frontiers in
NEUROINFORMATICS

REVIEW ARTICLE
published: 26 March 2014
doi: 10.3389/fninf.2014.00029

# A review of multivariate analyses in imaging genetics

***Jingyu Liu*[1,2] ** and Vince D. Calhoun*[1,2]**

[1] *The Mind Research Network and Lovelace Biomedical and Environmental Research Institute, Albuquerque, NM, USA*

[2] *Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM, USA*

***Edited by:***
*Venkata Satyanand Mattay, Lieber Institute for Brain Development, USA*

***Reviewed by:***
*Yi-Yuan Tang, Institute of Neuroinformatics, Dalian University of Technology, China*
*Fabio Sambataro, Istituto Italiano di Tecnologia, Italy*
*Qiang Chen, Lieber Institute for Brain Development, USA*

****Correspondence:***
*Jingyu Liu, Mind Research Network and Lovelace Biomedical and Environmental Research Institute, 1101 Yale Boulevard NE, Albuquerque, NM 87106, USA*
*e-mail: jliu@mrn.org*

Recent advances in neuroimaging technology and molecular genetics provide the unique opportunity to investigate genetic influence on the variation of brain attributes. Since the year 2000, when the initial publication on brain imaging and genetics was released, imaging genetics has been a rapidly growing research approach with increasing publications every year. Several reviews have been offered to the research community focusing on various study designs. In addition to study design, analytic tools and their proper implementation are also critical to the success of a study. In this review, we survey recent publications using data from neuroimaging and genetics, focusing on methods capturing multivariate effects accommodating the large number of variables from both imaging data and genetic data. We group the analyses of genetic or genomic data into either *a priori* driven or data driven approach, including gene-set enrichment analysis, multifactor dimensionality reduction, principal component analysis, independent component analysis (ICA), and clustering. For the analyses of imaging data, ICA and extensions of ICA are the most widely used multivariate methods. Given detailed reviews of multivariate analyses of imaging data available elsewhere, we provide a brief summary here that includes a recently proposed method known as independent vector analysis. Finally, we review methods focused on bridging the imaging and genetic data by establishing multivariate and multiple genotype-phenotype-associations, including sparse partial least squares, sparse canonical correlation analysis, sparse reduced rank regression and parallel ICA. These methods are designed to extract latent variables from both genetic and imaging data, which become new genotypes and phenotypes, and the links between the new genotype-phenotype pairs are maximized using different cost functions. The relationship between these methods along with their assumptions, advantages, and limitations are discussed.

**Keywords: imaging genetics, multivariate analyses, genotype, phenotype, intermediate phenotypes**

## INTRODUCTION

While most genetic studies have focused on phenotypes as diagnoses and clinical symptoms, it is relatively recent that intermediate phenotypes have become an ever increasing focus. Intermediate phenotypes refer to biological trait phenotypes conveying relatively closer association or higher penetration than traditional phenotypes (Meyer-Lindenberg and Weinberger, 2006; Rasetti and Weinberger, 2011). The best examples of approaches leveraging intermediate phenotypes come from studies of psychiatric disorders for which diagnoses are based mainly on clinical observations and interviews. Intermediate phenotypes derived from neuroimaging and signals directly assessing brain structure and function not only reduce the phenotypic heterogeneity common to many psychiatric disorders, but also increase detection power, given the genetic effects are not expressed directly as behaviors but as molecular and cellular functions mediating brain development and processes (Gottesman and Gould, 2003; Rose and Donohoe, 2013). The pioneer studies utilizing neuroimaging features to identify genetic impact were in the year 2000 (Bookheimer et al., 2000; Heinz et al., 2000; Small et al., 2000). They signified the birth of a new research approach using imaging genetics. As defined (Hariri et al., 2006; Meyer-Lindenberg et al., 2008; Silver et al., 2011; Meyer-Lindenberg, 2012), it combines genetic information and neuroimaging data in the same subjects to discover neuromechanisms linked to psychiatric disorders. The overall strength of imaging genetics and its impact on psychiatric disorder studies or broader have been stated clearly in several reviews (Meyer-Lindenberg and Weinberger, 2006; Glahn et al., 2007; Bigos and Weinberger, 2010; Meyer-Lindenberg, 2010; Rasetti and Weinberger, 2011).

The overwhelming growth of imaging genetics in recent years as summarized in recent studies (Roffman et al., 2006; Bigos and Weinberger, 2010), while providing abundant promising results, also reveals challenges embedded within study designs such as validity of candidate genes, control of non-genetic confounding factors, and selection of tasks to stimulate brain specific processes. Bigos and Weinberger (2010) have provided an excellent review with applications to demonstrate the principles in designing an imaging genetic study. Another big challenge faced by both imagers and geneticists is how to properly analyze the collected data, since both neuroimaging and genetics tend to generate a large amount of data. Different strategies, processing approaches, and validation methods such as false positive control (Silver et al., 2011) have been implemented and tailored for different conditions. But there is an even greater need in the future for the methodology development as pointed by Mayer-Lindenberg in his recent review (Meyer-Lindenberg, 2012), where complexity of epistasis, pleiotropy and genetic by environment interactions should been considered in particular in large scale genomic studies. The availability of imaging genetic analytic tools and their proper implementation are critical for both success of individual studies and the continuing growth of imaging genetics.

The earliest imaging genetic studies focused on candidate genetic variants using either a single or a few variables (Bookheimer et al., 2000; Heinz et al., 2000; Small et al., 2000; Egan et al., 2001). For example, the dopamine transporter gene (SLC6A) was analyzed with neuroimaging data from single-photon emission computed tomography (Heinz et al., 2000). Variation within the APOE gene was associated with activities in memory function affected by Alzheimer's disease (Bookheimer et al., 2000). COMT Val allele carriers showed increased activities in the prefrontal cortex compared to Met allele carriers (Egan et al., 2001). In parallel, the intermediate phenotypes from neuroimaging techniques can also be specified within selected brain regions or particular processes. Straightforward univariate analyses are often used and well suited for these studies. Candidate gene and candidate imaging phenotype studies in the last decade have proven the validity of imaging genetic approach as recapitulated in (Meyer-Lindenberg, 2012). But with the completion of human genome sequence and multimodal imaging practices, in conjunction with increased evidence of polygenicity and pleiotropy (Purcell et al., 2009; Sivakumaran et al., 2011; Whalley et al., 2012; Smoller et al., 2013), multivariate analysis methods are becoming more and more demanding. For instance, thousands of genetic variants have been suggested to be linked with the risk for schizophrenia (Purcell et al., 2009). Methods to capture the interactive or integrated genetic effects of a set of genetic variants, methods to extract brain networks formed from individual voxels or regions, and methods to detect, possibly, multiple genotype-phenotype connections have been developed with their limitations and advantages (Hardoon et al., 2009; Liu et al., 2009b; Vounou et al., 2010; Le Floch et al., 2012). We expect to see continued development of such powerful methods to face the challenges and promises from genome-wide whole brain association studies.

In this review, we focus on analysis approaches and, more specifically, on the multivariate analysis approaches. We will first give an overview of analysis strategies. Then, we will survey the methods and organize them according to their multivariate nature on genetic data, neuroimaging data or both.

## OVERVIEW OF ANALYSIS STRATEGIES IN IMAGING GENETICS

While various strategies can be applied to design and perform imaging genetic studies, several aspects of such studies require particular caution. Firstly, when an imaging feature is selected as the intermediate or endophenotype, useful criteria should be applied or at least considered. As summarized in (Gottesman and Gould, 2003) intermediate phenotypes should show association with illness in a population, certain level of heritability, and state-independent characters. A proper preprocessing or controlling for possible confounding factor should also be in place, such as scanning effects, age or gender difference, brain size, etc. The most often used software packages to process brain imaging data, particularly for magnetic resonance imaging (MRI) images, include FSL[1] , SPM[2], and AFNI[3] for functional and structural voxel-wise preprocessing, and FreeSurfer[4] for brain regional volume and cortical thickness. Secondly, genetic data either from single genetic mutation or genomic variants should be checked for family structure, population structure, and ethnicity differences. A rationale to pull samples together should be justified through, for instance, from a homogenous group, no indication of population structure, or a proper control of ethnicity difference. The most often used software package for single nucleotide polymorphism (SNP) data is plink[5], which provides tools to do various quality control, sample relatedness tests, filtering and population stratification. The most often used software packages (freely available) for calling copy number variation (CNV) include PennCNV[6], and BirdSuite[7]. Even though the effect of CNVs on brain imaging phenotypes is understudied now, it has been predicted to be an important extension in the future (Meyer-Lindenberg, 2012). Thirdly, methods to test the relation between genetics and imaging phenotypes heavily rely on the dimensionality of data, as explained explicitly in next paragraph. Finally, the interpretation of results depends on the study design and analysis approaches. Keep in mind that most imaging genetic studies test the association between genetic variants and imaging phenotypes, as the analytical method itself reveals later on. Any causal relation and underlying biological mechanism is only suggestive. Particular caution should be given to genome-wise association studies which result in a set of genetic variants interactively associated with imaging phenotypes. The interaction among them, linear, non-linear, dominate, recessive, two-way or n-way, etc., needs to be carefully explained and some methods test the overall effect without knowing the detailed interrelations. The verification or at least certain levels of cross evaluation for such findings as described in (Le Floch et al., 2012) plays a very crucial role.

Depending on the dimensionality of investigated genotypes and imaging intermediate phenotypes, we can classify imaging genetic studies into four categories, which is a concept borrowed from Vounou et al. (2010). As plotted in **Figure 1**, the first one includes studies with candidate phenotypes and candidate genotypes, where a direct univariate association test is applied to assess the hypothesized connection. A control for possible confounding factors (scanner, age, gender, medication, etc.) should be considered for imaging phenotypes. The second type includes studies investigating multiple genetic variants, ranging from a few to 100s of 1000s of variables in a genome-wide setting. Univariate tests corrected for multiple comparisons are straightforward (Potkin et al., 2009), but it may miss the well

[1] http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/

[2] http://www.fil.ion.ucl.ac.uk/spm/

[3] http://afni.nimh.nih.gov/afni/

[4] https://surfer.nmr.mgh.harvard.edu/

[5] http://pngu.mgh.harvard.edu/~purcell/plink/

[6] http://www.openbioinformatics.org/penncnv/

[7] http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/birdsuite/birdsuite

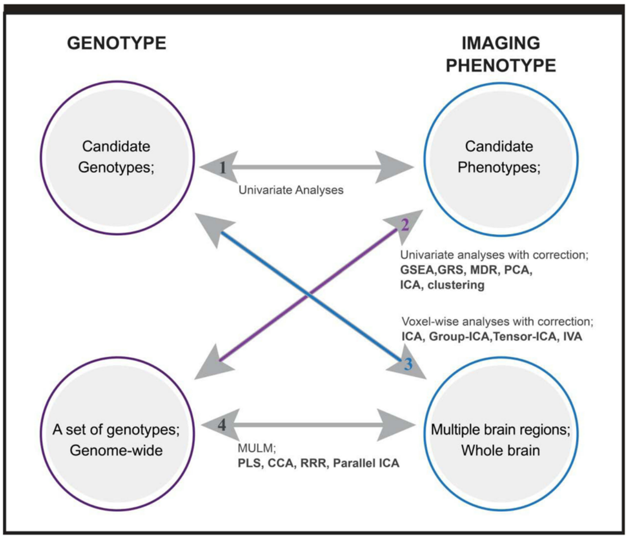
GENOTYPE
IMAGING PHENOTYPE
Candidate Genotypes;
Candidate Phenotypes;
1
Univariate Analyses
2
Univariate analyses with correction;
GSEA,GRS, MDR, PCA, ICA, clustering
Voxel-wise analyses with correction;
ICA, Group-ICA,Tensor-ICA, IVA
3
A set of genotypes; Genome-wide
4
MULM;
PLS, CCA, RRR, Parallel ICA
Multiple brain regions; Whole brain

**FIGURE 1 | Overview of imaging genetic studies and methods applied.** Category 1: candidate genotype with candidate phenotype. Category 2: sets of genotypes with candidate phenotype. Category 3: candidate genotype with multiple imaging phenotypes. Category 4: sets of genotypes with multiple imaging phenotypes. Methods written in bold are multivariate analysis methods. GSEA: gene set enrichment analysis; GRS: genetic risk score; MDR: multifactor dimensionality reduction; PCA: principal component analysis; ICA: independent component analysis; IVA: independent vector analysis; MULM: mass univariate linear model; PLS: partial least square; CCA: canonical component analysis; RRR: reduced rank regression.

documented gene–gene interactions. Data driven multivariate methods and *a priori* based gene-set or pathway analyses are the two main analytical approaches to capture the interactive or integrated genetic effect (Liu et al., 2010b; Walton et al., 2013). What type of interactive relation among genes can be captured depends on the analytic methods or specifically, the models that the methods are built on. The third type includes studies investigating multiple imaging phenotypes, which may come from one or more imaging modalities, such as structural, functional MRI, magnetic resonance spectroscopy, etc. The imaging phenotypes may cover whole brain or many brain regions or voxels. Except for voxel-wise analyses with multiple comparison correction, the strategy to analyze such phenotypes usually is to extract brain networks formed by interactive brain regions or voxels, thus not only accommodating interrelations but also reducing the number of tested phenotypes (Calhoun and Adali, 2006). The last group of studies involves associations between multiple genotypic variables and multiple phenotypic variables. A typical example is genome-wide whole brain studies. Although massive univariate approaches have been implemented such as a mass-univariate linear model (MULM) in studies (Stein et al., 2010), most utilize data reduction and factorization methods to effectively capture the interactive and complex relations within and between datasets. In the following, we present the analytical methods implemented in studies of the last three categories, category 2: sets of genotypes with candidate phenotype, category 3: candidate genotype with multiple imaging phenotypes, and category 4: sets of genotypes with multiple imaging phenotypes. We focus on the multivariate approaches for each category.

## *A priori* BASED MULTIVARIATE ANALYSES ON GENETIC/GENOMIC DATA (CATEGORY 2)

Gene set enrichment analysis (GSEA) is a computational method that determines whether a prior defined set of genetic variants shows statistically significant differences between two biological states (Mootha et al., 2003; Subramanian et al., 2005) or, more generally, significant associations with phenotypes compared to the null hypothesis. The GSEA was first introduced in cancer research and thereafter various modified versions have been introduced in studies of different diseases that includes psychiatric disorders (Subramanian et al., 2005; Holden et al., 2008; Suarez-Farinas et al., 2010; Oh et al., 2011; Weng et al., 2011). The basic principle of GSEA is that sets of genetic variants are first selected for tests. We will use SNPs as an example of genetic variants without loss of generality in this review. A set of SNPs are selected based on common biological attributes (gene ontology or pathways), chromosome location, or reported results in the literature. Then the overrepresentation, or "enrichment," of phenotype-association of this set of SNPs as one unit is calculated against the null hypothesis of normally distributed phenotype-association. Among many ways to decide the significance of enrichment (Abatangelo et al., 2009), the two most common methods are Fisher's exact test and enrichment score test (Subramanian et al., 2005). Fisher's exact test is fast but needs a pre-defined threshold, while enrichment score does not need a threshold but needs a permutation to get empirical *p* values. Specific issues associated, such as gene size bias (Mirina et al., 2012), linkage disequilibrium (LD) between adjacent SNPs, have been addressed by various modified versions (Liu et al., 2010b; Li et al., 2011). The rationale to select the set of SNPs comes from prior information, so this approach is indeed *a priori* driven test for the overall effect of multiple variables, without modeling the exact interaction among them. Another similar approach proposed by Walton et al. (2013) is to compute a cumulative genetic risk score ($GRS = \sum_{i=1}^{N} w^i x^i$), which combines the additive effects of multiple SNPs selected from the continuously updated meta-analysis of genetic studies. The authors showed that this multivariate score combined the impact of many genes with small effects, accounting for 3.6% of the total variance of brain activity at dorsal lateral prefrontal cortex (Walton et al., 2013). Similar approaches using polygenic risk scores have been implemented in several other studies (Whalley et al., 2012; Smoller et al., 2013).

## DATA DRIVEN MULTIVARIATE ANALYSES ON GENETIC/GENOMIC DATA (CATEGORY 2)

Unlike the approaches above, some studies have implemented purely data driven analyses without prior information, emphasizing the genetic patterns embedded in the datasets to capture the epistasis and polygenicity. Multifactor dimensionality reduction (MDR) was developed to identify combinations of gene–gene and gene-environmental factors that are predictive of a phenotype (Hu et al., 2011; Gui et al., 2013; Pan et al., 2013). The heart of MDR is an attribute construction algorithm that creates a new variable by pooling genotypes from multiple SNPs (Moore et al., 2010). In brief, values from any combination of multiple SNPs are classified into two distinct groups, high risk and low risk, effectively reducing the dimensionality from multidimensional to one-dimensional. Subsequently, the new variables are used to identify, from all potential combinations, the specific combination of SNPs showing the strongest association with the phenotype. This method with no particular model assumption is well suited for capturing epistasis and has been used in genetics studies of various disease status (Ritchie et al., 2001; Moore and Williams, 2002; Ma et al., 2005; Lou et al., 2007; Gui et al., 2011). Extensions of the method have been developed for quantitative phenotypes and genome-wide data (Lou et al., 2007; Pattin et al., 2009; Cattaert et al., 2011; Oh et al., 2012; Winham, 2013). It is expected to see more broad applications of this method even in imaging genetics (Papassotiropoulos and de Quervain, 2011). Within the same line of estimating aggregated effect of multiple genetic variants, but based on a linear additive model, multiple regression and its penalized or modified versions have been implemented to assess the explanation power of gene variables (from a couple to genome-wide) to various of phenotypes (Wang and Abbott, 2008; Wu et al., 2009; Cule et al., 2011). Penalized regression, specifically LASSO multiple regression, are also often used to downsize variables (voxels or SNP) for further analyses (Vounou et al., 2012).

Other types of data-driven approaches, as reviewed in (Jombart et al., 2009), mainly include principal component analysis (PCA), principal coordinate analysis, non-metric dimensional scaling, and correspondence analysis, belonging to the category of matrix decomposition and extracting factors/components of weighted genetic variants. An addition to the review is independent component analysis (ICA). PCA provides a set of linearly orthogonal principal components, explaining maximal variance, while ICA is designed to extract statistically independent components (and thus uses higher order statistical information). PCA is often used in genome-wide SNP data, and the top PCs extracted most likely present the population structure helpful for population stratification (Price et al., 2006; Liu et al., 2010a). ICA has proven successful in a variety of biological inquiries when applied to gene expression data (Kong et al., 2008), including identifying tumor-related pathways (Saidi et al., 2004; Sheng et al., 2011), classifying disease datasets (Huang and Zheng, 2006) and mining human gene expression modules (Engreitz et al., 2010).

The value of clustering methods has been established in various genetic studies, as reviewed by Jiang et al. (2004), as a means to group genetic variants according to their functional relatedness (D’haeseleer, 2005). In an example of using imaging as phenotypes, Sloan et al. (2010) applied a hierarchical clustering analysis on 834 SNPs and clinical and imaging phenotypes, including left, right hippocampal volume and gray matter density. The association between each SNP and each endpoint was first computed, and then the clustering was performed on the results, wherein both genotypes and phenotypes were grouped based on similarity. Subsequently, $p$-values for each cluster were estimated using bootstrap resampling. This study showed that (1) SNPs are frequently associated with imaging phenotypes and rarely associated with clinical scores and (2) most of the genes found within clusters are associated with either beta-amyloid production or apoptosis (Sloan et al., 2010). A noteworthy point of this study is that it combined a pathway-based approach and clustering analyses together, first by selecting SNPs based on pathways and then applying clustering on genotypes and phenotypes, and demonstrated that priori driven and data driven approaches can be integrated into one study.

## COMPONENT-BASED ANALYSES ON IMAGING DATA (CATEGORY 3)

Not only does the development of various neuroimaging techniques improve the precision of measurement of brain attributes, but it also stimulates the growth of analysis approaches. The most common imaging modalities include functional MRI (fMRI), measuring the dynamic brain activity based on blood-oxygenation-level dependent contrast; structural MRI, assessing the volume and density of gray matter, white matter, and cerebrospinal fluid; diffusion (tensor) imaging, depicting the white matter tract connections; and magnetic resonance spectroscopy, obtaining biochemical information about the tissues of brain. Furthermore, collecting multiple types of imaging data from the same individuals becomes a common practice in the hope of revealing additional information and increasing our knowledge. Thus, methods for multimodal analyses have also emerged and developed rapidly. Here, we limit ourselves to the component-based multivariate analysis approaches applied to imaging data, though there are many other multivariate approaches, such as unsupervised clustering, supervised pattern recognition, classification and projection, and others (Dimitriadou et al., 2004; Demirci et al., 2008; Hinrichs et al., 2009; Filipovych and Davatzikos, 2011).

ICA with various implementation algorithms (Cardoso, 1997; Hyvirinen and Oja, 1999; Bingham and Hyvarinen, 2000) and its modifications and extensions (Bach and Michael, 2002; Beckmann and Smith, 2004; Calhoun et al., 2005; Hong et al., 2005; Lin et al., 2010) are the most popular methods for multivariate analyses on imaging data. Several reviews have been offered to the imaging field (McKeown et al., 2003; Calhoun and Adali, 2006; Calhoun et al., 2009). Here, we briefly summarize the main points. A typical ICA model assumes that the source signals are not observable, statistically independent and non-Gaussian with an unknown but linear mixing process. Consider an observed $M$–dimensional random vector denoted by $X = [x_1, x_2,...,x_M]^T$, which is generated by the ICA model: $X = AS$, $S$ is the source matrix. The goal of ICA is to estimate an unmixing matrix $W$ such that $Y$ given by $Y = WX$ is a good approximation to the “true” sources. $Y$ is called the component matrix. In the context of imaging data, components are the independent brain networks embedded in the observed voxels. Furthermore, when MRI data from multiple subjects, each with their own temporal dynamics, are of interest, several ICA based multi-subject analysis approaches have been proposed (Calhoun et al., 2001; Schmithorst and Holland, 2004; Beckmann and Smith, 2005; Esposito et al., 2005; Erhardt et al., 2011; Calhoun and Adali, 2012). We refer to recent studies by Calhoun and Adali (2012); (Calhoun et al., 2009) for a more detailed explanation. A recent addition is independent vector analysis (IVA), which is a generalization of ICA for analysis of multiple datasets (Kim et al., 2006). It takes a model of $X^{[m]} = A^{[m]}S^{[m]}$, $Y^{[m]} = W^{[m]}X^{[m]}$, where $M$ is the number of datasets. Its cost function, the Kullback–Leibler divergence between two functions of dependence (joint probability density function of components and the product of marginal probability density function of components), allows maintaining the independency among components while increasing dependency of components between datasets (Lee et al., 2008a,b). Based on simulation (Lee et al., 2008b; Dea et al., 2011), IVA shows excellent performance in capturing inter-subject variability and the performance enhancement increases when the spatial variation of a given component across subjects is substantial.

For multimodal imaging analyses, a set of solutions with different emphases have been proposed and extensive reviews of these methods are also available (Biessmann et al., 2011; Sui et al., 2012a). Biessmann et al. (2011) reviewed the multimodal analyses from a variety of perspectives, including multimodal imaging study setup, the advances achieved in basic research and clinical applications, the methods for artifact removal, data-driven and model-driven analyses, and univariate and multivariate fusion. Sui et al. (2012a) focused on comparisons of the multivariate multimodal fusion methods rooted in ICA, canonical component analysis (CCA), and partial least squares (PLS) analysis. Similarity between methods fusing multimodal imaging data and multivariate analyses to bridge imaging and genetics are discussed in the next section.

## MULTIVARIATE ANALYSES BRIDGING IMAGING AND GENETICS (CATEGORY 4)

Given the characteristics of imaging and genetic data, multivariate multiple regression is a natural choice, where genetic variants are predictors along with other influencing factors such as age and gender, and imaging variables (regions or voxels of brain) are response variables. In practice with a set of SNPs and brain voxels (they are usually not independent to each other), regularization or modification of traditional multivariate multiple regression has to be taken in place. Wang et al. (2012a) proposed a group sparse regularization on multivariate regression. SNPs are grouped based on genes or LD blocks. A group sparsity to reduce to only genes or LD blocks relevant to all imaging phenotypes, and an individual sparsity to select only important SNPs are all enforced. Lin et al. (2012) presented a projection regression model that is also suitable for imaging genetics. The key of this model is to estimate the principal components of heritability (covariance between multiple phenotypes and genetics of interest), followed by a multivariate regression on the principle components.

When facing a very large number of genetic variants, such as genomic SNPs, and a large number of voxels in the brain, researchers in imaging genetics, very interestingly, has focused on a series of very closely related methods to capture interactive or integrated effects and possibly many genotype-phenotype pairs. These methods include PLS, CCA, reduced rank regression (RRR), and ICA (Hardoon et al., 2009; Liu et al., 2009b; Vounou et al., 2010, 2012; Le Floch et al., 2012; Meda et al., 2012; Chi et al., 2013). They are designed to simultaneously extract latent variables from both genetic and imaging data, which become new genotypes and phenotypes, and the connections of new geno-pheno variables are maximized using different cost functions.

We can use a typical imaging genetic example to illustrate the relation of these methods. We denote by $X$ an $n \times p$ matrix of genetic SNP data, and by $Y$ an $n \times q$ matrix of imaging data, where $n$ is the sample size, $p$ is the size of SNP loci, $q$ is the size of voxels, and $n << p$ or $q$. The latent variables are obtained through projecting the $X$ or $Y$ to new directions formed by the vectors in $U$ or $V$ matrices. **Figure 2** plots the cost function of each method and the condition under which two different methods become equivalent. PLS maximizes the covariance between latent variables of the two modalities, while CCA maximizes the correlation between them. In a high-dimensional problem where the number of variables is significantly larger than the number of samples, it is common to assume that the covariance matrices of $X$ and $Y$ are diagonal (Vounou et al., 2010; Le Floch et al., 2012). Under such a condition, CCA and PLS become equivalent. The RRR model takes a more general formation that begins from a multivariate linear regression from $X$ to $Y$, and reduces the rank of the project matrix, a product of $UV'$. Through minimizing the regression error noted as$(Y - XUV')\Gamma(Y - XUV')'$, RRR obtains the project matrices $U$ and $V$. When the function of $\Gamma$ is the identify matrix, RRR is equivalent to PLS, and when the function of $\Gamma$ is the inverse of covariance matrix $Y'Y$, RRR is equivalent to CCA. Note that the core computations of PLS, CCA and RRR all involve single value decomposition so that the latent variables or projection vectors within one modality (genetic or imaging) are orthogonal to each other. In contrast, ICA emphasizes that latent variables (components) are maximally independent from each other, which can be optimized through many forms of statistical measures, including minimization of mutual information and maximization of non-Gaussianity. One extension of ICA methods applied to imaging genetics is parallel ICA, which simultaneously maximizes both the independence of components and correlations between projection vectors of the two modalities (Liu et al., 2008b).

Parallel ICA was first introduced into imaging genetics in 2009 (Liu et al., 2009b) when applied to a genetic study of schizophrenia

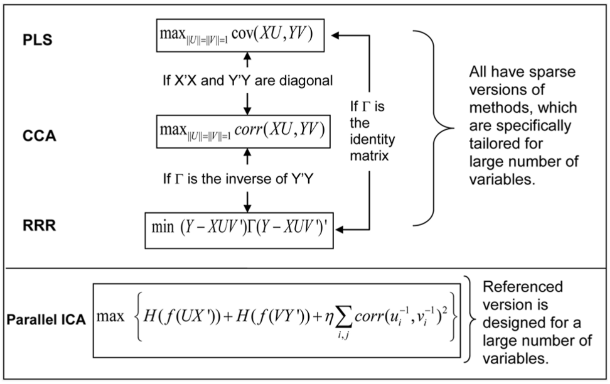
PLS
$\max_{\|U\|=\|V\|=1} \operatorname{cov}(XU, YV)$
If X'X and Y'Y are diagonal
CCA
$\max_{\|U\|=\|V\|=1} corr(XU, YV)$
If Γ is the inverse of Y'Y
RRR
$\min\ (Y - XUV')\Gamma(Y - XUV')'$
If Γ is the identity matrix
All have sparse versions of methods, which are specifically tailored for large number of variables.
Parallel ICA
$\max \left\{ H(f(UX')) + H(f(VY')) + \eta \sum_{i,j} corr(u_i^{-1}, v_i^{-1})^2 \right\}$
Referenced version is designed for a large number of variables.

**FIGURE 2 | Cost functions of four multivariate association methods, and their relation and extensions for a large number of variables.**

with a 384 SNP array and auditory oddball fMRI data. Since then, this method has been made available for the public through the fusion ICA toolbox[8]. This approach has been utilized by various other groups (Jagannathan et al., 2010; Meda et al., 2010, 2012; Meier et al., 2012). A noteworthy point is that parallel ICA can also be applied onto other types of data in addition to genetics and images (Liu and Calhoun, 2007; Liu et al., 2009a; Wu et al., 2011; Meier et al., 2012). A simulation study showed that parallel ICA performs better within a certain range of sample size vs. genetic variable ratio (Liu et al., 2008a). When a genome-wide high-density large genetic array (e.g., >100K SNP loci) is in place with a relatively small sample size, new extensions of parallel ICA are proposed to improve the performance by incorporating prior information about genetic or imaging data called parallel ICA with reference (Liu et al., 2012a; Chen et al., 2013). As showed by Chen et al. (2013), this approach leverages prior knowledge of known genetic functions to guide ICA for specific components. Thus, a specific SNP factor centered at gene ANK3, which is a schizophrenia susceptibility gene (Ripke et al., 2011), was extracted from a large SNP array (>700K loci). While this method does help extract particular genetic components, which may not be extracted otherwise (Liu et al., 2012a; Chen et al., 2013), its performance relies on the accuracy of reference (Liu et al., 2012a).

As noted above, PLS, CCA, and RRR are closely related. They all introduced the sparse version of algorithms – sparse PLS (Le Floch et al., 2012), sparse CCA (Boutte and Liu, 2010; Chi et al., 2013), and sparse RRR (Vounou et al., 2010, 2012) – when applied onto a large number of variables in imaging genetics. Not only does the increase of sparsity make the interpretation more plausible, but also strengthens the stability of results by avoiding the over-fitting problem. Le Floch et al. (2012) showed through simulation that different levels of regularization on sparsity may produce different results for CCA and PLS, and the two methods converge together with the corresponding regularization strength. Similarly, for RRR, sparsity affects the performance (Vounou et al., 2010), and how to choose sparsity is critical in real applications. Up to now, only sparse PLS (combined with a filtering step) and sparse RRR have been applied to real imaging genetic data with larger than 100k loci (Le Floch et al., 2012; Vounou et al., 2012).

The differences among these methods besides mathematical models listed above also include settings in practice. First, the number of latent variables (components or ranks) to test is chosen differently. CCA, PLS, and RRR extract same numbers of components for genetic and imaging data, and pair-wise connections are tested. Though guidance is discussed for the choice of component number, users of these methods tend to be very conservative. Silver et al. (2012) only investigated the components from first rank in their RRR application, and Vounou et al. (2010, 2012) investigated the top three ranks. In the application of CCA, Hardoon et al. (2009) tested the top pairs of components, and Le Floch et al. (2012) examined the first two pairs of components for both CCA and PLS methods. In contrast, parallel ICA, following the principle of Infomax ICA (Bell and Sejnowski, 1995; Cardoso, 1999), first estimates the number of components embedded in genetic and imaging data. Estimation is either based on information theory (Akaike, 1974; Li et al., 2007) or stability (Chen et al., 2012a), with the goal of reliably, maximally explaining the variance of data. The number of components for genetic and imaging data can be different, and the pairs of related components between the two modalities are driven by data. Sometimes pair-wise correlations are not necessary (Meda et al., 2012). Judging from this aspect, parallel ICA carries advantage of exploring more possible connections between the two modalities, while other methods target only the top correlated components.

Second, all methods are limited in handling a large number of variables (particularly SNP loci). CCA, PLS, and RRR methods may run into over-fitting problems, where cross evaluation performance drops (Le Floch et al., 2012). Parallel ICA fails to identify the connections between modalities (Liu et al., 2008a). The ways to overcome this limitation are also different. Pre-filtering SNP loci to reduce the dimensionality is successfully implemented for CCA and PLS. Le Floch et al. (2012) presented a comprehensive comparison of PLS and CCA combined with different filtering methods. They showed that incorporating a filtering step before the multivariate association test (with the goal of removing irrelevant SNPs) can improve the performance for both methods. Their real data application makes clear that the dimension reduction (which reduced 700k SNPs down to 1000 SNPs) is an important step for avoiding over-fitting with such large genetic data. Although various means can be used to pre-filter SNPs, we recommend leveraging large population genetic data as a reference, such as Psychiatric Genomics Consortium[9]. For RRR, enhancing the sparsity to select only a small number of SNPs is an effective way to increase stability. Yet, the choice of sparsity is not easy (Vounou et al., 2010). N-fold cross evaluation can be used to decide the best parameter. Vounou et al. (2012) chose to test a range of sparsity settings and select resultant SNPs with high probability. Parallel ICA leverages prior information (a referential SNP set) to increase chances of extracting relevant genetic components associated with imaging phenotypes from large SNP data. The difficulty with this approach lies in how to decide the reference. In particular, what we should do when we do not have any prior knowledge about genetics regarding a particular phenotype? While prior information helps interpret the genetic result in a degree, parallel ICA need to threshold the resultant latent variable to select the most weighted SNPs, since no sparsity is in place (Chen et al., 2013).

Third, verification of results from latent variables is very important to guard against false discoveries. N-fold cross evaluation has been utilized for CCA and PLS, and sub-sampling is used in RRR, not exactly verification but increasing the stability (Silver et al., 2012; Vounou et al., 2012). Permutation and leave-one-out evaluation are used in parallel ICA (Liu et al., 2009b; Chen et al., 2012b). We strongly recommend future users to incorporate certain verification steps in their studies, given the complexity of the methods mentioned. To date, only parallel ICA has a ready-to-use package available[10].

[8] http://mialab.mrn.org/software/fit

[9] https://pgc.unc.edu/

[10] http://mialab.mrn.org/software/fit/

Except for multivariate analyses based on latent variables, methods in machine learning category, i.e., training algorithms with known knowledge and using them to predict the unseen data, have also been applied to imaging genetics. For instance, support vector machine on ICA factors of genetic and fMRI data together achieved better separation of schizophrenia patients from controls than using either type of data alone, suggesting that genetic and brain functions capture different, but partially complementary schizophrenic features (Yang et al., 2010). Within the same line, Wang et al. (2012b) proposed a multimodal multitask learning algorithm that combines genetic and multimodal imaging features to predict simultaneously diagnoses and cognitive function. In this algorithm, classification and regression are performed jointly, and a group L1-norm regularization is used for feature selection to integrate heterogeneous imaging genetic data. One of strengths of this approach is that genetic markers and imaging biomarkers relevant for both diagnosis and cognitive function are identified. Another new application of learning algorithms in imaging genetics is random forest on distance matrices, where by employing distance measures between input variables, various interactions (away from original space) are modeled and random forest search is used for selection of best sets of features (Sim et al., 2013). While it provides promising results, the requirement for intensive computation and sophisticated modeling may hinder further applications, which is true for other methods too.

## CHALLENGE AND FUTURE DEVELOPMENTS

During the last decade, imaging genetics has rapidly developed into a promising, high impact research field and extended into a body of studies on mental disorders, including both human and animal studies. As Meyer-Lindenberg (2012) stated, future imaging genetic studies have to confront the complexity of epistasis, pleiotropy and gene-by-environment interactions, and this issue will become even more pressing as the field moves into whole genome sequencing. Although methods reviewed here attempt to tackle this complex problem, limitations are clear. For example, none of the methods can really address the genome-wide whole brain association without filtering or dimension reduction. Some multivariate methods such as MDR and prior knowledge guided approaches have not been fully incorporated into imaging genetics yet. Methods of CCA, PLS and RRR, facing over-fitting issues when handling large genetic variables, may be improved by leveraging prior information. Methods of parallel ICA may need to enhance sparsity within the independent genetic components. Such limitation in fact relates to a common problem across multivariate analyses, which is the difficulty in interpreting results (i.e., results are lack of direct biological meaning). For instance, GSEA does not model the exact interaction among SNPs. The latent component does not necessarily hold direct biological reason why multiple genetic variants form into one factor, or why hundreds of voxels group into one brain network. One way to alleviate this problem is to incorporate additional information, such as known biological information, cellular level information, or behavioral specific information, into analyses. Further developing current methods and integrating more information will continue to be an important research frontier.

As matter of fact, another pressing demand raised by Meyer-Lindenberg (2012) in the future of imaging genetics is to integrate various types of data relevant to imaging genetics, beyond just two modalities. The new data can be proteomic, gene expression, epigenetic, behavioral and environmental variables. Studies have shown their relevance to brain structural and functional changes, genetic mutations, and psychiatric disorders (Clark et al., 2006; Serretti et al., 2007; Maric and Svrakic, 2012; Liu et al., 2013). The relationship among these data is by no means simple and pairwise. To date, very few methods have been applied in imagine genetics to tackle the relation beyond two modalities (expect for *post hoc* analyses with behavior or diagnosis). It is very promising to see that some studies have stepped into this direction, though only for multimodal imaging data (Correa et al., 2010; Sui et al., 2012b). How to integrate such data in a systemic way with embeded biological hierarchy is still an untouched land. Methods and models incorporating multiple levels of biological variables (here including behavioral or environmental variables) into broader imaging genetics are another research direction of great potential and impact.

To date, very few studies focused on CNV's effect on brain-based phenotypes (Yeo et al., 2011; Boutte et al., 2012; Liu et al., 2012b), even though many studies have identified a relationship between CNVs with psychiatric disorders (McCarroll and Altshuler, 2007; Bassett et al., 2008; Guilmatre et al., 2009). Meyer-Lindenberg (2012) has indicated that the future of imaging genetics will recognize the importance of the sizeable amount of variation in CNVs. Given the low incidence of individual CNVs, in particular large and rare CNVs, such studies are more likely from multi-site collaborations, where increasing numbers of imaging genetic studies are heading for (Schumann et al., 2010; Thompson et al., 2014). Methods to encompass data from multi-sites, controlling for not only different equipments or experiments but also different local populations or environments, are in great need, which have to consider both computational feasibility and mathematical (model) validity.

Given that the future focus of imaging genetics is expected to be multi-site, large scale, genome-wide whole brain, multiple level association studies, we believe that more effort should be focused on the development of methods that can confront these challenges.

## ACKNOWLEDGMENTS

This work was supported by NIH, grant: NIBIB R01EB005846 and P20RR021938 to Vince D. Calhoun, and grant: R33DA027626 to Jingyu Liu.

## REFERENCES

Abatangelo, L., Maglietta, R., Distaso, A., D'Addabbo, A., Creanza, T. M., Mukherjee, S., et al. (2009). Comparative study of gene set enrichment methods. *BMC Bioinformatics* 10:275. doi: 10.1186/1471-2105-10-275

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* 19, 716–723.

Bach, F., and Michael, J. (2002). Kernel independent component analysis. *J. Mach. Learn. Res.* 3, 1–48.

Bassett, A. S., Marshall, C. R., Lionel, A. C., Chow, E. W., and Scherer, S. W. (2008). Copy number variations and risk for schizophrenia in 22q11.2 deletion syndrome. *Hum. Mol. Genet.* 17, 4045–4053. doi: 10.1093/hmg/ddn307

Beckmann, C. F., and Smith, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans. Med. Imaging* 23, 137–152. doi: 10.1109/TMI.2003.822821

Beckmann, C. F., and Smith, S. M. (2005). Tensorial extensions of independent component analysis for multisubject FMRI analysis. *Neuroimage* 25, 294–311. doi: 10.1016/j.neuroimage.2004.10.043

Bell, A. J., and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* 7, 1129–1159. doi: 10.1162/neco.1995.7.6.1129

Biessmann, F., Plis, S., Meinecke, F. C., Eichele, T., and Muller, K. R. (2011). Analysis of multimodal neuroimaging data. *IEEE Rev. Biomed. Eng.* 4, 26–58. doi: 10.1109/RBME.2011.2170675

Bigos, K. L., and Weinberger, D. R. (2010). Imaging genetics – days of future past. *Neuroimage* 53, 804–809. doi: 10.1016/j.neuroimage.2010.01.035

Bingham, E., and Hyvarinen, A. (2000). A fast fixed-point algorithm for independent component analysis of complex valued signals. *Int. J. Neural Syst.* 10, 1–8. doi: 10.1142/S0129065700000028

Bookheimer, S. Y., Strojwas, M. H., Cohen, M. S., Saunders, A. M., Pericak-Vance, M. A., Mazziotta, J. C., et al. (2000). Patterns of brain activation in people at risk for Alzheimer's disease. *N. Engl. J. Med.* 343, 450–456. doi: 10.1056/NEJM200008173430701

Boutte, D., Calhoun, V. D., Chen, J., Sabbineni, A., Hutchison, K., and Liu, J. (2012). Association of genetic copy number variations at 11 q14.2 with brain regional volume differences in an alcohol use disorder population. *Alcohol* 46, 519–527. doi: 10.1016/j.alcohol.2012.05.002

Boutte, D., and Liu, J. (2010). "Sparse canonical correlation analysis applied to fMRI and genetic data fusion," in *2010 IEEE International Conference on Bioinformatics and Biomedicine*, Hong Kong, 422–426. doi: 10.1109/BIBM.2010.5706603

Calhoun, V. D., and Adali, T. (2006). Unmixing fMRI with independent component analysis. *IEEE Eng. Med. Biol. Mag.* 25, 79–90. doi: 10.1109/MEMB.2006.1607672

Calhoun, V. D., and Adali, T. (2012). Multisubject independent component analysis of fMRI: a decade of intrinsic networks, default mode, and neurodiagnostic discovery. *IEEE Rev. Biomed. Eng.* 5, 60–73. doi: 10.1109/RBME.2012.2211076

Calhoun, V. D., Adali, T., Pearlson, G. D., and Pekar, J. J. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Hum. Brain Mapp.* 14, 140–151. doi: 10.1002/hbm.1048

Calhoun, V. D., Adali, T., Stevens, M. C., Kiehl, K. A., and Pekar, J. J. (2005). Semi-blind ICA of fMRI: a method for utilizing hypothesis-derived time courses in a spatial ICA analysis. *Neuroimage* 25, 527–538. doi: 10.1016/j.neuroimage.2004.12.012

Calhoun, V. D., Liu, J., and Adali, T. (2009). A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *Neuroimage* 45, S163–S172. doi: 10.1016/j.neuroimage.2008.10.057

Cardoso, J. F. (1997). Infomax and maximum likelihood for blind source separation. *IEEE Signal Process. Lett.* 4, 112–114. doi: 10.1109/97.566704

Cardoso, J. F. (1999). High-order contrasts for independent component analysis. *Neural Comput.* 11, 157–192. doi: 10.1162/089976699300016863

Cattaert, T., Calle, M. L., Dudek, S. M., Mahachie John, J. M., Van Lishout, F., Urrea, V., et al. (2011). Model-based multifactor dimensionality reduction for detecting epistasis in case-control data in the presence of noise. *Ann. Hum. Genet.* 75, 78–89. doi: 10.1111/j.1469-1809.2010.00604.x

Chen, J., Calhoun, V. D., and Liu, J. (2012a). ICA order selection based on consistency: application to genotype data. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2012, 360–363. doi: 10.1109/EMBC.2012.6345943

Chen, J., Calhoun, V. D., Pearlson, G. D., Ehrlich, S., Turner, J. A., Ho, B. C., et al. (2012b). Multifaceted genomic risk for brain function in schizophrenia. *Neuroimage* 61, 866–875. doi: 10.1016/j.neuroimage.2012.03.022

Chen, J., Calhoun, V. D., Pearlson, G. D., Perrone-Bizzozero, N., Sui, J., Turner, J. A., et al. (2013). Guided exploration of genomic risk for gray matter abnormalities in schizophrenia using parallel independent component analysis with reference. *Neuroimage* 83, 384–396. doi: 10.1016/j.neuroimage.2013.05.073

Chi, E. C., Allen, G. I., Zhou, H., Kohannim, O., Lange, K., and Thompson, P. M. (2013). "Imaging genetics via sparse canonical correlation analysis," in *Biomedical Imaging (ISBI), IEEE 10th International Symposium*, San Francisco, CA. doi: 10.1109/ISBI.2013.6556581

Clark, D., Dedova, I., Cordwell, S., and Matsumoto, I. (2006). A proteome analysis of the anterior cingulate cortex gray matter in schizophrenia. *Mol. Psychiatry* 11, 459–470. doi: 10.1038/sj.mp.4001806

Correa, N. M., Adali, T., Li, Y. O., and Calhoun, V. D. (2010). Canonical correlation analysis for data fusion and group inferences: examining applications of medical imaging data. *IEEE Signal Process. Mag.* 27, 39–50. doi: 10.1109/MSP.2010.936725

Cule, E., Vineis, P., and De Iorio, M. (2011). Significance testing in ridge regression for genetic data. *BMC Bioinformatics* 12:372. doi: 10.1186/1471-2105-12-372

Dea, J. T., Anderson, M., Allen, E., Calhoun, V. D., and Adali, T. (2011). "IVA for multi-subject FMRI analysis: a comparative study using a new simulation toolbox," in *Machine Learning for Signal Processing, IEEE International Workshop*, Beijing, China, 1–6.

Demirci, O., Clark, V. P., and Calhoun, V. D. (2008). A projection pursuit algorithm to classify individuals using fMRI data: application to schizophrenia. *Neuroimage* 39, 1774–1782. doi: 10.1016/j.neuroimage.2007.10.012

D'haeseleer, P. (2005). How does gene expression clustering work? *Nat. Biotechnol.* 23, 1499–1501. doi: 10.1038/nbt1205-1499

Dimitriadou, E., Barth, M., Windischberger, C., Hornik, K., and Moser, E. (2004). A quantitative comparison of functional MRI cluster analysis. *Artif. Intell. Med.* 31, 57–71. doi: 10.1016/j.artmed.2004.01.010

Egan, M. F., Goldberg, T. E., Kolachana, B. S., Callicott, J. H., Mazzanti, C. M., Straub, R. E., et al. (2001). Effect of COMT Val108/158 Met genotype on frontal lobe function and risk for schizophrenia. *Proc. Natl. Acad. Sci. U.S.A.* 98, 6917–6922. doi: 10.1073/pnas.111134598

Engreitz, J. M., Daigle, B. J. Jr., Marshall, J. J., and Altman, R. B. (2010). Independent component analysis: mining microarray data for fundamental human gene expression modules. *J. Biomed. Inform.* 43, 932–944. doi: 10.1016/j.jbi.2010.07.001

Erhardt, E. B., Rachakonda, S., Bedrick, E. J., Allen, E. A., Adali, T., and Calhoun, V. D. (2011). Comparison of multi-subject ICA methods for analysis of fMRI data. *Hum. Brain Mapp.* 32, 2075–2095. doi: 10.1002/hbm.21170

Esposito, F., Scarabino, T., Hyvarinen, A., Himberg, J., Formisano, E., Comani, S., et al. (2005). Independent component analysis of fMRI group studies by self-organizing clustering. *Neuroimage* 25, 193–205. doi: 10.1016/j.neuroimage.2004.10.042

Filipovych, R., and Davatzikos, C. (2011). Semi-supervised pattern classification of medical images: application to mild cognitive impairment (MCI). *Neuroimage* 55, 1109–1119. doi: 10.1016/j.neuroimage.2010.12.066

Glahn, D. C., Thompson, P. M., and Blangero, J. (2007). Neuroimaging endophenotypes: strategies for finding genes influencing brain structure and function. *Hum. Brain Mapp.* 28, 488–501. doi: 10.1002/hbm.20401

Gottesman, I. I., and Gould, T. D. (2003). The endophenotype concept in psychiatry: etymology and strategic intentions. *Am. J. Psychiatry* 160, 636–645. doi: 10.1176/appi.ajp.160.4.636

Gui, J., Andrew, A. S., Andrews, P., Nelson, H. M., Kelsey, K. T., Karagas, M. R., et al. (2011). A robust multifactor dimensionality reduction method for detecting gene-gene interactions with application to the genetic analysis of bladder cancer susceptibility. *Ann. Hum. Genet.* 75, 20–28. doi: 10.1111/j.1469-1809.2010.00624.x

Gui, J., Moore, J. H., Williams, S. M., Andrews, P., Hillege, H. L., Van Der Harst, P., et al. (2013). A simple and computationally efficient approach to multifactor dimensionality reduction analysis of gene-gene interactions for quantitative traits. *PLoS ONE* 8:e66545. doi: 10.1371/journal.pone.0066545

Guilmatre, A., Dubourg, C., Mosca, A. L., Legallic, S., Goldenberg, A., Drouin-Garraud, V., et al. (2009). Recurrent rearrangements in synaptic and neurodevelopmental genes and shared biologic pathways in schizophrenia, autism, and mental retardation. *Arch. Gen. Psychiatry* 66, 947–956. doi: 10.1001/archgenpsychiatry.2009.80

Hardoon, D. R., Ettinger, U., Mourao-Miranda, J., Antonova, E., Collier, D., Kumari, V., et al. (2009). Correlation-based multivariate analysis of genetic influence on brain volume. *Neurosci. Lett.* 450, 281–286. doi: 10.1016/j.neulet.2008.11.035

Hariri, A. R., Drabant, E. M., and Weinberger, D. R. (2006). Imaging genetics: perspectives from studies of genetically driven variation in serotonin function and corticolimbic affective processing. *Biol. Psychiatry* 59, 888–897. doi: 10.1016/j.biopsych.2005.11.005

Heinz, A., Goldman, D., Jones, D. W., Palmour, R., Hommer, D., Gorey, J. G., et al. (2000). Genotype influences *in vivo* dopamine transporter availability in human striatum. *Neuropsychopharmacology* 22, 133–139. doi: 10.1016/S0893-133X(99)00099-8

Hinrichs, C., Singh, V., Mukherjee, L., Xu, G., Chung, M. K., and Johnson, S. C. (2009). Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset. *Neuroimage* 48, 138–149. doi: 10.1016/j.neuroimage.2009.05.056

Holden, M., Deng, S., Wojnowski, L., and Kulle, B. (2008). GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics* 24, 2784–2785. doi: 10.1093/bioinformatics/btn516

Hong, B., Pearlson, G. D., and Calhoun, V. D. (2005). Source density-driven independent component analysis approach for fMRI data. *Hum. Brain Mapp.* 25, 297–307. doi: 10.1002/hbm.20100

Hu, T., Sinnott-Armstrong, N. A., Kiralis, J. W., Andrew, A. S., Karagas, M. R., and Moore, J. H. (2011). Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinformatics* 12:364. doi: 10.1186/1471-2105-12-364

Huang, D. S., and Zheng, C. H. (2006). Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics* 22, 1855–1862. doi: 10.1093/bioinformatics/btl190

Hyvirinen, A., and Oja, E. (1999). A fast fixed-point algorithm for independent component analysis. *Neural Comput.* 9, 1483–1492. doi: 10.1162/neco.1997.9.7.1483

Jagannathan, K., Calhoun, V. D., Gelernter, J., Stevens, M. C., Liu, J., Bolognani, F., et al. (2010). Genetic associations of brain structural networks in schizophrenia: a preliminary study. *Biol. Psychiatry* 68, 657–666. doi: 10.1016/j.biopsych.2010.06.002

Jiang, D., Tang, C., and Zahng, A. (2004). Cluster analysis for gene expression data: a survey. *IEEE Trans. Knowl. Data Eng.* 16, 1370–1386. doi: 10.1109/TKDE.2004.68

Jombart, T., Pontier, D., and Dufour, A. B. (2009). Genetic markers in the playground of multivariate analysis. *Heredity* 102, 330–341. doi: 10.1038/hdy.2008.130

Kim, T., Lee, I., and Lee, T.-W. (2006). "Independent vector analysis: definition and algorithms," in *Signals, Systems and Computers, ACSSC '06. Fortieth Asilomar Conference on*, Pacific Grove, CA, 1393–1396. doi: 10.1109/ACSSC.2006.354986

Kong, W., Vanderburg, C. R., Gunshin, H., Rogers, J. T., and Huang, X. (2008). A review of independent component analysis application to microarray gene expression data. *Biotechniques* 45, 501–520. doi: 10.2144/000112950

Lee, J. H., Lee, T. W., Jolesz, F. A., and Yoo, S. S. (2008a). Independent vector analysis (IVA) for group fMRI processing of subcortical area. *Int. J. Imaging Syst. Tech.* 18, 29–41. doi: 10.1002/ima.20141

Lee, J. H., Lee, T. W., Jolesz, F. A., and Yoo, S. S. (2008b). Independent vector analysis (IVA): multivariate approach for fMRI group study. *Neuroimage* 40, 86–109. doi: 10.1016/j.neuroimage.2007.11.019

Le Floch, E., Guillemot, V., Frouin, V., Pinel, P., Lalanne, C., Trinchera, L., et al. (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, 11–24. doi: 10.1016/j.neuroimage.2012.06.061

Li, M. X., Gui, H. S., Kwan, J. S., and Sham, P. C. (2011). GATES: a rapid and powerful gene-based association test using extended simes procedure. *Am. J. Hum. Genet.* 88, 283–293. doi: 10.1016/j.ajhg.2011.01.019

Li, Y. O., Adali, T., and Calhoun, V. D. (2007). Estimating the number of independent components for functional magnetic resonance imaging data. *Hum. Brain Mapp.* 28, 1251–1266. doi: 10.1002/hbm.20359

Liu, J., Bixler, J. N., and Calhoun, V. D. (2008a). "A multimodality ICA study – integrating genomic single nucleotide polymorphisms with functional neuroimaging data," in *Bioinformatics and Biomedicine Workshops, 2008. BIBMW 2008* (IEEE International Conference on), Philadelphia, PA, 151–157. doi: 10.1109/BIBMW.2008.4686229

Liu, J., and Calhoun, V. (2007). "Parallel independent component analysis for multimodel analysis: application to fMRI and EEG data," *in Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on*, Washington, DC, 1028–1031. doi: 10.1109/ISBI.2007.357030

Liu, J., Chen, J., Ehrlich, S., Walton, E., White, T., Perrone-Bizzozero, N., et al. (2013). Methylation patterns in whole blood correlate with symptoms in schizophrenia patients. *Schizophr. Bull.* doi: 10.1093/schbul/sbt080 [Epub ahead of print].

Liu, J., Demirci, O., and Calhoun, V. D. (2008b). A parallel independent component analysis approach to investigate genomic influence on brain function. *IEEE Signal Process. Lett.* 15, 413–416. doi: 10.1109/LSP.2008.922513

Liu, J., Ghassemi, M. M., Michael, A. M., Boutte, D., Wells, W., Perrone-Bizzozero, N., et al. (2012a). An ICA with reference approach in identification of genetic variation and associated brain networks. *Front. Hum. Neurosci.* 6:21. doi: 10.3389/fnhum.2012.00021

Liu, J., Hutchison, K., Perrone-Bizzozero, N., Morgan, M., Sui, J., and Calhoun, V. (2010a). Identification of genetic and epigenetic marks involved in population structure. *PLoS ONE* 5:e13209. doi: 10.1371/journal.pone.0013209

Liu, J., Kiehl, K. A., Pearlson, G., Perrone-Bizzozero, N. I., Eichele, T., and Calhoun, V. D. (2009a). Genetic determinants of target and novelty-related event-related potentials in the auditory oddball response. *Neuroimage* 46, 809–816. doi: 10.1016/j.neuroimage.2009.02.045

Liu, J., Pearlson, G., Windemuth, A., Ruano, G., Perrone-Bizzozero, N. I., and Calhoun, V. (2009b). Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. *Hum. Brain Mapp.* 30, 241–255. doi: 10.1002/hbm.20508

Liu, J., Ulloa, A., Perrone-Bizzozero, N., Yeo, R., Chen, J., and Calhoun, V. D. (2012b). A pilot study on collective effects of 22q13.31 deletions on gray matter concentration in schizophrenia. *PLoS ONE* 7:e52865. doi: 10.1371/journal.pone.0052865

Lin, J. A., Zhu, H., Knickmeyer, R., Styner, M., Gilmore, J., and Ibrahim, J. G. (2012). Projection regression models for multivariate imaging phenotype. *Genet. Epidemiol.* 36, 631–641. doi: 10.1002/gepi.21658

Liu, J. Z., Mcrae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., et al. (2010b). A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* 87, 139–145. doi: 10.1016/j.ajhg.2010.06.009

Lin, Q. H., Liu, J., Zheng, Y. R., Liang, H., and Calhoun, V. D. (2010). Semiblind spatial ICA of fMRI using spatial constraints. *Hum. Brain Mapp.* 31, 1076–1088. doi: 10.1002/hbm.20919

Lou, X. Y., Chen, G. B., Yan, L., Ma, J. Z., Zhu, J., Elston, R. C., et al. (2007). A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *Am. J. Hum. Genet.* 80, 1125–1137. doi: 10.1086/518312

Ma, D. Q., Whitehead, P. L., Menold, M. M., Martin, E. R., Ashley-Koch, A. E., Mei, H., et al. (2005). Identification of significant association and gene-gene interaction of GABA receptor subunit genes in autism. *Am. J. Hum. Genet.* 77, 377–388. doi: 10.1086/433195

Maric, N. P., and Svrakic, D. M. (2012). Why schizophrenia genetics needs epigenetics: a review. *Psychiatr. Danub.* 24, 2–18.

McCarroll, S. A., and Altshuler, D. M. (2007). Copy-number variation and association studies of human disease. *Nat. Genet.* 39, S37–S42. doi: 10.1038/ng2080

McKeown, M. J., Hansen, L. K., and Sejnowsk, T. J. (2003). Independent component analysis of functional MRI: what is signal and what is noise? *Curr. Opin. Neurobiol.* 13, 620–629. doi: 10.1016/j.conb.2003.09.012

Meda, S. A., Jagannathan, K., Gelernter, J., Calhoun, V. D., Liu, J., Stevens, M. C., et al. (2010). A pilot multivariate parallel ICA study to investigate differential linkage between neural networks and genetic profiles in schizophrenia. *Neuroimage* 53, 1007–1015. doi: 10.1016/j.neuroimage.2009.11.052

Meda, S. A., Narayanan, B., Liu, J., Perrone-Bizzozero, N. I., Stevens, M. C., Calhoun, V. D., et al. (2012). A large scale multivariate parallel ICA method reveals novel imaging-genetic relationships for Alzheimer's disease in the ADNI cohort. *Neuroimage* 60, 1608–1621. doi: 10.1016/j.neuroimage.2011.12.076

Meier, T. B., Wildenberg, J. C., Liu, J., Chen, J., Calhoun, V. D., Biswal, B. B., et al. (2012). Parallel ICA identifies sub-components of resting state networks that covary with behavioral indices. *Front. Hum. Neurosci.* 6:281. doi: 10.3389/fnhum.2012.00281

Meyer-Lindenberg, A. (2010). Imaging genetics of schizophrenia. *Dialogues Clin. Neurosci.* 12, 449–456.

Meyer-Lindenberg, A. (2012). The future of fMRI and genetics research. *Neuroimage* 62, 1286–1292. doi: 10.1016/j.neuroimage.2011.10.063

Meyer-Lindenberg, A., Nicodemus, K. K., Egan, M. F., Callicott, J. H., Mattay, V., and Weinberger, D. R. (2008). False positives in imaging genetics. *Neuroimage* 40, 655–661. doi: 10.1016/j.neuroimage.2007.11.058

Meyer-Lindenberg, A., and Weinberger, D. R. (2006). Intermediate phenotypes and genetic mechanisms of psychiatric disorders. *Nat. Rev. Neurosci.* 7, 818–827. doi: 10.1038/nrn1993

Mirina, A., Atzmon, G., Ye, K., and Bergman, A. (2012). Gene size matters. *PLoS ONE* 7:e49093. doi: 10.1371/journal.pone.0049093

Moore, J. H., Asselbergs, F. W., and Williams, S. M. (2010). Bioinformatics challenges for genome-wide association studies. *Bioinformatics* 26, 445–455. doi: 10.1093/bioinformatics/btp713

Moore, J. H., and Williams, S. M. (2002). New strategies for identifying gene-gene interactions in hypertension. *Ann. Med.* 34, 88–95. doi: 10.1080/07853890252953473

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., et al. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* 34, 267–273. doi: 10.1038/ng1180

Oh, S., Lee, J., Kwon, M. S., Weir, B., Ha, K., and Park, T. (2012). A novel method to identify high order gene-gene interactions in genome-wide association studies: gene-based MDR. *BMC Bioinformatics* 13(Suppl. 9):S5. doi: 10.1186/1471-2105-13-S9-S5

Oh, S. J., Ahn, J. Y., and Chung, D. H. (2011). Comparison of invariant NKT cells with conventional T cells by using gene set enrichment analysis (GSEA). *Immune Netw.* 11, 406–411. doi: 10.4110/in.2011.11.6.406

Pan, Q., Hu, T., and Moore, J. H. (2013). Epistasis, complexity, and multifactor dimensionality reduction. *Methods Mol. Biol.* 1019, 465–477. doi: 10.1007/978-1-62703-447-0-22

Papassotiropoulos, A., and de Quervain, D. J. (2011). Genetics of human episodic memory: dealing with complexity. *Trends Cogn. Sci.* 15, 381–387. doi: 10.1016/j.tics.2011.07.005

Pattin, K. A., White, B. C., Barney, N., Gui, J., Nelson, H. H., Kelsey, K. T., et al. (2009). A computationally efficient hypothesis testing method for epistasis analysis using multifactor dimensionality reduction. *Genet. Epidemiol.* 33, 87–94. doi: 10.1002/gepi.20360

Potkin, S. G., Guffanti, G., Lakatos, A., Turner, J. A., Kruggel, F., Fallon, J. H., et al. (2009). Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease. *PLoS ONE* 4:e6501. doi: 10.1371/journal.pone.0006501

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909. doi: 10.1038/ng1847

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'donovan, M. C., Sullivan, P. F., et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460, 748–752. doi: 10.1038/nature08185

Rasetti, R., and Weinberger, D. R. (2011). Intermediate phenotypes in psychiatric disorders. *Curr. Opin. Genet. Dev.* 21, 340–348. doi: 10.1016/j.gde.2011.02.003

Ripke, S., Sanders, A. R., Kendler, K. S., Levinson, D. F., Sklar, P., Holmans, P. A., et al. (2011). Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* 43, 969–976. doi: 10.1038/ng.940

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., et al. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.* 69, 138–147. doi: 10.1086/321276

Roffman, J. L., Weiss, A. P., Goff, D. C., Rauch, S. L., and 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, 78–91. doi: 10.1080/10673220600642945

Rose, E. J., and Donohoe, G. (2013). Brain vs behavior: an effect size comparison of neuroimaging and cognitive studies of genetic risk for schizophrenia. *Schizophr. Bull.* 39, 518–526. doi: 10.1093/schbul/sbs056

Saidi, S. A., Holland, C. M., Kreil, D. P., Mackay, D. J., Charnock-Jones, D. S., Print, C. G., et al. (2004). Independent component analysis of microarray data in the study of endometrial cancer. *Oncogene* 23, 6677–6683. doi: 10.1038/sj.onc.1207562

Schmithorst, V. J., and Holland, S. K. (2004). Comparison of three methods for generating group statistical inferences from independent component analysis of functional magnetic resonance imaging data. *J. Magn. Reson. Imaging* 19, 365–368. doi: 10.1002/jmri.20009

Schumann, G., Loth, E., Banaschewski, T., Barbot, A., Barker, G., Buchel, C., et al. (2010). The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. *Mol. Psychiatry* 15, 1128–1139. doi: 10.1038/mp.2010.4

Serretti, A., Olgiati, P., and De Ronchi, D. (2007). Genetics of Alzheimer's disease. a rapidly evolving field. *J. Alzheimers Dis.* 12, 73–92.

Sheng, J., Deng, H. W., Calhoun, V. D., and Wang, Y. P. (2011). Integrated analysis of gene expression and copy number data on gene shaving using independent component analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 8, 1568–1579. doi: 10.1109/TCBB.2011.71

Silver, M., Janousova, E., Hua, X., Thompson, P. M., and Montana, G. (2012). Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression. *Neuroimage* 63, 1681–1694. doi: 10.1016/j.neuroimage.2012.08.002

Silver, M., Montana, G., and Nichols, T. E. (2011). False positives in neuroimaging genetics using voxel-based morphometry data. *Neuroimage* 54, 992–1000. doi: 10.1016/j.neuroimage.2010.08.049

Sim, A., Tsagkrasoulis, D., and Montana, G. (2013). Random forests on distance matrices for imaging genetics studies. *Stat. Appl. Genet. Mol. Biol.* 12, 757–786. doi: 10.1515/sagmb-2013-0040

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J. G., Zgaga, L., Manolio, T., et al. (2011). Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.* 89, 607–618. doi: 10.1016/j.ajhg.2011.10.004

Sloan, C. D., Shen, L., West, J. D., Wishart, H. A., Flashman, L. A., Rabin, L. A., et al. (2010). Genetic pathway-based hierarchical clustering analysis of older adults with cognitive complaints and amnestic mild cognitive impairment using clinical and neuroimaging phenotypes. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 153B, 1060–1069. doi: 10.1002/ajmg.b.31078

Small, G. W., Ercoli, L. M., Silverman, D. H., Huang, S. C., Komo, S., Bookheimer, S. Y., et al. (2000). Cerebral metabolic and cognitive decline in persons at genetic risk for Alzheimer's disease. *Proc. Natl. Acad. Sci. U.S.A.* 97, 6037–6042. doi: 10.1073/pnas.090106797

Smoller, J. W., Craddock, N., Kendler, K., Lee, P. H., Neale, B. M., Nurnberger, J. I., et al. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* 381, 1371–1379. doi: 10.1016/S0140-6736(12)62129-1

Stein, J. L., Hua, X., Lee, S., Ho, A. J., Leow, A. D., Toga, A. W., et al. (2010). Voxelwise genome-wide association study (vGWAS). *Neuroimage* 53, 1160–1174. doi: 10.1016/j.neuroimage.2010.02.032

Suarez-Farinas, M., Lowes, M. A., Zaba, L. C., and Krueger, J. G. (2010). Evaluation of the psoriasis transcriptome across different studies by gene set enrichment analysis (GSEA). *PLoS ONE* 5:e10247. doi: 10.1371/journal.pone.0010247

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Sui, J., Adali, T., Yu, Q., Chen, J., and Calhoun, V. D. (2012a). A review of multivariate methods for multimodal fusion of brain imaging data. *J. Neurosci. Methods* 204, 68–81. doi: 10.1016/j.jneumeth.2011.10.031

Sui, J., He, H., Pearlson, G. D., Adali, T., Kiehl, K. A., Yu, Q., et al. (2012b). Three-way (N-way) fusion of brain imaging data based on mCCA + jICA and its application to discriminating schizophrenia. *Neuroimage* 66C, 119–132. doi: 10.1016/j.neuroimage.2012.10.051

Thompson, P. M., Stein, J. L., Medland, S. E., Hibar, D. P., Vasquez, A. A., Renteria, M. E., et al. (2014). The ENIGMA consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging Behav.* doi: 10.1007/s11682-013-9269-5 [Epub ahead of print].

Vounou, M., Janousova, E., Wolz, R., Stein, J. L., Thompson, P. M., Rueckert, D., et al. (2012). Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. *Neuroimage* 60, 700–716. doi: 10.1016/j.neuroimage.2011.12.029

Vounou, M., Nichols, T. E., and Montana, G. (2010). Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach. *Neuroimage* 53, 1147–1159. doi: 10.1016/j.neuroimage.2010.07.002

Walton, E., Turner, J., Gollub, R. L., Manoach, D. S., Yendiki, A., Ho, B. C., et al. (2013). Cumulative genetic risk and prefrontal activity in patients with schizophrenia. *Schizophr. Bull.* 39, 703–711. doi: 10.1093/schbul/sbr190

Wang, K., and Abbott, D. (2008). A principal components regression approach to multilocus genetic association studies. *Genet. Epidemiol.* 32, 108–118. doi: 10.1002/gepi.20266

Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S. L., et al. (2012a). Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. *Bioinformatics* 28, 229–237. doi: 10.1093/bioinformatics/btr649

Wang, H., Nie, F., Huang, H., Risacher, S. L., Saykin, A. J., and Shen, L. (2012b). Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics* 28, i127–i136. doi: 10.1093/bioinformatics/bts228

Weng, L., Macciardi, F., Subramanian, A., Guffanti, G., Potkin, S. G., Yu, Z., et al. (2011). SNP-based pathway enrichment analysis for genome-wide association studies. *BMC Bioinformatics* 12:99. doi: 10.1186/1471-2105-12-99

Whalley, H. C., Papmeyer, M., Sprooten, E., Romaniuk, L., Blackwood, D. H., Glahn, D. C., et al. (2012). The influence of polygenic risk for bipolar disorder on neural activation assessed using fMRI. *Trans. Psychiatry* 2, e130. doi: 10.1038/tp.2012.60

Winham, S. (2013). Applications of multifactor dimensionality reduction to genome-wide data using the R package "MDR." *Methods Mol. Biol.* 1019, 479–498. doi: 10.1007/978-1-62703-447-0-23

Wu, L., Eichele, T., and Calhoun, V. (2011). "Parallel independent component analysis using an optimized neurovascular coupling for concurrent EEG-fMRI sources," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Boston, Massachusetts, 2542–2545.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* 25, 714–721. doi: 10.1093/bioinformatics/btp041

Yang, H., Liu, J., Sui, J., Pearlson, G., and Calhoun, V. D. (2010). A hybrid machine learning method for fusing fMRI and genetic data: combining both improves classification of schizophrenia. *Front. Hum. Neurosci.* 4:192. doi: 10.3389/fnhum.2010.00192

Yeo, R. A., Gangestad, S. W., Gasparovic, C., Liu, J., Calhoun, V. D., Thoma, R. J., et al. (2011). Rare copy number deletions predict individual variation in human brain metabolite concentrations in individuals with alcohol use disorders. *Biol. Psychiatry.* 70, 537–544. doi: 10.1016/j.biopsych.2011.04.019

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

*Received: 31 October 2013; accepted: 04 March 2014; published online: 26 March 2014.*
*Citation: Liu J and Calhoun VD (2014) A review of multivariate analyses in imaging genetics. Front. Neuroinform.* ***8****:29. doi: 10.3389/fninf.2014.00029*
*This article was submitted to the journal Frontiers in Neuroinformatics.*