1000 SNPs selected with the univariate filter and a sPLS selection rate of 5% ($k=1000$, $s_{\lambda_{1X}}=0.05$). It should noted at this point that the significance of the multivariate model has to be considered as a whole and not SNP by SNP, thus we have to be very careful with the interpretation of the results.

After the univariate step, one may observe that each phenotype is associated with at least one of the 1000 best ranked SNPs, if one refers to the univariate p-values (Fig. 5). This reinforces the idea that the problem is multivariate on both the imaging and genetic sides and that there may exist interactions both between SNPs and between phenotypes, which suggests that the second step of multivariate analysis is useful.

Figs. 6 and 7 provide an illustration of the sPLS weights of SNPs and phenotypes in the genetic and imaging components respectively, after this second step. The two intra-block correlation matrices are shown below the graphs. One may notice that all phenotypes do not contribute to the same extent to the first component and that there seems to be to a stronger involvement of the phenotypes obtained from the "reading" contrast.

Figs. 8 and 9 show the location of the selected SNPs and of the phenotypes respectively. The distribution of the 1000 SNPs having the lowest univariate p-values along the 22 autosomes is illustrated in Fig. 8. The 5% of those SNPs that were selected by sPLS are highlighted in red. As can be seen, they spread over all autosomal chromosomes and some of them seem to be in linkage disequilibrium. Among the 50 SNPs selected by fsPLS, some of them were located within a gene (see Table 4). Eighteen genes were thus identified (Table 5), such as PPP2R2B and RBFOX1, which have been reported to be linked with ataxia and a poor coordination of speech and body

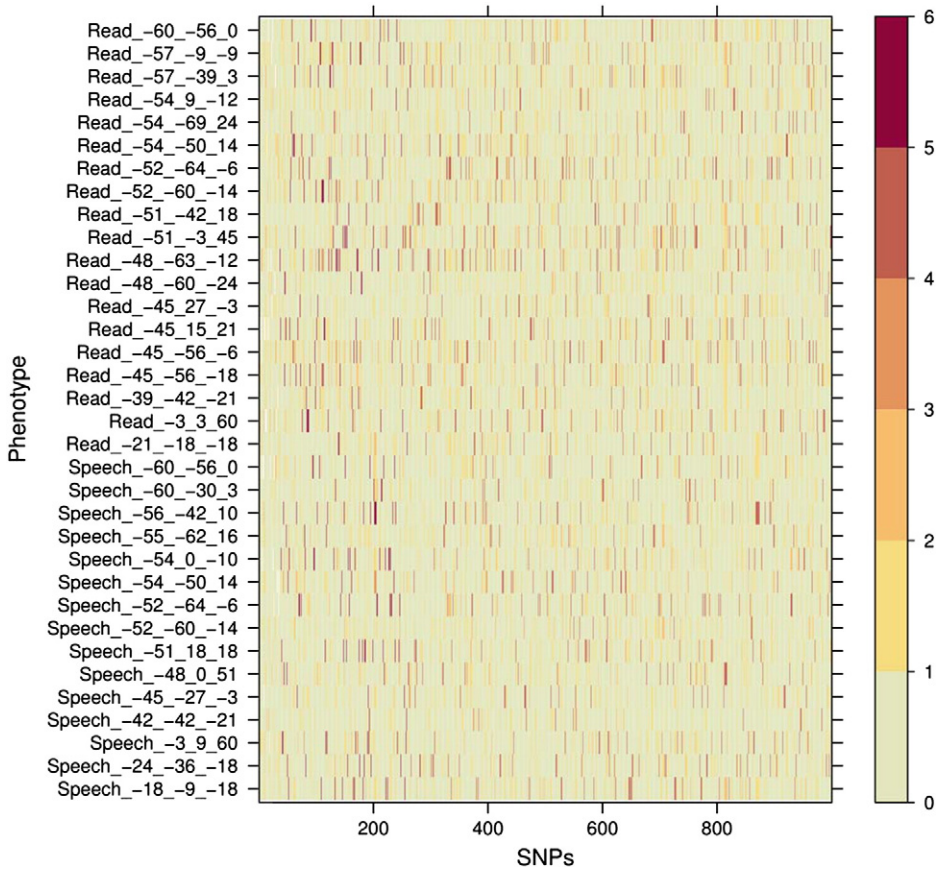

Fig. 5. Distribution of the *p*-values ($-\log(p)$) for the 1000 best ranked SNPs after univariate filtering with each of the 34 phenotypes (MNI coordinates are reported in brackets for the corresponding task, Reading or Speech comprehension).

movements, or also PDE4B which has been associated with schizophrenia and bipolar disorder.

Fig. 9 shows the location of the phenotypes where lateralisation indexes were computed, for both contrast maps of interest "reading" and "speech comprehension". The weights assigned by sPLS to the imaging phenotypes are illustrated according to the colourbar. The phenotypes that obtained the largest weights (in absolute value) mainly come from the "reading" contrast, especially from the temporal lobe.

Taken altogether, our results show that fsPLS could establish a significant link between a subset of SNPs distributed across the genome and a functional brain network activated during a reading task, some of these SNPs being probably indirectly linked to the neuroimaging phenotypes due to linkage disequilibrium. This suggests that individual variability in the entire genome contains predictors of the observed variability in brain activation during language tasks.

5. Discussion

5.1. Performance of the two-step method fsPLS

The originality of this work is to investigate a two-step approach combining univariate filtering with sparse PLS and to show that it performs much better than the other regularisation or dimension reduction strategies combined with PLS or KCCA on both simulated and real high-dimensional imaging genetics data. Indeed even though sparse PLS performs better than PLS and (regularised) KCCA, it does not seem able to overcome the overfitting issue by itself, which suggests that a first step of dimension reduction is also necessary. Univariate filtering appears to be the best solution, especially when combined with sPLS, while PC-based methods fail in that respect. Moreover, our results on the experimental dataset show that fsPLS was sensitive enough to uncover a generalisable and significant multivariate link between genetic and neuroimaging data.

5.2. Influence of the parameters of univariate filtering and L1 regularisation

We performed a sensitivity analysis in order to assess the influence of the parameters of univariate filtering and sPLS selection on the generalisability of the link found by fsPLS between the two blocks of data, which explains why we repeated the cross-validation procedure for all pairs of parameters. We also tried to add a nested CV loop in order to select, at each fold of our external 10-fold CV, the best pair of model parameters (filtering and sPLS selection rate) corresponding to that fold. The role of the external 10-fold CV then became the assessment of the generalisability of the whole procedure: fsPLS and parameter selection. But because of the computational load of such a procedure, we could not assess by permutations the significance of the out-of-sample correlation coefficient of the external CV.

Our main results on the experimental dataset show that fsPLS extracted the most generalisable and significant neuroimaging/genetics link when considering 1000 SNPs after univariate filtering and 5% of these SNPs selected by sPLS. The intersection between the 50 best SNPs after the univariate ranking step and of the 50 SNPs finally selected by fsPLS is of 6 SNPs. Those results as well as those obtained on simulated data raise the question of the relative contribution of the univariate filtering and the sparsity constraint to select relevant features. A relatively large number of SNPs kept after filtering seems to be required, up to a trade-off between the numbers of true and false positives, to allow sPLS to extract a robust association between a multivariate

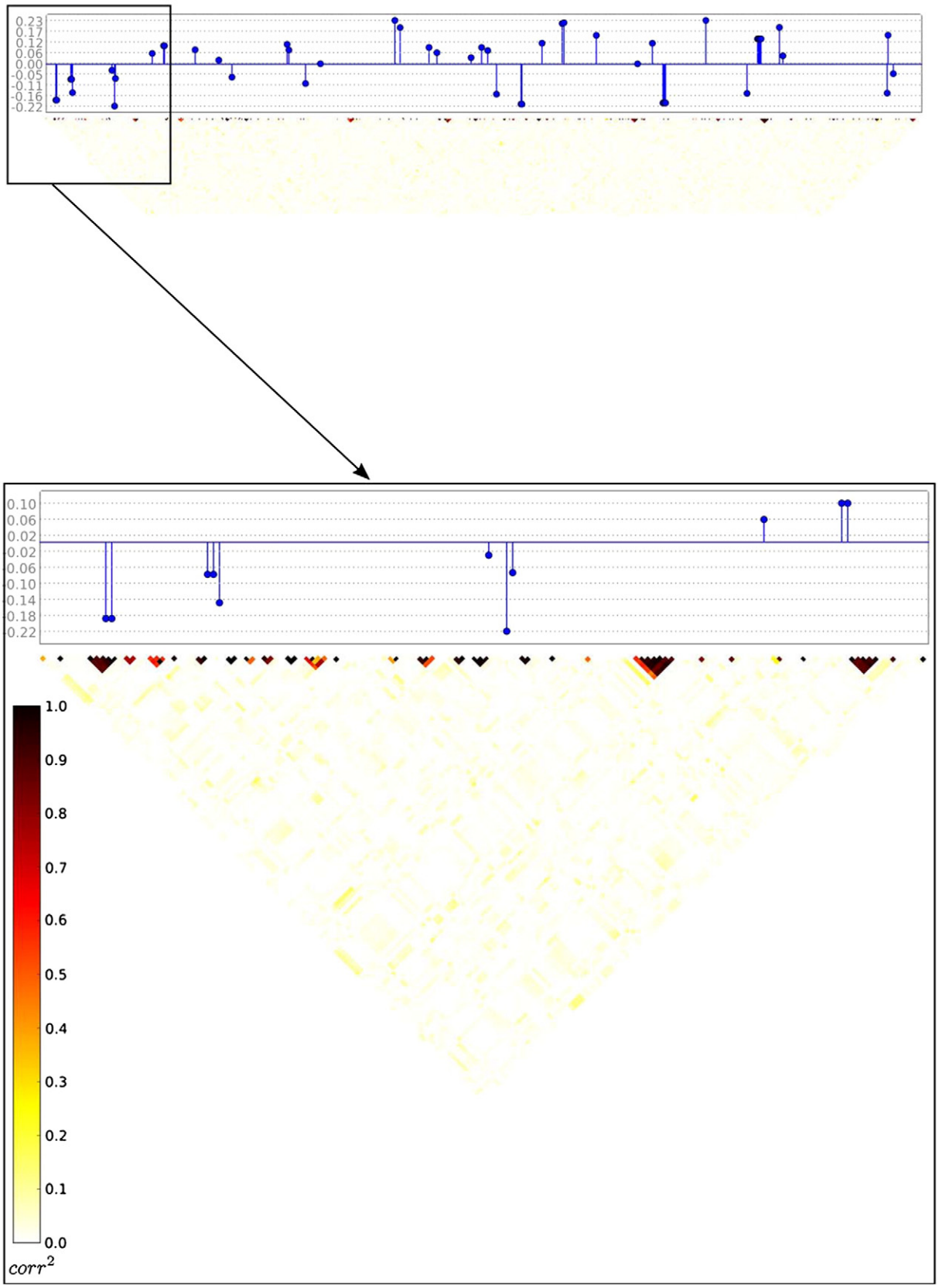

Fig. 6. sPLS weights for SNPs, when considering $k = 1000$ SNPs ordered here according to their position along the genome, with $s_{\lambda_{1X}} = 5\%$ of selected SNPs. Here we zoom only on the first 150 SNPs for visualisation purposes. The matrix of squared pairwise correlations is shown below.

pattern of SNPs and a multivariate neuroimaging pattern. However, univariate filtering appears to be a mandatory step to filter out the vast majority of irrelevant features. Indeed, the results on both simulated and experimental datasets demonstrated that a looser threshold on filtering (more than 1000 SNPs) always leads to an overfitting behaviour of PLS regardless of sparsity. Another reason to

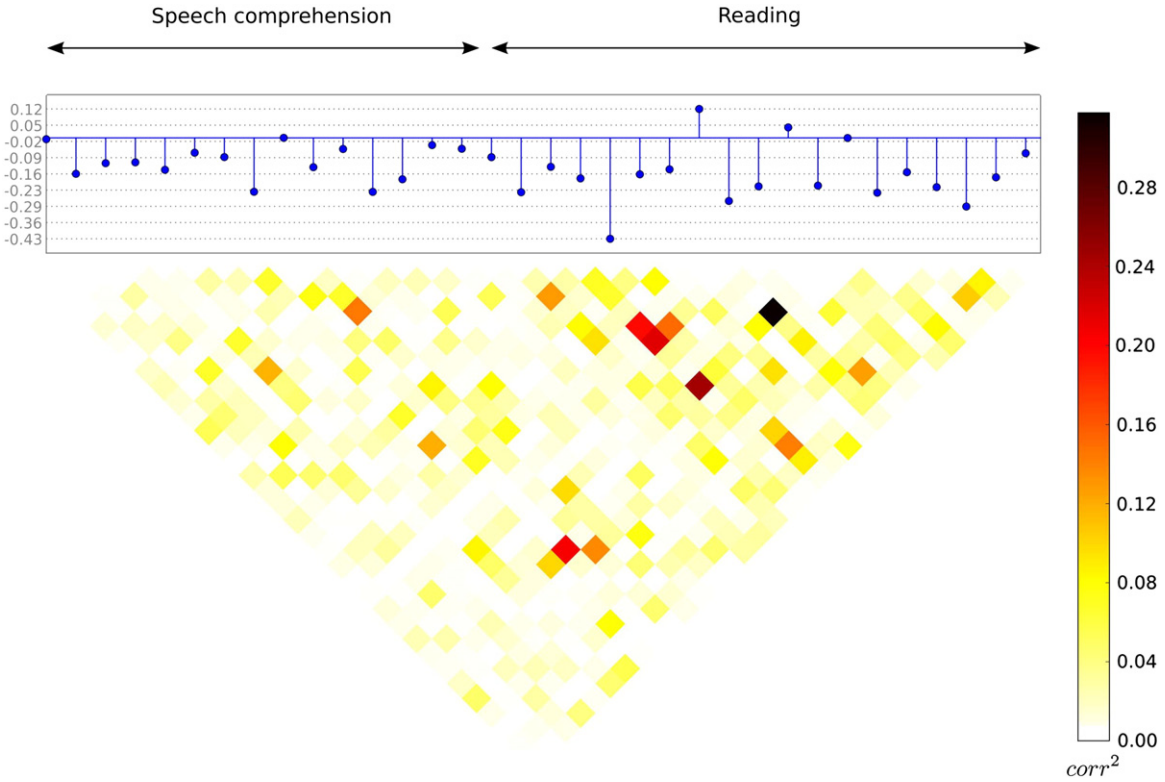

Fig. 7. sPLS weights for phenotypes, when considering $k = 1000$ SNPs with $s_{\lambda_{1X}} = 5\%$ of selected SNPs. The matrix of squared pairwise correlations is shown below.

perform univariate filtering is that PLS and even sparse PLS are too sensitive to a large number of irrelevant features, as they try to explain the variance of each block while they try to find some link between the blocks. Indeed, let us remind the criterion that is maximised by PLS regression:

$$max_{\|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1} \underbrace{\mathrm{corr}(\mathbf{Xu},\mathbf{Yv})}_{\text{Inter-block corr}} \underbrace{\sqrt{\mathrm{var}(\mathbf{Xu})}}_{\text{Intra-block stdev}} \underbrace{\sqrt{\mathrm{var}(\mathbf{Yv})}}_{\text{Intra-block stdev}},$$

where the first term is the inter-block correlation between the two latent variables of each block and the two last terms the intra-block standard deviations of the latent variable of each block. In the case of very large blocks, the two terms of intra-block standard deviations weigh too much compared to the term of inter-block correlation, as discussed by Tenenhaus and Tenenhaus (2011). Univariate filtering helps to solve this problem by reducing the number of SNPs and selecting the ones that are more correlated to the imaging phenotypes.

5.3. Potential limitations of fsPLS

However, although common practice in genome wide association studies, univariate tests may not be the best filter and it could be interesting to consider multivariate filters that account for specific interactions between potential predictors (e.g., for a review

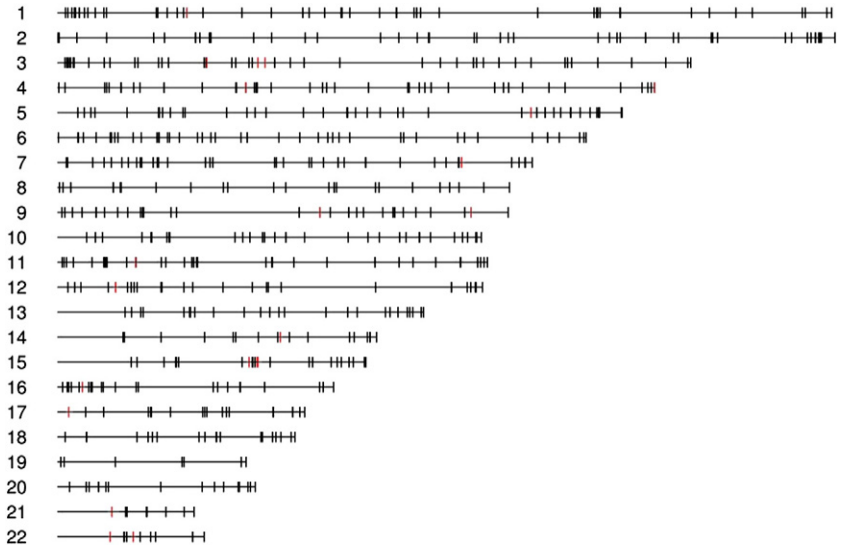

Fig. 8. Distribution of the 1000 most significant SNPs (univariate tests) across the genome. The 50 SNPs selected by sPLS are highlighted in red.

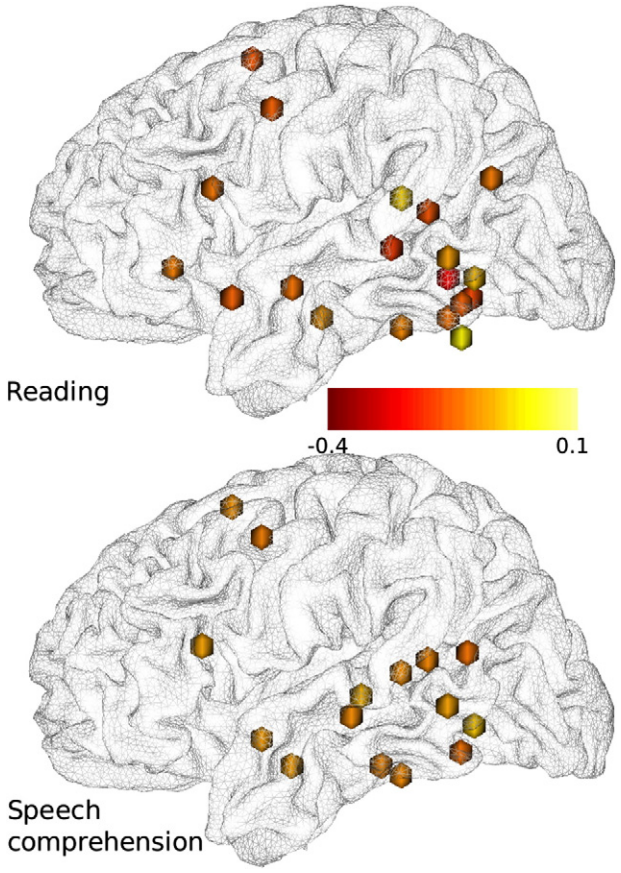

Fig. 9. Location of the 19 phenotypes extracted from the "reading" contrast map and the 15 phenotypes extracted from the "speech comprehension" contrast map. The weights assigned by sPLS to the phenotypes are illustrated according to the colourbar. (The signal that appears outside of the cortical surface belongs to the cerebellum.)

Díaz-Uriarte and Alvarez de Andrés, 2006). For instance a limitation of univariate filtering may be that it filters out suppressor variables. Indeed such variables are useful to remove the non-specific variability of the relevant SNPs, improving their predictive power, while being themselves not correlated (and thus not detectable) with imaging phenotypes.

As for penalisation, even though it is well known that it plays an important role when trying to uncover specific relationships among high-dimensional data, the choice of the penalisation is also important. For instance, an L1, L2 or L1-L2 (elastic net) penalisation scheme does not give rise to the same results when data are correlated. Indeed in the case of correlated variables grouping into a few clusters, L1 penalisation tends to select one "representative" variable of each cluster, which facilitates the interpretation of the results but may lead to an unstable solution, whereas L2 penalisation and the elastic net criterion tend to emphasise the whole set of correlated predictors. However, in our case, we observed that L1 penalisation could not offset the PLS tendency to select blocks of correlated variables, as PLS tends to maximise at each step the variances of the latent variables of each block while maximising the correlation between them. Indeed, in Fig. 6 one may observe one part of the correlation matrix of the 1000 best ranked SNPs ordered according to their position along the genome and the weights of the 5% of these SNPs that are selected by sPLS are shown in blue. It shows that sPLS still tends to select several SNPs from the same block (dark red blocks) that are spatially correlated due to linkage disequilibrium (LD). One could investigate a more sophisticated penalisation that takes into account the correlation structure of the data. Then, in Fig. 7, we plotted the correlation matrix of the 34 phenotypes. We may notice that there exists a structure of correlation between the variables obtained from the "reading" contrast (the last 19 variables of the matrix) which happen to be the variables that got the largest weights.

It should be noticed that such multivariate methods do not provide any variable-wise degree of significance or any explicit control for false positives. In further work, selection stability could be investigated instead, using bootstrapping techniques for instance. Another limitation of our method may be that on the experimental dataset it could not distinguish between different pairs of covarying sub-networks on the first pair of PLS components. Even on further dimensions, subtle sub-networks were not visible in such high-dimensional settings.

Moreover, it should be noted that some non-linear effects of the number of minor alleles may also be missed by fsPLS, with the additive genetic coding that we used. A different genetic coding, such as

Table 4
SNPs selected by fsPLS.

Reference SNP ID	Ensembl Gene ID	Location within gene	Chrom.	Position
rs13047077	C21orf34	Within Non Coding Gene	21	17794297
rs2070477	C22orf36	Upstream	22	24990916
rs5751901	C22orf36	Upstream	22	24992266
rs5760489	C22orf36	Upstream	22	24990646
rs6519519	C22orf36	Upstream	22	24991863
rs874852	C22orf36	Intronic	22	24987964
rs1894702	F5	Intronic	1	169530837
rs3934552	FBXL22	Downstream	15	63894400
rs12891349	GALNTL1	Intronic	14	69790389
rs4902713	GALNTL1	Intronic	14	69770939
rs8017671	GALNTL1	Intronic	14	69771213
rs2070477	GGT1	Intronic	22	24990916
rs5751901	GGT1	Intronic	22	24992266
rs5760489	GGT1	Intronic	22	24990646
rs5760492	GGT1	Intronic	22	24995202
rs6519519	GGT1	Intronic	22	24991863
rs874852	GGT1	Intronic	22	24987964
rs10519223	HERC1	Intronic or Splice Site	15	63935149
rs11630720	HERC1	Intronic or Splice Site	15	63984772
rs11635117	HERC1	Intronic	15	64112732
rs2228510	HERC1	Non Synonymous Coding	15	63970456
rs3764186	HERC1	Intronic	15	64056437
rs8034342	HERC1	Intronic	15	64038870
rs8034675	HERC1	Intronic	15	64039050
rs9972527	HERC1	Upstream	15	64127531
rs564249	HPCAL4	Intronic	1	40155623
rs2187522	NELL1	Intronic	11	21357112
rs4257797	ODZ2	Intronic	5	166869195
rs7688580	PAPSS1	Intronic	4	108518005
rs12081185	PDE4B	Intronic	1	66321193
rs4609402	PDE4B	Intronic	1	66318628
rs6684621	PDE4B	Intronic	1	66315450
rs1480149	PPP2R2B	Intronic	5	146448551
rs1480150	PPP2R2B	Intronic	5	146454825
rs6580448	PPP2R2B	Intronic	5	146438035
rs6872842	PPP2R2B	Upstream	5	146462839
rs1871394	PTPRG	Intronic	3	61931534
rs12598550	RBFOX1	Intronic	16	7683677
rs3785228	RBFOX1	Intronic	16	7679580
rs999566	RP11-343J18.2	Within Non Coding Gene	9	128835802
rs439339	SLC13A3	Intronic	20	45238334
rs7178762	USP3	Intronic	15	63871292
rs10834273			11	24091682
rs11043662			12	17928813
rs13086717			3	46139499
rs1480162			5	146471808
rs1534101			7	125149625
rs17680472			13	71273644
rs2120252			15	64136472
rs4241767			4	184353138
rs4341595			12	18033101
rs4477486			12	17939279
rs4820001			22	17827684
rs4865243			4	58421116
rs7044535			9	81969574

Table 5
Genes selected by fsPLS.

Gene name	Function
C21orf34	Non-coding
C22orf36	Unknown
F5	Central regulator of hemostasis. It serves as a critical cofactor for the prothrombinase activity of factor Xa that results in the activation of prothrombin to thrombin
FBXL22	Recognises and binds to some phosphorylated proteins and promotes their ubiquitination and degradation
GALNTL1	May catalyse the initial reaction in O-linked oligosaccharide biosynthesis, the transfer of an *N*-acetyl-D-galactosamine residue to a serine or threonine residue on the protein receptor (By similarity) GGT1 Initiates extracellular glutathione (GSH) breakdown, provides cells with a local cysteine supply and contributes to maintain intracellular GSH level. It is part of the cell antioxidant defense mechanism. Catalyses the transfer of the glutamyl moiety of glutathione to amino acids and dipeptide acceptors.
HERC1	This protein is thought to be involved in membrane transport processes.
HPCAL4	May be involved in the calcium-dependent regulation of rhodopsin phosphorylation
NELL1	Involved in the control of cell growth and differentiation
ODZ2	May function as a cellular signal transducer
PAPSS1	Bifunctional enzyme with both ATP sulfurylase and APS kinase activity, which mediates two steps in the sulfate activation pathway
PDE4B	Hydrolyses the second messenger cAMP, which is a key regulator of many important physiological processes. May be involved in mediating central nervous system effects of therapeutic agents ranging from antidepressants to antiasthmatic and anti-inflammatory agents
PPP2R2B	The B regulatory subunit might modulate substrate selectivity and catalytic activity, and also might direct the localisation of the catalytic enzyme to a particular subcellular compartment. Defects in this gene cause autosoma dominant spinocerebellar ataxia 12 (SCA12), a disease caused by degeneration of the cerebellum, sometimes involving the brainstem and spinal cord, and in resulting in poor coordination of speech and body movements.
PTPRG	Possesses tyrosine phosphatase activity
RBFOX1	RNA-binding protein that regulates alternative splicing events by binding to 5′-UGCAUGU-3′ elements. Regulates alternative splicing of tissue-specific exons and of differentially spliced exons during erythropoiesis. This protein binds to the C-terminus of ataxin-2 and may contribute to the restricted pathology of spinocerebellar ataxia type 2 (SCA2). Ataxin-2 is the gene product of the SCA2 gene which causes familial neurodegenerative diseases.
RP11-343J18.2	Non-coding
SLC13A3	High-affinity sodium-dicarboxylate cotransporter that accepts a range of substrates with 4–5 carbon atoms
USP3	Hydrolase that deubiquitinates monoubiquitinated target proteins such as histone H2A and H2B. Required for proper progression through S phase and subsequent mitotic entry. May regulate the DNA damage response (DDR) checkpoint through deubiquitination of H2A at DNA damage sites. Associates with the chromatin

dominant/recessive or genotypic coding, could be investigated in further work.

5.4. Conclusion

To conclude, in this study, we investigated a two-step method combining univariate filtering and sparse PLS, called fsPLS, and we showed that it performed much better than other regularisation or dimension reduction strategies combined with PLS or KCCA, on both simulated and real high-dimensional imaging genetics data. Moreover, on the experimental dataset, it allowed us to detect a significant link between a set of SNPs and a functional brain network activated during a reading task, in a whole genome analysis framework. This suggests that individual variability in the genome contains predictors of the observed variability in brain activation during language tasks. We showed that we could generalise our model on left out subjects, and that this two-step multivariate technique is useful to select associated SNPs that may not be detected by a univariate screening only. However the interpretation of the results is still a very difficult issue and the neuroscientific relevance of these findings should be investigated in further research. As for the fsPLS method itself, more elaborated filtering rules and more sophisticated types of penalisation should also be investigated, which could hopefully help for the interpretation of the results.

Acknowledgments

This work was supported by CEA and the Karametria grant for the French National Agency for Research (ANR). Support was also partially provided by the IMAGEN project, which receives research funding from the European Community's Sixth Framework Programme (LSHM-CT-2007-037286). This manuscript reflects only the author's views and the Community is not liable for any use that may be made of the information contained therein.

Appendix A

Multivariate Reduced-Rank Regression (RRR) (Reinsel and Velu, 1998) consists in transforming the classical multivariate multiple linear regression model of a $n*q$ response matrix $\mathbf{Y}$ on a $n*p$ design matrix $\mathbf{X}$, by imposing a rank $R \leq \min(p, q)$ on regression coefficients and taking into account the multivariate nature of the response matrix. The criterion optimised by multivariate RRR is:

$$\hat{\mathbf{U}}, \hat{\mathbf{V}} = \underset{\mathbf{U},\mathbf{V}}{arg\ min}\ \ \mathrm{Tr}\{(\mathbf{Y}-\mathbf{XUV})\boldsymbol{\Gamma}(\mathbf{Y}-\mathbf{XUV})\} \tag{8}$$

where regression coefficients are decomposed into a matrix $\mathbf{U}$ with R linearly independent columns and a matrix $\mathbf{V}$ with R linearly independent rows. Γ is a weight matrix, commonly set to be the identity matrix. The solutions for $\mathbf{U}$ and $\mathbf{V}$ are derived from the Singular Value Decomposition (SVD) of the matrix $\left(\mathbf{X}'\mathbf{X}\right)^{-\frac{1}{2}}\mathbf{X}'\mathbf{Y}\boldsymbol{\Gamma}^{\frac{1}{2}}$.

In the implementation of sparse (multivariate) RRR by Vounou et al. (2010), $\boldsymbol{\Gamma}$ is set to be the identity matrix and $\mathbf{X}'\mathbf{X}$ is approximated by the identity matrix because of its very high dimensionality, which makes RRR equivalent to PLS–SVD. However, instead of performing an SVD in high-dimensional settings, they recast the PLS–SVD problem into an iterative procedure using NIPALS algorithm and they apply L1-penalisation on weight vectors $\mathbf{u}$ and $\mathbf{v}$ for each new rank, using soft-thresholding within the NIPALS inner-loop. Indeed, for the rank-one model, the criterion optimised becomes:

$$\hat{\mathbf{u}}, \hat{\mathbf{v}} = \underset{\mathbf{u},\mathbf{v}}{arg\ min} -2\mathbf{v}\mathbf{Y}'\mathbf{X}\mathbf{u} + \mathbf{v}\mathbf{v}'\mathbf{u}'\mathbf{u} + \lambda_1\|\mathbf{u}\|_1 + \lambda_2\|\mathbf{v}'\|_1 \tag{9}$$

Further ranks are obtained by optimising the same criterion on the residuals of the matrices $\mathbf{X}$ and $\mathbf{Y}$ after regression on their own latent variables, which departs from the PLS–SVD problem and becomes equivalent to PLS regression in its canonical mode. In conclusion, the sparse multivariate RRR approach, under some approximations commonly made in high dimensional settings, becomes equivalent to the sparse PLS approach. This suggests that such multivariate methods may be appropriate to exploratory imaging genetics studies.

References

Calhoun, V., Liu, J., Adali, T., 2009. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. NeuroImage 45 (Suppl. 1), S163–S172.

Chun, H., Keleş, S., 2010. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J. R. Stat. Soc. B 72 (1), 3–25.
Clayton, D., Cheung, H.-T., 2007. An R package for analysis of whole-genome association studies. Hum. Hered. 64, 45–51.
de Bakker, P., Yelensky, R., Peer, I., Gabriel, S., Daly, M., Altshuler, D., 2005. Efficiency and power in genetic association studies. Nat. Genet. 37 (11), 1217–1223.
Díaz-Uriarte, R., Alvarez de Andrés, S., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinforma. 7 (3).
Furlanello, C., Serafini, M., Merler, S., Jurman, G., 2003. An accelerated procedure for recursive feature ranking on microarray data. Neural Netw. 16, 641–648.
Glahn, D.C., Thompson, P.M., Blangero, J., 2007. Neuroimaging endophenotypes: strategies for finding genes influencing brain structure and function. Hum. Brain Mapp. 28, 488–501.
Feature Extraction: Foundations and Applications. In: Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A. (Eds.), Springer-Verlag.
Hibar, D., Stein, J., Kohannim, O., Jahanshad, N., Saykin, A., Shen, L., Kim, S., Pankratz, N., Foroud, T., Huentelman, M., Potkin, S., Jack Jr., C., Weiner, M., Toga, A., Thompson, P., the Alzheimer's Disease Neuroimaging Initiative, 2011. Voxelwise gene-wide association study (vgenewas): multivariate gene-based association testing in 731 elderly subjects. NeuroImage 56, 1875–1891.
Hotelling, H., 1936. Relations between two sets of variates. Biometrika 28, 321–377.
Lê Cao, K.-A., Rossouw, D., Robert-Granié, C., Besse, P., 2008. A sparse PLS for variable selection when integrating omics data. Stat. Appl. Genet. Mol. Biol. 7 (1).
Lê Cao, K.-A., Martin, P.G., Robert-Granié, C., Besse, P., 2009. Sparse canonical methods for biological data integration: application to a cross-platform study. BMC Bioinforma. 10 (34).
Li, J., Chen, Y., 2008. Generating samples for association studies based on hapmap data. BMC Bioinforma. 9 (44).
McAllister, T.W., Flashman, L.A., McDonald, B.C., Saykin, A.J., 2006. Mechanisms of cognitive dysfunction after mild and moderate TBI (MTBI): evidence from functional MRI and neurogenetics. J. Neurotrauma 23 (10), 1450–1467.
McIntosh, A., Bookstein, F., Haxby, J., Grady, C., 1996. Spatial pattern analysis of functional brain images using partial least squares. NeuroImage 3, 143–157.
Parkhomenko, E., Tritchler, D., Beyene, J., 2007. Genome-wide sparse canonical correlation of gene expression with genotypes. BMC Proc. 1 (Suppl. 1), S119.
Parkhomenko, E., Tritchler, D., Beyene, J., 2009. Sparse canonical correlation analysis with application to genomic data integration. Stat. Appl. Genet. Mol. Biol. 8 (1) Article 1.
Paulesu, E., Demonet, J.-F., Fazio, F., McCrory, E., Chanoine, V., Brunswick, N., Cappa, F., Cossu, G., Habib, M., Frith, C., Frith, U., 2001. Dyslexia: cultural diversity and biological unity. Science 291, 2165–2167.
Pinel, P., Dehaene, S., 2009. Beyond hemispheric dominance: brain regions underlying the joint lateralization of language and arithmetic to the left hemisphere. J. Cogn. Neurosci. http://dx.doi.org/10.1162/jocn.2009.21184 Posted Online January 13, 2009.
Pinel, P., Thirion, B., Meriaux, S., Jobert, A., Serres, J., Le Bihan, D., Poline, J.-B., Dehaene, S., 2007. Fast reproducible identification and large-scale databasing of individual functional cognitive networks. BMC Neurosci. 8 (91).
R Development Core Team, 2009. R: A language and environment for statistical computing. http://www.R-project.org.
Reinsel, G., Velu, R., 1998. Multivariate Reduced-Rank Regression, Theory and Applications. Springer, New York.
Roffman, J.L., Weiss, A.P., Goff, D.C., Rauch, S.L., Weinberger, D.R., 2006. Neuroimaging-genetic paradigms: a new approach to investigate the pathophysiology and treatment of cognitive deficits in schizophrenia. Harv. Rev. Psychiatry 14 (2), 78–91.
Sanna, S., Jackson, A., Nagaraja, R., Willer, C., Chen, W., Bonnycastle, L., Shen, H., Timpson, N., Lettre, G., Usala, G., Chines, P., Stringham, H., Scott, L., Dei, M., Lai, S., Albai, G., Crisponi, L., Naitza, S., Doheny, K., Pugh, E., Ben-Shlomo, Y., Ebrahim, S., Lawlor, D., Bergman, R., Watanabe, R., Uda, M., Tuomilehto, J., Coresh, J., Hirschhorn, J., Shuldiner, A., Schlessinger, D., Collins, F., Davey Smith, G., Boerwinkle, E., Cao, A., Boehnke, M., Abecasis, G., Mohlke, K., 2008. Common variants in the GDF5-UQCC region are associated with variation in human height. Nat. Genet. 40, 198–203.
Soneson, C., Lilljebjörn, H., Fioretos, T., Fontes, M., 2010. Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. BMC Bioinforma. 11 (191).
Stein, J., Hua, X., Lee, S., Ho, A., Leow, A., Toga, A., Saykin, A., Shen, L., Foroud, T., Pankratz, N., Huentelman, M., Craig, D., Gerber, J., Allen, A., Corneveaux, J., DeChairo, B., Potkin, S., Weiner, M., Thompson, P., 2010. Voxelwise genome-wide association study (vGWAS). NeuroImage 53, 1160–1174.
Tenenhaus, A., Tenenhaus, M., 2011. Regularized generalized canonical correlation analysis. Psychometrika 76 (2), 257–284.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B 58 (1), 267–288.
Tucker, L., 1958. An inter-battery method of factor analysis. Psychometrika 23 (2), 111–136.
Vounou, M., Nichols, T., Montana, G., 2010. Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank approach. NeuroImage 53, 1147–1159.
Waaijenborg, S., Verselewel de Witt Hamer, P., Zwinderman, A., 2008. Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. Stat. Appl. Genet. Mol. Biol. 7 (1) (Article 3).
Westfall, P., Young, S. (Eds.), 1993. Resampling-Based Multiple Testing. Wiley, New York.
Willer, C., Sanna, S., Jackson, A., Scuteri, A., Bonnycastle, L., Clarke, R., Heath, S., Timpson, N., Najjar, S., Stringham, H., Strait, J., Duren, W., Maschio, A., Busonero, F., Mulas, A., Albai, G., Swift, A., Morken, M., Narisu, N., Bennett, D., Parish, S., Shen, H., Galan, P., Meneton, P., Hercberg, S., Zelenika, D., Chen, W., Li, Y., Scott, L., Scheet, P., Sundvall, J., Watanabe, R., Nagaraja, R., Ebrahim, S., Lawlor, D., Ben-Shlomo, Y., Davey-Smith, G., Shuldiner, A., Collins, R., Bergman, R., Uda, M., Tuomilehto, J., Cao, A., Collins, F., Lakatta, E., Lathrop, G., Boehnke, M., Schlessinger, D., Mohlke, K., Abecasis, G.R., 2008. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat. Genet. 40, 161–169.
Witten, D., Tibshirani, R., 2009. Extensions of sparse canonical correlation analysis, with applications to genomic data. Stat. Appl. Genet. Mol. Biol. 8 (1) (Article 28).
Wold, H., 1966. Multivariate Analysis. Estimation of Principal Components and Related Models by Iterative Least Squares. Academic Press, New York, Ch, pp. 391–420.
Wold, S., Martens, H., Wold, H., 1983. The multivariate calibration problem in chemistry solved by the PLS method. In: Ruhe, A., Kastrøm, B. (Eds.), Proceedings Conference Matrix Pencils. : Vol. Lecture Notes in Mathematics. Springer-Verlag, pp. 286–293.
Zou, H., Hastie, T., Tibshirani, R., 2006. Sparse principal component analysis. J. Comput. Graph. Stat. 15 (2), 265–286.