Accepted Manuscript

Title: Strategies for integrated analysis in Imaging Genetics studies

Authors: Natàlia Vilor-Tejedor, Silvia Alemany, Alejandro Cáceres, Mariona Bustamante, Jesús Pujol, Jordi Sunyer, Juan R. González

PII: S0149-7634(17)30882-5
DOI: <https://doi.org/10.1016/j.neubiorev.2018.06.013>
Reference: NBR 3156

To appear in:

Received date: 29-11-2017
Revised date: 30-4-2018
Accepted date: 15-6-2018

Please cite this article as: Vilor-Tejedor N, Alemany S, Cáceres A, Bustamante M, Pujol J, Sunyer J, González JR, Strategies for integrated analysis in Imaging Genetics studies, *Neuroscience and Biobehavioral Reviews* (2018), <https://doi.org/10.1016/j.neubiorev.2018.06.013>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Title: Strategies for integrated analysis in Imaging Genetics studies.

Authors:

Natàlia Vilor-Tejedor^{*1-4}; Silvia Alemany¹⁻³; Alejandro Cáceres¹⁻⁴; Mariona Bustamante¹⁻⁴; Jesús Pujol⁵; Jordi Sunyer^{1-3,6}; Juan R. González^{*1-3}

¹Barcelona Research Institute for Global Health (ISGlobal), Barcelona, Spain.

²Universitat Pompeu Fabra (UPF), Barcelona, Spain

³CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain.

⁴Center for Genomic Regulation (CRG), the Barcelona Institute of Science and Technology, Barcelona, Spain

⁵MRI Research Unit, Hospital del Mar, and Centro de Investigación Biomédica en Red de Salud Mental, CIBERSAM G21, Barcelona, Spain

⁶IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain

***Correspondence to:**

Natàlia Vilor-Tejedor, Center for Genomic Regulation (CRG). C. Doctor Aiguader 88, Edif. PRBB 08003 Barcelona, Spain. E-mail: natalia.vilortejedor@crg.eu

ORCID: 0000-0003- 4935-6721

Juan R. González, Barcelona Institute for Global Health (ISGlobal). C. Doctor Aiguader 88, 08003 Barcelona, Spain. E-mail: juanr.gonzalez@isglobal.org

ORCID: 0000-0003-3267-2146

Research highlights

What is the key question?

- Despite the increasing volume of research on Imaging Genetics, it is unclear how to statistically analyze the results of these studies.

What is the bottom line?

- Establish a clear consensus on the analytical procedures and methods implemented in Imaging Genetics.

Why read on?

- This is a systematic review that will help catalog methods and strategies that can serve as a reference for researchers in the field.

Abstract

Imaging Genetics (IG) integrates neuroimaging and genomic data from the same individual, deepening our knowledge of the biological mechanisms behind neurodevelopmental domains and neurological disorders. Although the literature on IG has exponentially grown over the past years, the majority of studies have mainly analyzed associations between candidate brain regions and individual genetic variants. However, this strategy is not designed to deal with the complexity of neurobiological mechanisms underlying behavioral and neurodevelopmental domains. Moreover, larger sample sizes and increased multidimensionality of this type of data represents a challenge for standardizing modeling procedures in IG research. This review provides a systematic update of the methods and strategies currently used in IG studies, and serves as an analytical framework for researchers working in this field. To complement the functionalities of the Neuroconductor framework, we also describe existing R packages

that implement these methodologies. In addition, we present an overview of how these methodological approaches are applied in integrating neuroimaging and genetic data.

Keywords: *Analytical strategies; genetics; imaging genetics; Neuroconductor; neuroimaging*

1. Introduction

Information from neuroimaging is used to understand neurodevelopment, cognition and behavior at the level of the brain. The effects of complex neurological diseases and behavioral processes are likely to be reflected in the structure and/or function of the brain (Kremen & Jacobson, 2010). In addition, some of the consistent brain responses to stimuli have a heritable basis (Bigos & Hariri, 2007; Blokland et al., 2011; Yancey, Venables, Hicks, & Patrick, 2013).

Genetics plays a pivotal role in brain structure and function, which in turn, are related to neurological diseases and behavioral processes. Brain imaging phenotypes offer an accurate measurement of biological mechanisms, all the way from gene-based pathways to neurodevelopmental domains. Moreover, it has been argued that genetic variation has a greater effect on brain imaging phenotypes than behavior. Therefore, magnetic resonance imaging (MRI)-based features can be considered as potential phenotypes (intermediate imaging markers) for a disease or neurodevelopmental domain (Gallinat, Bauer, & Heinz, 2008; Gottesman & Gould, 2003). Recent advances in neuroimaging and genetics have paved the way for further insights into neurodevelopmental domains and complex brain disorders by identifying novel imaging biomarkers and genetic risk variants. However, the combined analysis of both fields, known as Imaging Genetics (IG), provides even more potential and robustness. More specifically, by integrating neuroimaging and genomic data from the same individual, IG studies offer the opportunity to deepen our understanding of the biological mechanisms behind

neurodevelopmental domains and complex brain disorders (Meyer-Lindenberg, 2012; Meyer-Lindenberg & Weinberger, 2006). Thus, the main goal of IG consists of leveraging these interrelations to gain knowledge about neurological diseases that would otherwise have been untapped by studying neuroimaging information or genetics separately. The continual growth of both neuroimaging data and genetic information, along with the need to combine their analyses, has led to various methodological challenges (Nymberg et al., 2013). In this respect, although the literature on IG studies has exponentially grown in recent years, a clear consensus on analytical procedures has yet to emerge.

In this review, we systematically summarize the current analytical methods used in IG studies. We provide a catalogue of methods and strategies that can serve as a reference for researchers in the field, and explain how they are implemented in R. We also discuss several examples in which these methodological approaches are applied to integrate neuroimaging and genetic data.

2. Methods

2.1 Literature search and selection criteria

We searched articles indexed in PubMed (National Library of Medicine) using the following terms: “Imaging Genetics”, “Neuroimaging and Genetics”, or “Neuroimage and Genetics” combined with “Statistical Methods”, “Analytical strategies”, or “Multivariate Methods”. We only selected studies that met the following criteria: a) they included both genetic and neuroimaging data, b) they were written in English, and c) they were original research articles with accessible information. We excluded abstracts only, case reports, letters to the editor, and duplicated articles. Moreover, the search was limited to studies focused on the application of multivariate methods in IG, therefore other relevant clinical research or basic studies were out of the scope of the review.

2.2 Article summary and classification

A total of 367 articles were identified in PUBMED. Using these, we initially screened titles and abstracts only to identify all potentially eligible articles based on the following inclusion criteria: (a) the article must use multivariate statistical methods to analyze

genetic and neuroimaging data, and (b) the article must use both, genetic and neuroimaging data in the analysis. As a result of this screening, we evaluated the full-text of 41 articles. These articles are cataloged in Table 1, which highlights key information regarding the phenotype of interest, the neuroimaging modality, the analytical approach, and the main statistical method. All studies included in this review were performed in participants of European descent.

3. Results

3.1. *Statistical methods and strategies in Imaging Genetics*

IG studies typically look for three different types of genetic effects: (1) individual genetic variants that independently affect phenotypes/imaging markers, (2) multiple genetic variants that jointly affect one phenotype/imaging marker, and (3) multiple genetic variants that jointly affect multiple phenotypes/imaging markers. Of note, neuroimaging data can be considered as either an outcome (i.e., genetic variants versus imaging markers) or an independent variable (in combination with genetic data) related to an outcome.

From a statistical point of view, data involving multiple sources can be formulated as a multi-dataset framework problem. Indeed, this is the case for IG studies, in which neuroimaging and genetic data are treated as two independent datasets. Taking this as our starting point, we defined a classification scheme for the IG studies based on the analytical strategies and statistical methods used, the dimensionality of the data, and the existence or not of a hypothesis (a hypothesis-free design (e.g., entire genome) versus a hypothesis-driven design (e.g., targeted genetic markers)) [Figure 1].

3.2. *Data dimensionality*

3.2.1 *Analysis of individual genetic variants affecting one imaging marker*

One common strategy evaluates relationships between candidate single nucleotide polymorphisms (SNPs) and brain regions of interest (ROI)/measure taken at a given voxel level related to behavioral dimensions by fitting a univariate correlation/association analysis or performing an analysis of covariance (Hoogman et al., 2014). This type of strategy examines differences in brain structures (mean values) in relation to genotypic variation (Potkin et al., 2009; Shen et al., 2010). However, such

studies ignore the multidimensionality and complexity of data and are limited to identifying independent genetic effects (Liu & Calhoun, 2014; Nickl-Jockschat, Janouschek, Eickhoff, & Eickhoff, 2015).

3.2.2 Analysis of multiple genetic variants in relation to one imaging marker

Genome-wide association analysis

More recently, genome-wide association studies (GWAS) have been used to search the entire genome for genetic polymorphisms that influence brain structure (Adams et al., 2016; Hibar et al., 2015, 2017). GWAS has led to the discovery of associations between certain genetic variants and complex diseases, providing a better understanding of the genetic architecture of complex traits. However, GWAS is not considered to be the most feasible approach for IG as these types of studies require large samples to discover genetic effects that survive stringent multiple comparison corrections (P -values $< 10^{-8}$). An additional limitation is that a significant number of the SNPs reported in GWAS studies are often intergenic, and therefore create certain ambiguity when linking the causal gene to the phenotype (Visscher et al., 2017).

Gene-set analysis

Given that complex phenotypes are likely to be influenced by numerous polymorphic genes of common pathways, gene-set analysis (GSA) has yielded important biological insights in IG (Hibar et al., 2011; Inkster et al., 2010; Vilor-Tejedor, Gonzalez, & Calle, 2015; Khadka et al., 2016; Lin, Calhoun, & Wang, 2014a; Mattingsdal et al., 2013; Mous et al., 2015; Pérez-Palma et al., 2014). GSA aggregates the association evidence across a set of SNPs into one combined test statistic, and then tests whether this statistic is larger than what would be expected by chance. As most reported associations with common variants have small effect sizes, combining the effect of multiple variants can improve power (over other techniques) to detect genetic risk factors for complex diseases. However, apart from being limited by the biological knowledge of the disease, the results of GSA are also highly dependent on the definitions of the gene sets themselves and the statistical methods used. Thus, GSA is generally used as an exploratory analysis to contextualize and further explore the results obtained from GWAS (C. A. de Leeuw, Mooij, Heskes, & Posthuma, 2015; L. Wang, Jia, Wolfinger, Chen, & Zhao, 2011).

Polygenic risk scores

Some studies calculate polygenic risk scores (PRS) to quantify the aggregate genetic influence of a trait. A PRS is the sum of all “risk” alleles weighted by their effect size as previously determined in an independent GWAS. In this way, PRSs help identify neural mechanisms that are correlated with genetic risk. This individual score, which is basically an index of genetic susceptibility for a given disorder or trait, is then checked for associations with neuroimaging phenotypes. Since a cumulative effect may be detected if enough small effects are present, testing for cumulative effects may account for the genetic heterogeneity and polygenic architecture of a disease (Mooney & Wilmot, 2015). Moreover, PRSs may incorporate information from both GWAS hits and SNPs that are not significant at the genome-wide level. As a greater proportion of the phenotype variance is explained by PRS, they may represent an improvement over candidate gene and gene-set analyses (Dima & Breen, 2015; Wray et al., 2014). For instance, Whalley *et al.* (2012) found an association between the PRS for bipolar disorder and neural activation as assessed using functional MRI (fMRI). In addition, a composite multivariate score including both genetic variants and neuroimaging features was proposed as a method to predict Alzheimer’s disease (Filipovych, Gaonkar, & Davatzikos, 2012). However, a crucial aspect of polygenic risk scores approach concerns the significance threshold used as it determines the number of genetic variants examined (Dudbridge, 2013). The fact that different criteria is used to select one or more thresholds difficult comparability across studies.

Data-driven Methods

Multidimensionality reduction methods

Data-driven methods, such as feature selection methods and multidimensionality reduction (MDRed), are another type of approach used in IG. The use of such methods is motivated by the fact that the original datasets contain a large number of genetic and neuroimaging features. Thus, MDRed strategies factorize neuroimaging and genetic data sets into a meaningful representation of reduced dimensionality (McIntosh & Mišić, 2013). The reduction in dimensionality consists of finding the minimum number of parameters needed to account for the observed properties of the data, for the purpose of facilitating classification, visualization and comprehension of high-dimensional data (Van Der Maaten, Postma, & Van Den Herik, 2009).

Feature selection methods

Feature selection methods are used to select genetic variants that exhibit the strongest effect on the phenotype of interest. The most popular of these methods is the Least Absolute Shrinkage and Selection Operator (LASSO) multiple regression, which intrinsically performs feature selection and regularization, and shrinking coefficients of non-selected predictors towards zero (Tibshirani, 2011). LASSO is commonly used in genetic studies (Papachristou, Ober, & Abney, 2016; Pineda et al., 2014), and has also been used in IG (Kohannim et al., 2012).

In addition, some MDRed methods explore the multivariate structure of the data, and aim to select representative components. These components may reveal structures or patterns in the data that are difficult to identify *a priori*, and extract latent factors that represent a set of genetic variants/effects or neuroimaging-based features. To reduce the number of predictor variables and solve the multi-collinearity problem of a unique dataset, principal component analysis (PCA) is the traditional statistical method used. This multivariate method is based on maximizing the variability of the data (Ma & Dai, 2011). Another method is independent component analysis (ICA), which can be seen as an extension of PCA when the assumptions of orthogonality and normality do not hold. ICA creates a linear representation of the data based on independent sub-components, removing correlations and higher order dependence. In addition, ICA has greater potential and provides a more useful representation than PCA, specifically in neuroimaging studies.

Another MDRed strategy is multifactorial analysis (MFA). MFA is based on a dimensionality reduction model that focuses on determining the most significant features of the analyzed data. In the context of IG, the main goal of MFA is to evaluate how much the whole set of genetic and neuroimaging intermediate phenotypes contribute to the variability of the extracted latent components, which in turn is correlated with the disease outcome. MFA is a generalization of PCA that is tailored to handle multiple tables of data. Therefore, a remarkable advantage of MFA over PCA and ICA is its ability to reduce the dimensionality of the data and simultaneously combine multiple sources of information (e.g., both genetics and neuroimaging features) into the same analysis (Abdi, Williams, & Valentin, 2013; Eslami, Qannari, Kohler, & Bougeard, 2014).

Epistasis

The study of SNP-SNP and/or gene-gene interactions is another strategy used to explore the effects of genetic variation on brain structure and function, and human behavior and neurodevelopment. Current methods are based on screening pair-wise interactions. For instance, Hibar *et al.* (2013) found a significant gene-gene interaction associated with temporal lobe volume. Moreover, Meda *et al.* (2012) found a total of 109 significant interactions associated with right hippocampal atrophy, and 125 significant interactions associated with right entorhinal cortex atrophy. However, in addition to being difficult, the search for pairs of significant interactions also incurs high computational costs (Ehrenreich, 2017). Moreover, while most work on epistasis focuses on pairs of interacting variants (Bloom *et al.*, 2015), higher-order interactions can also be present (Taylor & Ehrenreich, 2014, 2015).

3.2.3 Analysis of multiple genetic variants in relation to multiple neuroimaging phenotypes

Currently, efforts in statistical modeling in IG studies seek to integrate multiple neuroimaging phenotypes (imaging-based features or neurodevelopmental domains) with multiple genotypes at the same time; this is in contrast to making pairwise associations or models of hierarchical structure. Unlike methods that reduce the dimensionality of genetic and/or neuroimaging datasets, joint multivariate methods simultaneously capture complex relationships between genes and brain measurements.

Massive univariate analysis

Massive univariate linear modeling (MULM) is the method of choice for modeling IG data because its results can be more easily interpreted. However, due to the high dimensionality and heterogeneity of the datasets, this strategy still produces challenges. To overcome these challenges, sparse-based statistical methods have been developed (Lin, Calhoun, & Wang, 2014b).

Sparse approaches aim to select *a priori* features that reduce the dimensionality of the data and facilitate interpretation of the results. We classified sparse strategies into three groups: sparse multivariate regression models (sMRM), sparse block methods (sBM), and sparse classification methods (sCM). Each of these strategies requires the modification of common methods and leads to a more accurate detection of the risk factors involved in neurodevelopmental domains.

Sparse multivariate regression

sMRM simultaneously searches multiple markers that are highly predictive of multiple imaging phenotypes. The most commonly used method is the sparse reduced-rank regression (sRRR), which is an efficient way a model with a multivariate responses and multiple features (Vounou et al., 2012; Vounou, Nichols, Montana, & Alzheimer's Disease Neuroimaging Initiative, 2010). Based on LASSO, Wang *et al.* (2012) proposed the group-sparse multi-task regression and feature-selection (G-SMuRFS) method to identify quantitative trait loci for multiple disease-relevant traits in an IG study of Alzheimer's disease. Importantly, G-SMuRFS accounts for linkage disequilibrium between SNPs. On the other hand, Wang *et al.* (2013) proposed a parallel version of random forest regression (PaRFR) to optimize computational cost. PaRFR ranks SNPs associated with the phenotype, and produces pair-wise measures of genetic proximity that can be directly compared to pair-wise measures of phenotypic proximity. Compared to other sparse methods, these methods have the advantage of being able to simultaneously perform dimension reduction and coefficient estimation, thereby improving predictive performance in the analysis (Kharratzadeh & Coates, 2016).

Sparse block methods

sBM are used to account for pleiotropic effects (i.e., when one genotype influences two or more seemingly unrelated brain structures or neurodevelopmental domains). In sBM, data are not differentiated in terms of regressors and responses, but rather each column of the data matrix is considered as an independent variable. The most commonly used methods are sparse Partial Least Squares (PLS; Wold, Martens, and Wold 1983; Krishnan et al. 2011), sparse Canonical Correlation Analysis (CCA; Correa et al. 2008; Sui et al. 2010), and parallel ICA (Meda et al., 2010; Pearlson, Liu, & Calhoun, 2015). PLS regression is used to model associations between two blocks of variables in which latent variables are searched such that their covariance is maximal. The CCA method, which maximizes the correlation between latent variables, is quite similar. Both methods are closely related and are often used in IG studies (Grellmann et al., 2015; Le Floch et al., 2012). Furthermore, some methodological extensions have been implemented, including incorporating correlation and group-like structure into the model (Du et al. 2014), and introducing group constraints in which the fMRI voxels are grouped by ROIs and the SNPs by genes (Lin, Calhoun, et al., 2014a). Moreover, to solve the time-consuming problems associated with high-volume datasets in real

applications, pICA was proposed as an optimized extension of ICA. This optimized method uses multiple processors in parallel and is based on logarithmic-model performance to speed up the procedure. pICA methods have had a great impact on IG studies. For instance, Chen *et al.*, (2012) employed a pICA method to jointly detect significant correlations between fMRI components and SNP components. Meda *et al.* (2012b) also applied pICA in their study, and found novel IG relationships for Alzheimer's disease. Extensions of pICA have also been proposed. For example, Chen *et al.* (2013) developed an upgrade for pICA in which knowledge is incorporated *a priori* to help focus the analysis on particular components of interest. Subsequently, Vergara *et al.* (2014) proposed a three-way pICA approach to analyze links between genetics, brain structure and brain function. This method is novel in the sense that it incorporates three different sources of information into one comprehensive analysis (SNP, MRI and fMRI). The main advantage of these methods is the possibility to fully integrate multiple sources of datasets, and include a preliminary step of dimensionality reduction.

Sparse classification methods

sCMs have also been applied in IG to achieve better discrimination of disease status based on multiple imaging and multiple genetic features. These methods include machine learning techniques, which can be also used to define gene-sets or pathways that best predict the imaging phenotype (Cao, Duan, Lin, Calhoun, & Wang, 2013), Kernel machine-based methods (Ge *et al.*, 2015; Ge, Feng, Hibar, Thompson, & Nichols, 2012; Zhang, Huang, Shen, & Alzheimer's Disease Neuroimaging Initiative, 2014) and Bayesian methods (Batmanghelich, Dalca, Sabuncu, Polina, & ADNI, 2013; Stingo, Guindani, Vannucci, & Calhoun, 2013; Zhe, Xu, Qi, Yu, & ADNI, 2014).

In general, sCMs apply a sparse representation coefficient during classification, which contains very important discriminating information. Therefore, sCM have more power to discriminate than other previously described methods (Li & Ngom, 2013; Wen, Jia, Lian, Zhou, & Lu, 2016).

3.3 Statistical methods used in analyzing different data modalities

Extensions to multivariate methodologies increase the analytical possibilities for IG studies. However, we showed that most (62.5%) studies that combine genetic and imaging data are based on sBM methods, data-driven analyses and sMRM. In terms of sBM, the most widely used methods are sCCA (20%) and pICA (15%). Among the data-driven methods, the most frequently used are LASSO (7.5%) and PCA (6%) [Figure 2].

3.3.1. Statistical methods by neuroimaging technique

Most IG studies discussed in this review imaged the brain using either structural MRI (sMRI) or fMRI, while a few studies were also based on diffusion tensor imaging (DTI) or other methods [Figure 3a]. Stratifying the statistical methods according to neuroimaging technique (sMRI or fMRI), we found that sBS methods are still the most widely used method for analyzing both sMRI (55.6%) and fMRI (30.8%) data. While both neuroimaging methods techniques coincide in a nearly similar use of sCM, data-driven, and GSA, they differ with respect to the use of sMRM, which seems to be specific for analyzing fMRI data [Figure 3b].

3.3.2. Statistical methods by imaging marker/disease

Regarding outcomes, IG studies can be divided into two types: studies that have (1) disease or functional capacity outcomes (30.6%), where neuroimaging data is used in combination with genetic data to predict the disease; and (2) neuroimaging feature outcomes (i.e., image-based features) such as, diffusivity values and/or fractional anisotropy (48.9%) [Figure 4a]. Stratifying the classification of statistical methods by taking into account neuroimaging intermediate phenotypes as principal phenotypes of the study, we observed that sBM (42.2%) and sMRM (21.1%) are still the most popular methods. The most commonly analyzed diseases were Alzheimer's disease (16.3%) and Schizophrenia (8.2%). While for Schizophrenia, sBM (66.7%) and sCM (33.3%) were the most popular methods, Alzheimer's disease studies displayed much more heterogeneity in terms of the specific methodology used [Figure 4b].

3.4. Multiple comparison correction strategies

The dimensionality of the data in IG studies is the principal handicap in acquiring analytical perspectives, exacerbating existing issues regarding the reliability and interpretability of the results. Moreover, independently testing millions of SNPs (population-wide) for associations on the one hand, and millions of neuroimaging-based measures on the other, reduces statistical power due to multiple testing correction. Thus, large-scale analysis of genetic versus brain data requires new strategies for controlling Type-I error (incorrect rejection of a true null hypothesis, also known as a false positive). While most types of research accept a confidence rate of 95%, the hundreds of thousands of statistical tests composing an IG analysis increases the likelihood of a Type-I error. Hence, a confidence rate of 95% seems insufficient to properly account for multiple testing in IG studies. The most widely known and used method designed to counteract this problem, especially in genetics, is the Bonferroni correction, which defines a new confidence level by taking into account the number of tests performed in the analysis. However, this method has also received criticism for being very conservative (Perneger, 1998), particularly in situations where there are a large number of tests and/or the test statistics are correlated (e.g., in linkage disequilibrium (LD) in the genetics context). In such a situation, Bonferroni correction comes at the cost of increasing the likelihood of producing false negatives, and in turn, decreases the power of the test (i.e., the probability of correctly rejecting a false null hypothesis). For this reason, significance thresholds have been established in the genetics context using simulation studies that account for LD between SNPs. In this case, genome-wide significance was set to $\alpha=5 \times 10^{-8}$ and suggestive evidence of association was set to $\alpha=10^{-5}$ (Duggal, Gillanders, Holmes, & Bailey-Wilson, 2008; Pe'er, Yelensky, Altshuler, & Daly, 2008). The challenge of choosing an appropriate threshold becomes even more difficult in the neuroimaging context. In genetics, the Bonferroni correction is an extended option available for correcting for multiple testing. However, because of the significant spatial correlation of neuroimaging data, such as in the simultaneous study of grey matter volume and subcortical structures, or cortical thickness and cortical surface areas, this method is not optimal in this context either. Alternatives include the use of Gaussian random field theory (Brett, Penny, & Kiebel, 2003), and/or non-parametric permutation correction techniques (Nichols & Holmes, 2002). Other methods such as the false discovery rate (FDR), which controls the proportion of false positives among all rejected tests (Benjamini & Hochberg, 1995) are often used in the neuroimaging

context. FDR can be less stringent and lead to a gain in power. In addition, since FDR work only on the p-value, it can be applied to any valid statistical test. However, FDR cannot be applied in the genomic context because it assumes that many features do not follow the null hypothesis, and this is not the case for genetic data where only a few SNPs are expected to be significantly associated with the outcome of interest (Dudbridge, Gusnanto, & Koeleman, 2006). FDR methods together with the use of arbitrary uncorrected thresholds (e.g., $p < 0.001$) are an extended and generalized practice to correct multiple comparisons problem in the neuroimaging field. The main disadvantage is that null findings are hard to disseminate, hence it is difficult to refute false positives established in the literature (Christley, 2010; Forstmeier, Wagenmakers, & Parker, 2017).

3.5. *R packages*

Table 2 shows R and Bioconductor packages designed to perform multivariate analyses of one, two, or more datasets. This information is not exhaustive, but aims to provide a list of packages that can be used to perform the data analyses described in this review.

We are aware that some IG researchers analyze their data using Matlab software. However, we prefer to provide a list of R packages due to the following advantages: (1) the functions and packages are freely available through an open source project such as R and Bioconductor; (2) R not only contains functions and packages that deal with the integrative analyses described here, but also includes a lot of other packages capable of performing different types of neuroimaging data analyses (R CRAN Task; <https://cran.r-project.org/web/views/MedicalImaging.html>); (3) R functions can be combined with Bioconductor packages that are designed to deal with genomic data; and (4) there is a big initiative called Neuroconductor (<https://neuroconductor.org/>) that is implementing methods and tools to analyze imaging data. Neuroconductor started with 58 interoperable packages that cover multiple areas of imaging including visualization, data processing and storage, and statistical inference. It enables the IG community to submit new packages addressing novel issues and methods. Finally, another advantage of working under the R environment is that current techniques are continuously evolving and most of the R packages that deal with new data problems are available at GitHub repositories where developmental version can be easily installed into R.

3.5.1 Analysis of one table

3.5.1.1 Dimensionality reduction methods

A wide variety of methods and packages are available to reduce the dimensionality of a table. Basically, all of these methods end up with a similar output: a set of variables (called principal components; PCs) that are a combination of original variables. The main difference between these methods is how PCA assumptions are held. PCA assumes that all variables are continuous (multidimensional normality), that PCs are linear combinations of the original variables, and that the PCs are independent (e.g., orthogonal). All of these assumptions can be relaxed by using other approaches. For instance, non-linear PCA (the *homals* function in the *homals* package; Leeuw & Mair, 2009) can be used when the assumption of linearity does not hold. Additionally, this package can also deal with data that is composed of different types of variables (categorical and numerical). The normality assumption, on the other hand, can be overcome by using independent component analysis (ICA) implemented in the *fastICA* package (Hyvärinen, 2013). All of these methods can be used to visualize variables and samples in two or three dimensions (i.e., PCs). While these methods can be used in a preliminary data analysis to discover patterns and possible associations, they are by no means a formal statistical test to decide which set of variables is relevant.

3.5.1.2 Variable selection methods

The problem of variable selection can be addressed by using LASSO or elastic net methods, implemented in the *glmnet* package for example (Friedman, Hastie, & Tibshirani, 2010; Simon, Friedman, Hastie, & Tibshirani, 2011; Tibshirani et al., 2012). Furthermore, a newer version of this technique capable of dealing with large datasets is implemented in the *biglasso* package (Zeng & Breheny, 2017).

3.5.2 Analysis of multiple tables

There is also a wide variety of packages that can perform multivariate data analysis by integrating two or more tables. CCA is the counterpart of PCA when analyzing two tables (*cc* function of the *cca* package; Legendre, Legendre, Legendre, & Legendre, 2012). The assumptions made when using this methodology are basically the same as those of PCA. However, in CCA the PCs also maximize the correlation between the two tables since it assumes normality. This assumption can be overcome by using coinertia

analysis (*cia* function of the *made4* package; Culhane, Thioulouse, Perriere, & Higgins, 2005). CIA can be seen as the non-parametric version of CCA, and estimates PCs by maximizing the covariance between tables. It is robust against outliers and can address the ($n \ll p$) problem without any extra task. In contrast, IG studies using CCA to analyze a large amount of SNP data require sparse or penalized versions of CCA. This problem is very well addressed in the *CCA* function of the *PMA* package (Witten, Tibshirani, & Hastie, 2009). This package also contains the *multiCCA* function, which can analyze more than two tables using a canonical correlation approach. Based on our experience, we highly recommend its use because it is quite fast even when analyzing thousands of features. The only limitation of this package is visualization. Nonetheless, we have developed some functions that create appropriate plots to help interpret *multiCCA* results. These functions are available on request.

CCA and CIA assume that both tables are interchangeable. That is, we are interested in the correlation between tables, or how they are associated. In some data analyses, one of the tables can be considered as the outcome. For instance, in IG one may argue that SNPs can explain observed differences in volumes. In this case, the relationship between the two tables cannot be considered symmetrical. This problem can be addressed by using redundancy analysis (RDA). RDA considers one of the tables as dependent (volumes) and the other one as independent (SNPs). The analysis of multiple tables using the ICA approach is implemented in the *pICA* function from the *PGICA* package (Eloyan, Crainiceanu, & Caffo, 2013).

3.5.2.1 Regularized methods

In some IG studies, it can be interesting to analyze the complete genome or a large number of features. In such a situation, multivariate methods must be extended to account for having more variables (p) than individuals (n) ($n \ll p$). To this end, several R packages implement a regularized or sparse version of standard multivariate methods. These methods require the estimation of a penalized parameter that is normally estimated using a cross-validation procedure. Sometimes, this can be very demanding in terms of computational time. However, these processes can be sped up by using parallel implementations of these algorithms.

We would like to highlight that the regularized generalized canonical correlation (RGCCA) method is a generalization of almost all of the multivariate methods described here (M. Tenenhaus, Tenenhaus, & Groenen, 2017). The *rgcca* package

illustrates how to perform, among others, PCA, CCA, RDA, GCCA and MCIA. Moreover, the *sgcca* function within this package makes it possible to combine these methods using a variable selection procedure (A. Tenenhaus et al., 2014).

4. Discussion

In this review, we summarize the current literature on statistical methods for Imaging Genetics. These studies demonstrate that the use of accurate statistical models for the joint analysis of imaging and genetics data facilitates our understanding of complex neurodevelopmental domains.

In contrast to single association analyses, where the p-values for association and effect size are computed for each variant independently, multivariate techniques aim to decompose the data into its most important sources of variation. Dimensionality reduction techniques make it possible to determine the degree to which the whole set of genetic and/or neuroimaging markers contributes to the variability of the symptomatology in a joint, rather than individual manner. Multivariate data reduction techniques select the best combinations of variables, offering a new reference basis for observing the dataset. IG components can therefore be associated with biological mechanisms (e.g., the identification of new target genes. Then, on one hand, by obtaining a set of components, is it possible to rank the importance of the original variables. This can be done by taking into account, for instance, their correlation/association with the obtained components. On the other hand, after select significant results, then univariate models can be made to quantify the effect and / or to validate it in other data sets. This procedure allows to perform a discovery step without relying on a univariate approach, and in turn, obtain a meaningful way to quantify the effect size.

The vast majority of the IG studies reviewed here used sparse representation methods. Sparse models overcome the difficulty of modeling IG data, which is usually characterized by a smaller sample size than the number of features by reducing the dimensionality of the data for easier interpretation of the results. Sparse models also provide a powerful approach for considering correlation structure, something which is absent in many other strategies. In addition, sparse models provide a flexible way to incorporate *a priori* biological knowledge into the model, and also facilitate the

discovery complex relationships between imaging and genetic features (Lin, Cao, Calhoun, & Wang, 2014).

There is no doubt that multivariate methods are highly suited for providing insight into the links between genes and neurodevelopmental domains via intermediate imaging markers (Liu & Calhoun, 2014). However, even though there is a clear need to implement multivariate methods in IG studies, these novel techniques have had limited impact in terms of real scientific discoveries. This could be because (1) traditional methods that perform well for moderate sample sizes do not scale up well to massive data, and (2) statistical methods that perform well for low-dimensional data face significant challenges in analyzing high-dimensional data. Multivariate association analysis methods are not easily applied, and despite the efforts thus far, novel methods still need to be developed and standardized to deal with the superdimensionality and interpretability of the data. To limit the analytical burden on human and computational resources, this would ideally make use of already available results from collaborative efforts.

On the other hand, brain imaging is increasingly recognized as providing intermediate phenotypes to understand the complex pathway between genetics and behavioral or clinical phenotypes (Lenzenweger, 2013; Meyer-Lindenberg & Weinberger, 2006). The intermediate phenotype strategy has certain advantages over that based on disease status, both in improving the power for discovering genetic risk variants, and in understanding underlying mechanisms of neurodevelopmental domains. For instance, the association between a particular behavioral feature and thickness in a specific cortical area may be a clue that the mechanisms involved in the development of that cortical region are also relevant to that behavior (Meyer-Lindenberg & Weinberger, 2006). Moreover, intermediate phenotypes are thought to be "simpler" at a genetic level. Genetic effects on these measures would be stronger and independent of clinical status, thereby making their detection possible within samples of the general population. From this, genetic associations could be captured more objectively and accurately (Braff & Tamminga, 2017; Geschwind & Flint, 2015; Gottesman & Gould, 2003; Iacono, Vaidyanathan, Vrieze, & Malone, 2014; Meyer-Lindenberg & Weinberger, 2006; Munafò & Flint, 2014).

Regarding this strategy, IG studies have started to analyze associations between candidate SNPs and/or candidate genes and candidate neuroimaging intermediate features (e.g., fMRI (Bookheimer et al., 2000), brain structure (Winkler et al., 2010),

functional connectivity (Dennis & Thompson, 2014). Specifically, in our review we found that more than half of the studies take advantage of intermediate imaging markers to elucidate relationships between genetic variants and neurodevelopmental outcomes. Despite this fact, there are still many challenges to overcome in modeling IG data (Bigos & Weinberger, 2010; Bogdan et al., 2017). For instance, while intermediate imaging markers have greater power to detect associations between genetic markers and neurodevelopmental and behavioral outcomes, larger datasets are needed to obtain sufficient power to detect genome-wide associations, and to determine how much of the phenotypic variance can be accounted for by those genetic factors. Moreover, the availability of multiscale and multimodal IG data creates computational challenges for developing powerful, efficient and biologically interpretable methods. Together with the low effect sizes of individual genetic variants and the heterogeneity of most neurodevelopmental domains, this challenges the discovery of significant effects (Franke et al., 2016; Murphy et al., 2013).

Importantly, as the associations described in this review are focused on European populations (as this was what we found in the literature), interpretation of the results cannot be universally applied to populations of different genetic backgrounds (Popejoy & Fullerton, 2016). In addition, replication of IG results is confronted with distinct replicability issues to those of genetic and neuroimaging studies, as well as those of complex analysis strategies. While replication in independent genetic studies is often required to validate a finding (Ioannidis, Munafò, Fusar-Poli, Nosek, & David, 2014; Schnack & Kahn, 2016), neuroimaging results are challenged by the lack of power for a reliable finding (Nord, Valton, Wood, & Roiser, 2017; Poldrack et al., 2017). In addition, complex analyses should be tested for robustness and reproducibility (Peng, 2011). A challenging question is then posed: How and to what extent should an IG finding be considered valid? While fully replicating an experiment, or performing a power analysis on exploratory studies may not be an option, general guidelines for exploratory analyses can be followed. Analytical flexibility can be met with an internal replication or robustness assessment using leave-one-out validation methods (Fletcher & Grafton, 2013). Reproducibility can be assessed by applying the same methods to a random subsample in which the exploratory analysis was not conducted (Picciotto, 2018). In addition, different methods and analytical strategies should be compared within studies. Given that IG is a young field, standards for sound research will begin to emerge as more studies are conducted.

Even with the disadvantages and problems that can be produced in statistical inference and interpretation, we highlight that recent advances in multivariate analysis, large data set analysis, and big data science have great potential to achieve new analytical perspectives. This will ultimately advance our understanding of the underlying biology of many neurodevelopmental domains. In this sense, collaborative efforts between many studies increases the statistical power needed to find and replicate those genetic variants related to changes in brain structure and function. Working with large datasets also has an impact on the tools used in the data analyses. This issue has been considered in the R framework by creating specific infrastructures to deal with big data problems. As an example, the *bigmemory* package is designed to manage large matrices with shared memory and memory-mapped files. Some packages implement multivariate methods using the big matrices base of the *bigmemory* infrastructure (Kane, 2015). For instance, the *bigpca* package is designed to perform PCA on very large tables, and uses objects of class *big.matrix* as input (Cooper, 2015). These methods not only dramatically reduce the computational time required to analyze complete GWAS data, but also allow the user to properly manage such amounts of information in R. As a result, one may use the *prcomp* or *princomp* functions of the *stats* package to compute a PCA with a small to moderate number of features and samples. However, performing PCA with thousands of variants in a large number of individuals with these same functions is not recommended. Instead, the *big.PCA* function implemented in the *bigpca* package can be used to solve memory and computing issues. Hence, the reinforcement and development of new, specific statistical methods and computational procedures will become crucial as more applications for IG arise in the future.

CONFLICT OF INTEREST

The authors declare no conflicts of interest

ACKNOWLEDGEMENTS

Natalia Vilor-Tejedor was funded by a pre-doctoral grant from the Agència de Gestió d'Ajuts Universitaris i de Recerca (2017 FI_B 00636), Generalitat de Catalunya – Fons Social Europeu. This work has been partially supported by a STSM Grant from EU COST Action 15120 Open Multiscale Systems Medicine (OpenMultiMed) and Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP). Silvia Alemany thanks the Institute of Health Carlos III for her Sara Borrell postdoctoral grant (CD14/00214). The research leading to these results has also received funding from the European Research Council under the ERC Grant Agreement number 268479 – the BREATHE project and by grant MTM2015-68140-R from the Ministerio de Economía e Innovación (Spain). ISGlobal is a member of the CERCA Programme, Generalitat de Catalunya.

REFERENCES

- Abdi, H., Williams, L. J., & Valentin, D. (2013). Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(2), 149–179. <https://doi.org/10.1002/wics.1246>
- Adams, H. H. H., Hibar, D. P., Chouraki, V., Stein, J. L., Nyquist, P. A., Rentería, M. E., ... Thompson, P. M. (2016). Novel genetic loci underlying human intracranial volume identified through genome-wide association. *Nature Neuroscience*, 19(12), 1569–1582. <https://doi.org/10.1038/nn.4398>
- Batmanghelich, N. K., Dalca, A. V., Sabuncu, M. R., Polina, G., & ADNI. (2013). Joint modeling of imaging and genetics. *Information Processing in Medical Imaging : Proceedings of the ... Conference*, 23, 766–77. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/24684016>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*. WileyRoyal Statistical Society. <https://doi.org/10.2307/2346101>
- Bigos, K. L., & Hariri, A. R. (2007). Neuroimaging: technologies at the interface of genes, brain, and behavior. *Neuroimaging Clinics of North America*, 17(4), 459–67, viii. <https://doi.org/10.1016/j.nic.2007.09.005>
- Bigos, K. L., & Weinberger, D. R. (2010). Imaging genetics—days of future past. *NeuroImage*, 53(3), 804–809. <https://doi.org/10.1016/j.neuroimage.2010.01.035>
- Blokland, G. A. M., McMahon, K. L., Thompson, P. M., Martin, N. G., de Zubicaray, G. I., & Wright, M. J. (2011). Heritability of working memory brain activation. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, 31(30), 10882–90. <https://doi.org/10.1523/JNEUROSCI.5334-10.2011>
- Bloom, J. S., Kotenko, I., Sadhu, M. J., Treusch, S., Albert, F. W., & Kruglyak, L. (2015). Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. *Nature Communications*, 6(1), 8712. <https://doi.org/10.1038/ncomms9712>
- Bogdan, R., Salmeron, B. J., Carey, C. E., Agrawal, A., Calhoun, V. D., Garavan, H., ... Goldman, D. (2017). Imaging Genetics and Genomics in Psychiatry: A Critical Review of Progress and Potential. *Biological Psychiatry*, 82(3), 165–175. <https://doi.org/10.1016/j.biopsych.2016.12.030>
- Bookheimer, S. Y., Strojwas, M. H., Cohen, M. S., Saunders, A. M., Pericak-Vance, M. A., Mazziotta, J. C., & Small, G. W. (2000). Patterns of Brain Activation in People at Risk for Alzheimer's Disease. *New England Journal of Medicine*, 343(7), 450–456. <https://doi.org/10.1056/NEJM200008173430701>
- Braff, D. L., & Tamminga, C. A. (2017). Endophenotypes, Epigenetics, Polygenicity

- and More: Irv Gottesman's Dynamic Legacy. *Schizophrenia Bulletin*, 43(1), 10–16. <https://doi.org/10.1093/schbul/sbw157>
- Brett, M., Penny, W., & Kiebel, S. (2003). An Introduction to Random Field Theory. Retrieved from <http://www.fil.ion.ucl.ac.uk/spm/doc/books/hbf2/pdfs/Ch14.pdf>
- Cao, H., Duan, J., Lin, D., Calhoun, V., & Wang, Y.-P. (2013). Integrating fMRI and SNP data for biomarker identification for schizophrenia with a sparse representation based variable selection method. *BMC Medical Genomics*, 6(Suppl 3), S2. <https://doi.org/10.1186/1755-8794-6-S3-S2>
- Chen, J., Calhoun, V. D., Pearlson, G. D., Ehrlich, S., Turner, J. A., Ho, B.-C., ... Liu, J. (2012). Multifaceted genomic risk for brain function in schizophrenia. *NeuroImage*, 61(4), 866–75. <https://doi.org/10.1016/j.neuroimage.2012.03.022>
- Chen, J., Calhoun, V. D., Pearlson, G. D., Perrone-Bizzozero, N., Sui, J., Turner, J. A., ... Liu, J. (2013). Guided exploration of genomic risk for gray matter abnormalities in schizophrenia using parallel independent component analysis with reference. *NeuroImage*, 83, 384–96. <https://doi.org/10.1016/j.neuroimage.2013.05.073>
- Christley, R. M. (2010). Power and Error: Increased Risk of False Positive Results in Underpowered Studies. *The Open Epidemiology Journal*, 3, 16–19. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.824.2966&rep=rep1&type=pdf>
- Cooper, N. (2015). bigpca: PCA, Transpose and Multicore Functionality for “big.matrix” Objects version 1.0.3 from CRAN. Retrieved November 17, 2017, from <https://rdrr.io/cran/bigpca/>
- Correa, N. M., Li, Y.-O., Adalı, T., & Calhoun, V. D. (2008). Canonical Correlation Analysis for Feature-Based Fusion of Biomedical Imaging Modalities and Its Application to Detection of Associative Networks in Schizophrenia. *IEEE Journal of Selected Topics in Signal Processing*, 2(6), 998–1007. <https://doi.org/10.1109/JSTSP.2008.2008265>
- Culhane, A. C., Thioulouse, J., Perriere, G., & Higgins, D. G. (2005). MADE4: an R package for multivariate analysis of gene expression data. *Bioinformatics*, 21(11), 2789–2790. <https://doi.org/10.1093/bioinformatics/bti394>
- de Leeuw, C. A., Mooij, J. M., Heskes, T., & Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. *PLoS Computational Biology*, 11(4), e1004219. <https://doi.org/10.1371/journal.pcbi.1004219>
- Dennis, E. L., & Thompson, P. M. (2014). Reprint of: Mapping connectivity in the developing brain. *International Journal of Developmental Neuroscience: The Official Journal of the International Society for Developmental Neuroscience*, 32, 41–57. <https://doi.org/10.1016/j.ijdevneu.2013.11.005>
- Dima, D., & Breen, G. (2015). Polygenic risk scores in imaging genetics: Usefulness and applications. *Journal of Psychopharmacology*, 29(8), 867–871. <https://doi.org/10.1177/0269881115584470>
- Du, X., An, Y., Yu, L., Liu, R., Qin, Y., Guo, X., ... Wang, Y. (2014). A genomic copy number variant analysis implicates the MBD5 and HNRNPU genes in Chinese

- children with infantile spasms and expands the clinical spectrum of 2q23.1 deletion. *BMC Medical Genetics*, 15, 62. <https://doi.org/10.1186/1471-2350-15-62>
- Dudbridge, F. (2013). Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics*, 9(3), e1003348. <https://doi.org/10.1371/journal.pgen.1003348>
- Dudbridge, F., Gusnanto, A., & Koeleman, B. P. C. (2006). Detecting multiple associations in genome-wide studies. *Human Genomics*, 2(5), 310–7. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/16595075>
- Duggal, P., Gillanders, E. M., Holmes, T. N., & Bailey-Wilson, J. E. (2008). Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC Genomics*, 9, 516. <https://doi.org/10.1186/1471-2164-9-516>
- Ehrenreich, I. M. (2017). Epistasis: Searching for Interacting Genetic Variants Using Crosses. *G3 (Bethesda, Md.)*, 7(6), 1619–1622. <https://doi.org/10.1534/g3.117.042770>
- Eloyan, A., Crainiceanu, C. M., & Caffo, B. S. (2013). Likelihood-based population independent component analysis. *Biostatistics (Oxford, England)*, 14(3), 514–27. <https://doi.org/10.1093/biostatistics/kxs055>
- Eslami, A., Qannari, E. M., Kohler, A., & Bougeard, S. (2014). Multivariate analysis of multiblock and multigroup data. *Chemometrics and Intelligent Laboratory Systems*, 133, 63–69. <https://doi.org/10.1016/J.CHEMOLAB.2014.01.016>
- Filipovych, R., Gaonkar, B., & Davatzikos, C. (2012). A Composite Multivariate Polygenic and Neuroimaging Score for Prediction of Conversion to Alzheimer's Disease. In *2012 Second International Workshop on Pattern Recognition in NeuroImaging* (pp. 105–108). IEEE. <https://doi.org/10.1109/PRNI.2012.9>
- Fletcher, P. C., & Grafton, S. T. (2013). Repeat after me: Replication in clinical neuroimaging is critical. *NeuroImage: Clinical*, 2, 247–248. <https://doi.org/10.1016/j.nicl.2013.01.007>
- Forstmeier, W., Wagenmakers, E.-J., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings - a practical guide. *Biological Reviews*, 92(4), 1941–1968. <https://doi.org/10.1111/brv.12315>
- Franke, B., Stein, J. L., Ripke, S., Anttila, V., Hibar, D. P., van Hulzen, K. J. E., ... Sullivan, P. F. (2016). Genetic influences on schizophrenia and subcortical brain volumes: large-scale proof of concept. *Nature Neuroscience*, 19(3), 420–31. <https://doi.org/10.1038/nn.4228>
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1–22. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/20808728>
- Gallinat, J., Bauer, M., & Heinz, A. (2008). Genes and Neuroimaging: Advances in Psychiatric Research. *Neurodegenerative Diseases*, 5(5), 277–285. <https://doi.org/10.1159/000135612>
- Ge, T., Feng, J., Hibar, D. P., Thompson, P. M., & 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(2), 858–873. <https://doi.org/10.1016/j.neuroimage.2012.07.012>
- Ge, T., Nichols, T. E., Ghosh, D., Mormino, E. C., Smoller, J. W., Sabuncu, M. R., & Alzheimer's Disease Neuroimaging Initiative. (2015). A kernel machine method for detecting effects of interaction between multidimensional variable sets: An imaging genetics application. *NeuroImage*, 109, 505–514. <https://doi.org/10.1016/j.neuroimage.2015.01.029>
- Geschwind, D. H., & Flint, J. (2015). Genetics and genomics of psychiatric disease. *Science (New York, N.Y.)*, 349(6255), 1489–94. <https://doi.org/10.1126/science.aaa8954>
- Gottesman, I. I., & Gould, T. D. (2003). The Endophenotype Concept in Psychiatry: Etymology and Strategic Intentions. *American Journal of Psychiatry*, 160(4), 636–645. <https://doi.org/10.1176/appi.ajp.160.4.636>
- Grellmann, C., Bitzer, S., Neumann, J., Westlye, L. T., Andreassen, O. A., Villringer, A., & Horstmann, A. (2015). Comparison of variants of canonical correlation analysis and partial least squares for combined analysis of MRI and genetic data. *NeuroImage*, 107, 289–310. <https://doi.org/10.1016/j.neuroimage.2014.12.025>
- Hibar, D. P., Adams, H. H. H., Jahanshad, N., Chauhan, G., Stein, J. L., Hofer, E., ... Ikram, M. A. (2017). Novel genetic loci associated with hippocampal volume. *Nature Communications*, 8, 13624. <https://doi.org/10.1038/ncomms13624>
- Hibar, D. P., Stein, J. L., Kohannim, O., Jahanshad, N., Saykin, A. J., Shen, L., ... Initiative, the A. D. N. (2011). Voxelwise gene-wide association study (vGeneWAS): multivariate gene-based association testing in 731 elderly subjects. *NeuroImage*, 56(4), 1875–91. <https://doi.org/10.1016/j.neuroimage.2011.03.077>
- Hibar, D. P., Stein, J. L., Renteria, M. E., Arias-Vasquez, A., Desrivieres, S., Jahanshad, N., ... Medland, S. E. (2015). Common genetic variants influence human subcortical brain structures. *Nature*, 520(7546), 224–9. <https://doi.org/10.1038/nature14101>
- Hibar, D. P., Stein, J. L., Ryles, A. B., Kohannim, O., Jahanshad, N., Medland, S. E., ... Alzheimer's Disease Neuroimaging Initiative. (2013). Genome-wide association identifies genetic variants associated with lentiform nucleus volume in N = 1345 young and elderly subjects. *Brain Imaging and Behavior*, 7(2), 102–15. <https://doi.org/10.1007/s11682-012-9199-7>
- Hoogman, M., Guadalupe, T., Zwiers, M. P., Klarenbeek, P., Francks, C., & Fisher, S. E. (2014). Assessing the effects of common variation in the FOXP2 gene on human brain structure. *Frontiers in Human Neuroscience*, 8, 473. <https://doi.org/10.3389/fnhum.2014.00473>
- Hyvärinen, A. (2013). Independent component analysis: recent advances. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, 371(1984), 20110534. <https://doi.org/10.1098/rsta.2011.0534>
- Iacono, W. G., Vaidyanathan, U., Vrieze, S. I., & Malone, S. M. (2014). Knowns and

- unknowns for psychophysiological endophenotypes: Integration and response to commentaries. *Psychophysiology*, 51(12), 1339–1347.
<https://doi.org/10.1111/psyp.12358>
- Inkster, B., Nichols, T. E., Saemann, P. G., Auer, D. P., Holsboer, F., Muglia, P., & Matthews, P. M. (2010). Pathway-based approaches to imaging genetics association studies: Wnt signaling, GSK3beta substrates and major depression. *NeuroImage*, 53(3), 908–917. <https://doi.org/10.1016/j.neuroimage.2010.02.065>
- Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in Cognitive Sciences*, 18(5), 235–41.
<https://doi.org/10.1016/j.tics.2014.02.010>
- Kane, M. J. (2015). The Bigmemory Project. Retrieved from
<http://www.bigmemory.org/>
- Khadka, S., Pearlson, G. D., Calhoun, V. D., Liu, J., Gelernter, J., Bessette, K. L., & Stevens, M. C. (2016). Multivariate Imaging Genetics Study of MRI Gray Matter Volume and SNPs Reveals Biological Pathways Correlated with Brain Structural Differences in Attention Deficit Hyperactivity Disorder. *Frontiers in Psychiatry*, 7.
<https://doi.org/10.3389/fpsyt.2016.00128>
- Kharratzadeh, M., & Coates, M. (2016). Sparse multivariate factor regression. In 2016 *IEEE Statistical Signal Processing Workshop (SSP)* (pp. 1–5). IEEE.
<https://doi.org/10.1109/SSP.2016.7551732>
- Kohannim, O., Hibar, D. P., Stein, J. L., Jahanshad, N., Hua, X., Rajagopalan, P., ... Alzheimer's Disease Neuroimaging Initiative. (2012). Discovery and Replication of Gene Influences on Brain Structure Using LASSO Regression. *Frontiers in Neuroscience*, 6, 115. <https://doi.org/10.3389/fnins.2012.00115>
- Kremen, W. S., & Jacobson, K. C. (2010). Introduction to the Special Issue, Pathways Between Genes, Brain, and Behavior. *Behavior Genetics*, 40(2), 111–113.
<https://doi.org/10.1007/s10519-010-9342-4>
- Krishnan, A., Williams, L. J., McIntosh, A. R., & Abdi, H. (2011). Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review. *NeuroImage*, 56(2), 455–475. <https://doi.org/10.1016/j.neuroimage.2010.07.034>
- Le Floch, E., Guillemot, V., Frouin, V., Pinel, P., Lalanne, C., Trinchera, L., ... Duchesnay, E. (2012). Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares. *NeuroImage*, 63(1), 11–24.
<https://doi.org/10.1016/j.neuroimage.2012.06.061>
- Leeuw, J. de, & Mair, P. (2009). Gifi Methods for Optimal Scaling in R: The Package **homals**. *Journal of Statistical Software*, 31(4), 1–21.
<https://doi.org/10.18637/jss.v031.i04>
- Legendre, P., Legendre, L., Legendre, L., & Legendre, P. (2012). *Numerical ecology*. Elsevier.
- Lenzenweger, M. F. (2013). Thinking clearly about the endophenotype–intermediate

- phenotype–biomarker distinctions in developmental psychopathology research. *Development and Psychopathology*, 25(4pt2), 1347–1357.
<https://doi.org/10.1017/S0954579413000655>
- Li, Y., & Ngom, A. (2013). Sparse representation approaches for the classification of high-dimensional biological data. *BMC Systems Biology*, 7 Suppl 4(Suppl 4), S6.
<https://doi.org/10.1186/1752-0509-7-S4-S6>
- Lin, D., Calhoun, V. D., & Wang, Y.-P. (2014a). Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. *Medical Image Analysis*, 18(6), 891–902. <https://doi.org/10.1016/j.media.2013.10.010>
- Lin, D., Calhoun, V. D., & Wang, Y.-P. (2014b). Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. *Medical Image Analysis*, 18(6), 891–902. <https://doi.org/10.1016/j.media.2013.10.010>
- Lin, D., Cao, H., Calhoun, V. D., & Wang, Y.-P. (2014). Sparse models for correlative and integrative analysis of imaging and genetic data. *Journal of Neuroscience Methods*, 237, 69–78. <https://doi.org/10.1016/j.jneumeth.2014.09.001>
- Liu, J., & Calhoun, V. D. (2014). A review of multivariate analyses in imaging genetics. *Frontiers in Neuroinformatics*, 8, 29. <https://doi.org/10.3389/fninf.2014.00029>
- Ma, S., & Dai, Y. (2011). Principal component analysis based methods in bioinformatics studies. *Briefings in Bioinformatics*, 12(6), 714–22.
<https://doi.org/10.1093/bib/bbq090>
- Mattingsdal, M., Brown, A. A., Djurovic, S., Snderby, I. E., Server, A., Melle, I., ... Andreassen, O. A. (2013). Pathway analysis of genetic markers associated with a functional MRI faces paradigm implicates polymorphisms in calcium responsive pathways. *NeuroImage*, 70, 143–149.
<https://doi.org/10.1016/j.neuroimage.2012.12.035>
- McIntosh, A. R., & Mii, B. (2013). Multivariate Statistical Analyses for Neuroimaging Data. *Annual Review of Psychology*, 64(1), 499–525.
<https://doi.org/10.1146/annurev-psych-113011-143804>
- Meda, S. A., Jagannathan, K., Gelernter, J., Calhoun, V. D., Liu, J., Stevens, M. C., & Pearlson, G. D. (2010). A pilot multivariate parallel ICA study to investigate differential linkage between neural networks and genetic profiles in schizophrenia. *NeuroImage*, 53(3), 1007–15. <https://doi.org/10.1016/j.neuroimage.2009.11.052>
- Meda, S. A., Narayanan, B., Liu, J., Perrone-Bizzozero, N. I., Stevens, M. C., Calhoun, V. D., ... Pearlson, G. D. (2012a). A large scale multivariate parallel ICA method reveals novel imaging-genetic relationships for Alzheimer’s disease in the ADNI cohort. *NeuroImage*, 60(3), 1608–21.
<https://doi.org/10.1016/j.neuroimage.2011.12.076>
- Meda, S. A., Narayanan, B., Liu, J., Perrone-Bizzozero, N. I., Stevens, M. C., Calhoun, V. D., ... Pearlson, G. D. (2012b). A large scale multivariate parallel ICA method reveals novel imaging-genetic relationships for Alzheimer’s disease in the ADNI cohort. *NeuroImage*, 60(3), 1608–21.
<https://doi.org/10.1016/j.neuroimage.2011.12.076>

- Meyer-Lindenberg, A. (2012). The future of fMRI and genetics research. *NeuroImage*, 62(2), 1286–1292. <https://doi.org/10.1016/j.neuroimage.2011.10.063>
- Meyer-Lindenberg, A., & Weinberger, D. R. (2006). Intermediate phenotypes and genetic mechanisms of psychiatric disorders. *Nature Reviews Neuroscience*, 7(10), 818–827. <https://doi.org/10.1038/nrn1993>
- Mooney, M. A., & Wilmot, B. (2015). Gene set analysis: A step-by-step guide. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics : The Official Publication of the International Society of Psychiatric Genetics*, 168(7), 517–27. <https://doi.org/10.1002/ajmg.b.32328>
- Mous, S. E., Hammerschlag, A. R., Polderman, T. J. C., Verhulst, F. C., Tiemeier, H., van der Lugt, A., ... Posthuma, D. (2015). A Population-Based Imaging Genetics Study of Inattention/Hyperactivity: Basal Ganglia and Genetic Pathways. *Journal of the American Academy of Child and Adolescent Psychiatry*, 54(9), 745–52. <https://doi.org/10.1016/j.jaac.2015.05.018>
- Munafò, M. R., & Flint, J. (2014). The genetic architecture of psychophysiological phenotypes. *Psychophysiology*, 51(12), 1331–1332. <https://doi.org/10.1111/psyp.12355>
- Murphy, S. E., Norbury, R., Godlewska, B. R., Cowen, P. J., Mannie, Z. M., Harmer, C. J., & Munafò, M. R. (2013). The effect of the serotonin transporter polymorphism (5-HTTLPR) on amygdala function: a meta-analysis. *Molecular Psychiatry*, 18(4), 512–520. <https://doi.org/10.1038/mp.2012.19>
- Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, 15(1), 1–25. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11747097>
- Nickl-Jockschat, T., Janouschek, H., Eickhoff, S. B., & Eickhoff, C. R. (2015). Lack of Meta-Analytic Evidence for an Impact of COMT Val158Met Genotype on Brain Activation During Working Memory Tasks. *Biological Psychiatry*, 78(11), e43–e46. <https://doi.org/10.1016/j.biopsych.2015.02.030>
- Nord, C. L., Valton, V., Wood, J., & Roiser, J. P. (2017). Power-up: A Reanalysis of “Power Failure” in Neuroscience Using Mixture Modeling. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, 37(34), 8051–8061. <https://doi.org/10.1523/JNEUROSCI.3592-16.2017>
- Nymberg, C., Jia, T., Lubbe, S., Ruggeri, B., Desrivieres, S., Barker, G., ... Schumann, G. (2013). Neural mechanisms of attention-deficit/hyperactivity disorder symptoms are stratified by MAOA genotype. *Biological Psychiatry*, 74(8), 607–14. <https://doi.org/10.1016/j.biopsych.2013.03.027>
- Papachristou, C., Ober, C., & Abney, M. (2016). A LASSO penalized regression approach for genome-wide association analyses using related individuals: application to the Genetic Analysis Workshop 19 simulated data. *BMC Proceedings*, 10(S7), 53. <https://doi.org/10.1186/s12919-016-0034-9>
- Pe’er, I., Yelensky, R., Altshuler, D., & Daly, M. J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants.

- Genetic Epidemiology*, 32(4), 381–385. <https://doi.org/10.1002/gepi.20303>
- Pearlson, G. D., Liu, J., & Calhoun, V. D. (2015). An introductory review of parallel independent component analysis (p-ICA) and a guide to applying p-ICA to genetic data and imaging phenotypes to identify disease-associated biological pathways and systems in common complex disorders. *Frontiers in Genetics*, 6, 276. <https://doi.org/10.3389/fgene.2015.00276>
- Peng, R. D. (2011). Reproducible research in computational science. *Science (New York, N.Y.)*, 334(6060), 1226–7. <https://doi.org/10.1126/science.1213847>
- Pérez-Palma, E., Bustos, B. I., Villamán, C. F., Alarcón, M. A., Avila, M. E., Ugarte, G. D., ... NIA-LOAD/NCRAD Family Study Group. (2014). Overrepresentation of Glutamate Signaling in Alzheimer's Disease: Network-Based Pathway Enrichment Using Meta-Analysis of Genome-Wide Association Studies. *PLoS ONE*, 9(4), e95413. <https://doi.org/10.1371/journal.pone.0095413>
- Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *BMJ (Clinical Research Ed.)*, 316(7139), 1236–8. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/9553006>
- Picciotto, M. (2018). Analytical Transparency and Reproducibility in Human Neuroimaging Studies. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, 38(14), 3375–3376. <https://doi.org/10.1523/JNEUROSCI.0424-18.2018>
- Pineda, S., Milne, R. L., Calle, M. L., Rothman, N., López de Maturana, E., Herranz, J., ... Malats, N. (2014). Genetic Variation in the TP53 Pathway and Bladder Cancer Risk. A Comprehensive Analysis. *PLoS ONE*, 9(5), e89952. <https://doi.org/10.1371/journal.pone.0089952>
- Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., ... Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, 18(2), 115–126. <https://doi.org/10.1038/nrn.2016.167>
- Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature*, 538(7624), 161–164. <https://doi.org/10.1038/538161a>
- Potkin, S. G., Turner, J. A., Guffanti, G., Lakatos, A., Fallon, J. H., Nguyen, D. D., ... Macciardi, F. (2009). A genome-wide association study of schizophrenia using brain activation as a quantitative phenotype. *Schizophrenia Bulletin*, 35(1), 96–108. <https://doi.org/10.1093/schbul/sbn155>
- Schnack, H. G., & Kahn, R. S. (2016). Detecting Neuroimaging Biomarkers for Psychiatric Disorders: Sample Size Matters. *Frontiers in Psychiatry*, 7, 50. <https://doi.org/10.3389/fpsy.2016.00050>
- Shen, L., Kim, S., Risacher, S. L., Nho, K., Swaminathan, S., West, J. D., ... Initiative, the A. D. N. (2010). Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort. *NeuroImage*, 53(3), 1051–63. <https://doi.org/10.1016/j.neuroimage.2010.01.042>

- Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, 39(5), 1–13. <https://doi.org/10.18637/jss.v039.i05>
- Stingo, F. C., Guindani, M., Vannucci, M., & Calhoun, V. D. (2013). An Integrative Bayesian Modeling Approach to Imaging Genetics. *Journal of the American Statistical Association*, 108(503), 876–891. <https://doi.org/10.1080/01621459.2013.804409>
- Sui, J., Adali, T., Pearlson, G., Yang, H., Sponheim, S. R., White, T., & Calhoun, V. D. (2010). A CCA+ICA based model for multi-task brain imaging data fusion and its application to schizophrenia. *NeuroImage*, 51(1), 123–34. <https://doi.org/10.1016/j.neuroimage.2010.01.069>
- Taylor, M. B., & Ehrenreich, I. M. (2014). Genetic Interactions Involving Five or More Genes Contribute to a Complex Trait in Yeast. *PLoS Genetics*, 10(5), e1004324. <https://doi.org/10.1371/journal.pgen.1004324>
- Taylor, M. B., & Ehrenreich, I. M. (2015). Higher-order genetic interactions and their contribution to complex traits. *Trends in Genetics*, 31(1), 34–40. <https://doi.org/10.1016/j.tig.2014.09.001>
- Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K.-A., Grill, J., & Frouin, V. (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics*, 15(3), 569–583. <https://doi.org/10.1093/biostatistics/kxu001>
- Tenenhaus, M., Tenenhaus, A., & Groenen, P. J. F. (2017). Regularized Generalized Canonical Correlation Analysis: A Framework for Sequential Multiblock Component Methods. *Psychometrika*, 82(3), 737–777. <https://doi.org/10.1007/s11336-017-9573-x>
- Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3), 273–282. <https://doi.org/10.1111/j.1467-9868.2011.00771.x>
- Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Tay-Lor, J., & Tibshirani, R. J. (n.d.). Strong Rules for Discarding Predictors in Lasso-type Problems. Retrieved from <https://statweb.stanford.edu/~tibs/ftp/strong.pdf>
- Van Der Maaten, L., Postma, E., & Van Den Herik, J. (2009). Tilburg centre for Creative Computing Dimensionality Reduction: A Comparative Review Dimensionality Reduction: A Comparative Review. Retrieved from <http://www.uvt.nl/ticc>
- Vergara, V. M., Ulloa, A., Calhoun, V. D., Boutte, D., Chen, J., & Liu, J. (2014). A three-way parallel ICA approach to analyze links among genetics, brain structure and brain function. *NeuroImage*, 98, 386–94. <https://doi.org/10.1016/j.neuroimage.2014.04.060>
- Vilor-Tejedor, N., Gonzalez, J. R., & Calle, M. L. (2015). Efficient and powerful method for combining p-values in Genome-wide Association Studies. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. <https://doi.org/10.1109/TCBB.2015.2509977>

- Visser, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, *101*(1), 5–22. <https://doi.org/10.1016/j.ajhg.2017.06.005>
- Vounou, M., Janousova, E., Wolz, R., Stein, J. L., Thompson, P. M., Rueckert, D., ... Alzheimer's Disease Neuroimaging Initiative. (2012). Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. *NeuroImage*, *60*(1), 700–716. <https://doi.org/10.1016/j.neuroimage.2011.12.029>
- Vounou, M., Nichols, T. E., Montana, G., & Alzheimer's Disease Neuroimaging Initiative. (2010). Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *NeuroImage*, *53*(3), 1147–1159. <https://doi.org/10.1016/j.neuroimage.2010.07.002>
- Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S. L., ... Alzheimer's Disease Neuroimaging Initiative. (2012). Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. *Bioinformatics (Oxford, England)*, *28*(2), 229–37. <https://doi.org/10.1093/bioinformatics/btr649>
- Wang, L., Jia, P., Wolfinger, R. D., Chen, X., & Zhao, Z. (2011). Gene set analysis of genome-wide association studies: Methodological issues and perspectives. *Genomics*, *98*(1), 1–8. <https://doi.org/10.1016/J.YGENO.2011.04.006>
- Wang, Y., Goh, W., Wong, L., & Montana, G. (2013). Random forests on Hadoop for genome-wide association studies of multivariate neuroimaging phenotypes. *BMC Bioinformatics*, *14*(Suppl 16), S6. <https://doi.org/10.1186/1471-2105-14-S16-S6>
- Wen, D., Jia, P., Lian, Q., Zhou, Y., & Lu, C. (2016). Review of Sparse Representation-Based Classification Methods on EEG Signal Processing for Epilepsy Detection, Brain-Computer Interface and Cognitive Impairment. *Frontiers in Aging Neuroscience*, *8*, 172. <https://doi.org/10.3389/fnagi.2016.00172>
- Whalley, H. C., Papmeyer, M., Sprooten, E., Romaniuk, L., Blackwood, D. H., Glahn, D. C., ... McIntosh, A. M. (2012). The influence of polygenic risk for bipolar disorder on neural activation assessed using fMRI. *Translational Psychiatry*, *2*(7), e130. <https://doi.org/10.1038/tp.2012.60>
- Winkler, A. M., Kochunov, P., Blangero, J., Almasy, L., Zilles, K., Fox, P. T., ... Glahn, D. C. (2010). Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetics studies. *NeuroImage*, *53*(3), 1135–1146. <https://doi.org/10.1016/j.neuroimage.2009.12.028>
- Witten, D. M., Tibshirani, R., & Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics (Oxford, England)*, *10*(3), 515–34. <https://doi.org/10.1093/biostatistics/kxp008>
- Wold, S., Martens, H., & Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method (pp. 286–293). Springer, Berlin, Heidelberg. <https://doi.org/10.1007/BFb0062108>

- Wray, N. R., Lee, S. H., Mehta, D., Vinkhuyzen, A. A. E., Dudbridge, F., & Middeldorp, C. M. (2014). Research Review: Polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry*, 55(10), 1068–1087. <https://doi.org/10.1111/jcpp.12295>
- Yan, J., Du, L., Kim, S., Risacher, S. L., Huang, H., Moore, J. H., ... Alzheimer's Disease Neuroimaging Initiative. (2014). Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm. *Bioinformatics (Oxford, England)*, 30(17), i564-71. <https://doi.org/10.1093/bioinformatics/btu465>
- Yancey, J. R., Venables, N. C., Hicks, B. M., & Patrick, C. J. (2013). Evidence for a Heritable Brain Basis to Deviance-Promoting Deficits in Self-Control. *Journal of Criminal Justice*, 41(5). <https://doi.org/10.1016/j.jcrimjus.2013.06.002>
- Zeng, Y., & Breheny, P. (2017). The biglasso Package: A Memory- and Computation-Efficient Solver for Lasso Model Fitting with Big Data in R. Retrieved from <http://arxiv.org/abs/1701.05936>
- Zhang, Z., Huang, H., Shen, D., & Alzheimer's Disease Neuroimaging Initiative. (2014). Integrative analysis of multi-dimensional imaging genomics data for Alzheimer's disease prediction. *Frontiers in Aging Neuroscience*, 6, 260. <https://doi.org/10.3389/fnagi.2014.00260>
- Zhe, S., Xu, Z., Qi, Y., Yu, P., & ADNI. (2014). Joint association discovery and diagnosis of Alzheimer's disease by supervised heterogeneous multiview learning. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 300–11. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/24297556>

LIST OF FIGURES

Figure 1. Framework of analytical strategies in IG studies.

Figure 2. Percentage of studies classified by statistical method and total number of studies classified by analytical approach. Legend: sCCA: sparse Canonical Correlation Analysis; sPLS: sparse Partial Least Squares; pICA: parallel Independent Component Analysis; IBFA: Bayesian Interbattery factor analysis; sRRR: sparse Rank Reduction Regression; RF: Random Forest; TGSL: Tree-guided sparse learning; sMML: sparse Multimodel learning; SVR: Support vector regression; LASSO: Least Absolute Shrinkage and Selection Operator; FS: Feature Selection.

Figure 3. Total number of studies classified by neuroimaging technique. a) Percentage of studies using sMRI classified by analytical approach. b) Percentage of studies using fMRI classified by analytical approach. Legend: sMRI: structural magnetic resonance imaging; fMRI: functional magnetic resonance imaging; DTI: diffusion tensor imaging.

Figure 4. Total number of studies classified by study outcome. a) Percentage of studies using brain imaging as the intermediate phenotype, classified by analytical approach. b) Percentage of Schizophrenia studies classified by analytical approach. c) Percentage of Alzheimer's disease studies classified by analytical approach.

Figures:

Structural Magnetic Resonance Imaging

Functional Magnetic Resonance Imaging

LIST OF TABLES

Table 1. Summary of the Imaging Genetics studies selected for this systematic review.

Author	Phenotype	Project	Neuroimaging Modality	Omics data	Analytical Approach	Statistical Modelling
Batmanghelich et al., (2013)	image-based features as an intermediate phenotype and AD status	ADNI	sMRI (voxels or ROIS)	SNPs	Sparse feature selection	Bayesian Hierarchical model
Cao et al., (2013)	Schizophrenia	MCIC	fMRI	SNPs	Sparse feature Selection	Independent variable selection and multivariate classification approach
Chen et al., (2012)	Schizophrenia	MCIC	fMRI	SNPs	Sparse block methods	Three-step analysis: (1) SNP selection (ANOVA) (2) parallel-ICA (3) Pathway Analysis
Chen et al., (2013)	Schizophrenia	MCIC and PGC	sMRI	SNPs	Sparse block methods	pICA-R
Chi et al., (2013)	diffusivity values as an intermediate		DTI	SNPs	Sparse block methods	Sparse CCA

	phenotype					
Clemm von Hohenberg et al., (2013)	diffusivity values as an intermediate phenotype	Dep. Psychiatry University Munich	DTI	SNPs	Others	ANCOVA
Du L et al., (2014)	Alzheimer	ADNI	sMRI	SNPs	Sparse block methods	New sparse CCA
Filipovych R et al., (2012)	Alzheimer	ADNI	sMRI	SNPs	Polygenic score	Polygenic score
Ge et al., (2012)	image-based features as an intermediate phenotype	ADNI	sMRI	SNPs	Sparse Feature Selection	random field theory
Grellmann C et al.,	image-based features as an intermediate phenotype		sMRI	SNPs	Sparse block methods	sCCA and PLSc and IBFA
Hao et al., (2014)	image-based features as an intermediate phenotype	ADNI	sMRi	SNPs	Sparse feature selection	(1) Feature selection via TGSL (2) Prediction analysis via SVR
Hao et al., (2017)	image-based features as an intermediate phenotype	ADNI	sMRI	SNPs	Sparse block methods	Three sparse CCA
Hibar et al., (2011)	image-based features as an	ADNI	sMRI	SNPs	GSA-voxelGeneWAS-Pathway Analysis-	Massive Gene-set Analysis

	intermediate phenotype				Network Analysis	
Hibar et al., (2013)	image-based features as an intermediate phenotype	ADNI, QTIM	sMRI	SNPs	Epistasis	Epistasis
Hibar et al., (2013)	image-based features as an intermediate phenotype	ADNI, QTIM	sMRI	SNPs	GWAS	SEM
Inkster et al., (2010)	Major Depressive Disorder		sMRI	SNPs	GSA-Pathway Analysis-Network Analysis	voxelGeneWAS
Kauppi et al., (2014)	Memory	BETULA	sMRI	SNPs	Epistasis	GLM (main effects and interaction) and ANOVA
Khadka et al., (2016)	ADHD		sMRI	SNPs	Sparse block methods and GSA-voxelGeneWAS-Pathway Analysis-Network Analysis	pICA and GSA
Kohannim et al., (2012)	Alzheimer	ADNI	sMRI	SNPs	Data-driven analysis	LASSO and partial F-tests
Le Floch et al., (2012)	image-based features as an intermediate phenotype		fMRI	SNPs	Sparse block methods	Sparse PLS and CCA
Lin et al., (2014)	Schizophrenia		fMRI	SNPs	Sparse block methods	Two-step Analysis: (1) GSA and (2)

						Sparse CCA
Mattingsdal et al., (2013)	image-based features as an intermediate phenotype		fMRI	SNPs	Data-driven analysis / GSA-voxelGeneWAS-Pathway Analysis-Network Analysis	PCA and pathway analysis
Meda et al., (2010)	image-based features as an intermediate phenotype		fMRI	SNPs	Sparse block methods	pICA
Meda et al., (2012)	Alzheimer	ADNI	sMRI	SNPs	Sparse block methods	ICA
Meda et al., (2013)	Alzheimer	ADNI	sMRI	SNPs	Epistasis	Epistasis
Mounce et al., (2014)	FA	MCIC	DTI	SNPs	Data-driven analysis	ICA
Mous et al., (2015)	ADHD	GEN-R	sMRI	SNPs	GSA-Pathway Analysis-Network Analysis	voxelGeneWAS
Roussotte et al., (2014)	Alzheimer	ADNI	sMRI	SNPs	Others	Linear discriminant analysis-based approach
Sheng et al., (2015)	image-based features as an intermediate phenotype	ADNI	sMRI	SNPs	Sparse block methods	Sparse CCA

Stingo et al., (2013)	image-based features as an intermediate phenotype	MCIC	fMRI	SNPs	Others	(1) Selection of ROIs that discriminate case/control outcome status (2) Group-specific distribution depending on selected SNPs (3) Network models
Strijbis et al., (2013)	Multiple Sclerosis	GeneMSA	sMRI	SNPs	Massive Univariate Analysis / Data-driven analysis / Sparse multivariate regression	(1) MULM (2) LASSO (3) Sparse RRR
Vergara et al., (2014)	image-based features as an intermediate phenotype		sMRI and fMRI	SNPs	Sparse block methods	pICA
Vilor-Tejedor et al., (submitted)	ADHD	BREATHE	sMRI	SNPs	Data-driven analysis	(1) LASSO (2) MFA
Vounou et al., (2010)	simulated image-based features as an intermediate phenotype	ADNI	sMRI	SNPs	Sparse multivariate regression	Sparse RRR
Vounou et al., (2012)	image-based features as an intermediate phenotype	ADNI	fMRI	SNPs	Sparse multivariate regression	Sparse RRR
Wang H et al., (2012)	image-based features as an intermediate phenotype	ADNI	sMRI	SNPs	Sparse multivariate regression	G-SMuRFs

Wang H et al., (2013)	image-based features as an intermediate phenotype	ADNI	sMRI	SNPs	Sparse multivariate regression	RF regression
Whalley et al., (2012)	Bipolar Disorder	PGC-BD	fMRI	SNPs	Polygenic score	Polygenic score
Yan et al., (2014)	image-based features as an intermediate phenotype	ADNI	sMRI and PET	SNPs	Sparse block methods	knowledge-guided sparse CCA
Zhang et al., (2014)	Alzheimer	ADNI	sMRI, PET, CSF	SNPs	Sparse feature selection	Sparse MML
Zhe et al., (2014)	Alzheimer	ADNI	sMRI	SNPs	Data-driven analysis	Bayesian sparse projection

Legend	
<i>ADHD</i>	Attention-Deficit/Hyperactivity Disorder
<i>ADNI</i>	Alzheimer's Disease Neuroimaging Initiative
<i>BREATHE</i>	BRain dEvelopment and Air polluTion ultrafine particles in scHool childrEn
<i>CCA</i>	Canonical correlation analysis
<i>CSF</i>	Cerebrospinal fluid flow
<i>DTI</i>	Diffusion Tensor Imaging
<i>FA</i>	fractional anisotropy
<i>fMRI</i>	functional Magnetic Resonance Imaging
<i>GEN-R</i>	Generation R
<i>GeneMSA</i>	Gene Multiple Sclerosis consortium

<i>GSA</i>	Gene-set Analysis
<i>G-SMuRFs</i>	Group-sparse Multi-task regression and feature selection
<i>GWAS</i>	Genome-wide Association Study
<i>ICA</i>	Independent Component Analysis
<i>LASSO</i>	Least Absolute Shrinkage and selection operator
<i>MCIC</i>	multisite Mind Clinical Imaging Consortium
<i>MML</i>	Multi model learning
<i>PCA</i>	Principal Component Analysis
<i>pICA</i>	parallel ICA
<i>pICA-R</i>	pICA including a Reference
<i>PET</i>	Positron emission tomography
<i>PGC-BD</i>	Psychiatric GWAS Consortium Bipolar Disorder Working Group
<i>PLS</i>	Partial Least Squares
<i>QTIM</i>	Queensland Twin Imaging Study
<i>RF</i>	Random Forest
<i>RRR</i>	Reduced Rank Regression
<i>ROI</i>	Region of Interest
<i>RVS</i>	Variable selection representation
<i>SEM</i>	Structural Equation Modelling
<i>sMRI</i>	structural Magnetic Resonance Imaging
<i>SNPs</i>	Single Nucleotide Polymorphism
<i>SVR</i>	Support vector regression
<i>TGSL</i>	Tree-Guided Sparse Learning

Table 2. Summary of R packages and functions for multivariate analysis of multiblock data.

Method	Description	R-function{R-Package}
<i>1 table</i>		
PCA	Principal component analysis	prcomp{stats}, princomp{stats}, dudi.pca{ade4}, pca{vegan}, PCA{FactoMineR}, principal{psych}
sPCA	Sparse PCA ($n < p$)	spca{mixOmics}, SPC{PMA}, big.PCA{bigpca}*
CA, COA	Correspondence analysis	ca{ca}, CA{FactoMineR}, dudi.coa{ade4}
CIA	Coinertia analysis	cia{made4}
MCA	Multiple correspondence analysis	dudi.acm{ade4}, mca{MASS}
ICA	Independent component analysis	fastICA{FastICA}
sIPCA	Sparse independent PCA (combines sPCA and ICA)	sipca{mixOmics} ipca{mixOmics}
LASSO	Least Absolute Shrinkage and Selection Operator	glmnet{glmnet}, ncvreg{ncvreg}, biglasso{biglasso}*
ENET	Elastic Net	glmnet{glmnet}, enet{elasticnet}, biglasso{biglasso}*
IBFA	Interbattery factor analysis	GFA{CCAGFA}
<i>2 tables</i>		
CCA	Canonical correlation analysis. [Limited to $n > p$]	cc{cca}; CCorA{vegan}
RDA	Redundancy analysis is a constrained PCA.	rda{vegan}
rCCA	Regularized canonical correlation	rcc{cca}, rcc{mixOmics}
sCCA	Sparse CCA	CCA{PMA}
pCCA	Penalized CCA	spCCA{spCCA} supervised version
PLS	Partial least squares of K-tables (multi-block PLS)	mbpls{ade4}, plsda{caret}

sPLS pPLS	Sparse PLS Penalized PLS	<code>spls{spls} spls{mixOmics} ppls{ppls}</code>
O2PLS	two-way ortogonal PLS	<code>o2m{github.com/selbouhaddani/OmicsPLS}</code>
RF	Random Forest	<code>randomForest{randomForest}</code>
sRRR	Sparse Reduced-Rank regression	<code>rssvd{rrpack}</code>
MKL	Multiple Kernel Learning	<code>komd{github.com/jmikko/EasyMKL}</code>
DISCOSCA, JIVE, O2PLS	Disco SCA, jive, two-way ortogonal PLS	<code>omicsCompAnalysis{STATegRa}</code>
<i>> 2 tables</i>		
gCCA	Generalized CCA	<code>regCCA{dmt}, multicCA{PMA}*</code>
MCIA	Multiple Coinertia Analysis	<code>mca{omicade4}</code>
RGCCA	Regularized generalized CCA	<code>regCCA{dmt} rgcca{RGCCA} wrapper.rgcca{mixOmics}</code>
sGCCA	Sparse generalized canonical correlation analysis	<code>sgcca{rgcca} wrapper.sgcca{mixOmics}</code>
MFA	Multiple Factor Analysis	<code>MFA{FactoMineR}</code>
pICA	parallelized ICA	<code>mica{PGICA}</code>
NPC	non parametric combination	<code>omicsNPC{STATegRa}</code>

*Specially designed for big data matrices