CHAPTER

Machine learning in brain imaging genomics

14

J. Yan[a], **L. Du, X. Yao, L. Shen**

Indiana University School of Medicine, Indianapolis, IN, United States

CHAPTER OUTLINE

14.1 INTRODUCTION

Brain imaging genomics has attracted increasing attention in recent years. It is an emerging research field that has arisen with the advances in high-throughput genotyping and multimodal imaging techniques. Its major task is to examine the association between genetic markers such as single nucleotide polymorphisms (SNPs) and quantitative traits (QTs) extracted from multimodal neuroimaging data (eg, anatomical, functional, and molecular imaging scans). Given the well-known importance of imaging and genomics in the brain study, bridging these two factors and exploring their connections have the potential to provide a better mechanistic understanding of normal or disordered brain functions. Also, changes in imaging phenotypes usually precede those in disease status and cognitive outcomes, and are believed to be closer to the underlying genetic mechanisms. Therefore associating

[a]Equal contributions by J. Yan, L. Du, and X. Yao. J. Yan contributed to Sections 14.1, 14.2.1, 14.2.2, and 14.4. L. Du contributed to Section 14.2.3. X. Yao contributed to Section 14.3.

Machine Learning and Medical Imaging. http://dx.doi.org/10.1016/B978-0-12-804076-8.00014-1

genetic data with imaging phenotypes, rather than disease status, is highly promising for discovery of influential genetic architecture and has the potential to help reveal the earliest brain changes for prognosis.

Early attempts in imaging genomics were mostly pairwise univariate analyses for quantitative genome-wide association studies (GWASs), which were performed to correlate high-throughput SNP data to large-scale image QT data (Shen et al., 2010; Stein et al., 2010). A simple regression model, for example, that implemented in PLINK (Purcell et al., 2007), was typically used to examine the additive effect of each single SNP on each single imaging QT. These simple regression models were often coupled with hypothesis testing, in which the significances of regression coefficients were learned simultaneously. Pairwise univariate analysis was used in traditional association studies to quickly provide important association information between SNPs and QTs. However, both SNPs and QTs were treated as independent and isolated units, and therefore the underlying correlating structure between the units was ignored.

Multiple regression models were later introduced to study the multilocus effects on imaging phenotypes. Instead of discovering individual SNPs with significant effects, multivariate models have enabled the identification of SNPs that jointly affect phenotypic changes. GCTA, a popular heritability analysis tool, utilizes a linear mixed regression model and has successfully demonstrated that ∼45% of the phenotypic variance for human height can be explained by the joint effect of common SNPs (Yang et al., 2011, 2013). Also considering the intercorrelated nature and high-dimensional setting of imaging and genetic data, sparse regression models are particularly favored. These models can not only address the correlation problem, but also help identify a small number of biologically meaningful genetic markers for easy interpretation. In addition, many recent efforts have been devoted to prior knowledge guided regression models, and many studies have confirmed the beneficial role of prior data structure in capturing more accurate imaging genomic relationships (Wang et al., 2012a; Silver et al., 2012a,b).

Bi-multivariate association analysis has recently received increasing attention for exploration of complex multi-SNP-multi-QT relationship. Existing bi-multivariate models widely used in imaging genomic studies can generally be classified into two types: canonical correlation analysis (CCA) type and reduced rank regression (RRR) type. Based on the assumption that a real imaging genomic signal typically involves a small number of SNPs and QTs, both CCA and RRR have their sparse versions. These sparse models are designed to better fit the imaging genomic study as they yield sparse patterns for easy interpretation. Example studies applying these methods in imaging genomic applications include Chi et al. (2013), Wan et al. (2011), and Vounou et al. (2012). Prior knowledge has also been examined in bi-multivariate models recently, and its beneficial role has also been reported in many studies (Yan et al., 2014a; Du et al., 2014; Lin et al., 2014).

Enrichment analysis has been widely studied in gene expression data analysis, and has recently been modified to analyze GWAS data to extract biological insights based on functional annotation and pathway databases

(Ramanan et al., 2012; Younesi and Hofmann-Apitius, 2013). Recently, it has been extended to the imaging genomics domain, to discover high-level associations based on prior knowledge, including meaningful gene sets (GSs) and brain circuits (BCs), which typically contain multiple genes and multiple QTs, respectively (Yao et al., 2015). By jointly considering the complex relationships between the interlinked genetic markers and correlated imaging phenotypes, imaging genomic enrichment analysis (IGEA) provides additional power for extracting biological insights on neurogenomic associations at a systems biology level.

In this chapter, we will summarize the widely used traditional and state-of-the-art statistical and machine learning methods, ranging from univariate (Section 14.2.1), multilocus (Section 14.2.2), bi-multivariate (Section 14.2.3) models to recent enrichment models (Section 14.3), and discuss their applications in brain imaging genomics.

14.2 MINING IMAGING GENOMIC ASSOCIATIONS VIA REGRESSION OR CORRELATION ANALYSIS

Genotype-phenotype associations are typically explored in three ways based on different assumptions: (1) one genotype affects one or more phenotypes independently, (2) multiple genotypes jointly affect one phenotype, and (3) multiple genotypes jointly affect a BC with multiple phenotypes. The first assumption is the simplest and is the most common practice for mining imaging genomic associations. Generally, traditional correlation or linear regression models are sufficient for the analysis, and mostly coupled with hypothesis testing. Regression models and bi-multivariate association models are appropriate methods for Assumptions 2 and 3, respectively, and have already been widely applied. In this section, we will discuss these statistical and machine learning models in detail and summarize their applications in the field of brain imaging genomics.

14.2.1 SINGLE-LOCUS ANALYSIS

Early brain imaging genomic studies mostly focused on exploring paired relationships between a single SNP and a single imaging QT. A simple linear regression model (Eq. 14.1), for example, that implemented in PLINK (Purcell et al., 2007), is usually used to examine the additive effect of each single SNP on each imaging QT. Example studies include Potkin et al. (2009) and Shen et al. (2010). Let X_{cov} indicate the covariates, such as age and gender, X_{s} be the genotype of the SNP to be examined, and y be the imaging QT as the response to be associated with the SNP. SNPs are typically coded as 0, 1, or 2, the number of minor alleles at a particular genome position.

$$y = X_{\text{cov}} w_{\text{cov}} + X_{\text{s}} w_{\text{s}} + \epsilon. \tag{14.1}$$

This additive model is generally applied for allelic association tests that only examine the association between each SNP and each imaging QT. The dominant, recessive, or their combination can also be investigated in the same way. A common practice of GWAS is to examine the allelic effects only, as it has the reasonable power to detect both dominant and recessive effects. But there are also plenty of studies examining all of them individually, the result of which will be subject to a further multiple comparison correction. In order to capture more phenotypic variance, the interaction effects between SNPs and environmental factors (ie, $X_s X_{env}$) or just between SNPs (ie, $X_{s1} X_{s2}$) have also been investigated in some studies. Another similar approach is the generalized linear model (GLM), a procedure that uses ANOVA, which considers allelic, dominant, and recessive effect measures as a categorical predictor variable.

Typically, linear regression is performed together with hypothesis testing, in which the regression coefficients and corresponding statistical value will be obtained simultaneously. For example, assuming the normal distribution of the error term $\boldsymbol{\epsilon}$, a t-statistic of w_s can be calculated through:

$$t = \frac{w_s}{\mathrm{SE}}, \tag{14.2}$$

where SE is the standard deviation of w_s, ie, $\mathrm{SE} = \frac{\sqrt{\frac{1}{n-2}\Sigma(y_i - \bar{y})^2}}{\sqrt{\Sigma(w_{si} - \overline{w_s})^2}}$. The p-value generated from the statistic evaluates the probability of observing an equal or greater statistic by chance. The smaller the p-value is, the more significant the genotype-phenotype relationship is considered to be. Fig. 14.1 shows an example of single-locus analysis across the entire genome.

14.2.1.1 Multiple comparison correction

In large univariate tests of all the pairwise SNP-QT associations, the p-value obtained from each single test is generally further corrected using various strategies. Bonferroni correction is the simplest one, which works by multiplying the p-value by the test number (ie, the number of SNPs $\times$ the number of QTs). However, this method hypothesizes that all experiments are independent and thus is considered to be overly conservative. It is more common in practical applications to combine it with the estimated number of independent tests (Bush and Moore, 2012; Yan et al., 2015b). Also, the false positive rate (FDR) introduced by Benjamini and Hochberg (1995) is also quite favored due to its less stringent threshold. Permutation, though potentially time-consuming, is another approach to acquire the significance with dependent genotypes, in which imaging QTs are randomly reassigned to other subjects to break the underlying SNP-QT relationships. By repeating this permutation procedure many times, it helps simulate a null distribution of test statistics for w_s. An empirical p-value can then be generated as the proportion of those random statistics equal to or greater than the original one.

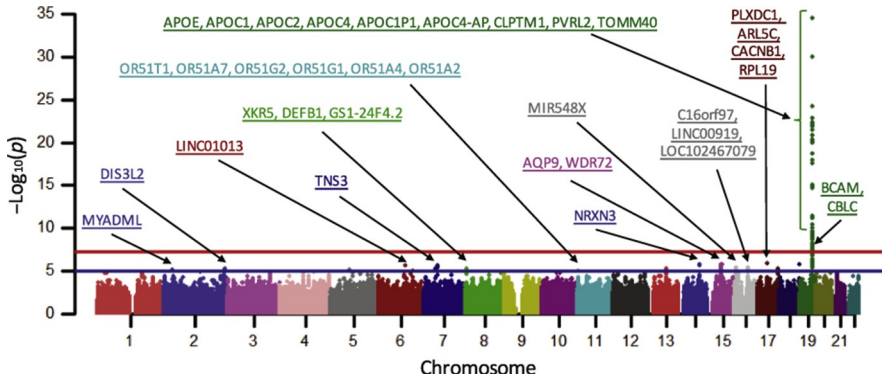

FIG. 14.1

Manhattan plot of a genome-wide association study (GWAS) of the average amyloid burden measure in right precuneus, using an imaging genomic data set from the Alzheimer's Disease Neuroimaging Initiative (ADNI); see Yao et al. (2015) for details of the analysis. The *x*-axis represents the chromosomes and the *y*-axis represents $-\log_{10}(p)$, where p is the SNP-based significance. Selected top hits are labeled with their corresponding gene names.

14.2.2 MULTILOCUS EFFECTS

Despite the great success of single-locus analysis, which has led to the identification of hundreds of candidate SNPs conferring the genetic contribution of complex disorders and diseases, the statistical model is relatively simple and not designed for identifying multilocus effects on imaging QTs. Due to the correlated structure of genetic data, a straightforward linear regression is rarely used for multilocus effect detection and thus is not included here. Note that most machine learning models are not capable of handling either genome-wide genotypes or brain-wide voxel phenotypes. Most existing multilocus imaging genomic studies using machine learning approaches are targeted analyses.

Principal component regression (PCR) is one of those early attempts to address the correlated structure. Unlike traditional regression techniques which operate directly on the original data, in PCR the principal components (PCs) are firstly extracted and linear regression is performed on top of them rather than the original features. Typically only a subset of all the PCs is selected for regression, which has higher variances represented by the large eigenvalues. With genome-wide SNP data, these top PCs extracted in PCR possibly represent the population structure helpful for population stratification (Liu et al., 2010a; Price et al., 2006). The primary advantage of PCR is to tackle the collinearity problem among predictors, since components with low variances are normally excluded in the final analysis. Also, by including only a subset of all the PCs, PCR can result in significant dimension reduction and therefore is very desirable in high-dimensional settings.

The linear mixed model (LMM) is an alternative method designed for purpose of correlated structure. The mixed model refers to the combination of both fixed and random effects. It has a similar form to linear regression:

$$\begin{aligned} &Y = Xw + U\gamma + \boldsymbol{\epsilon} = Xw + \boldsymbol{\epsilon}^*, \\ &\gamma \sim N(0, G), \boldsymbol{\epsilon} \sim N(0, I\sigma^2), \\ &\boldsymbol{\epsilon}^* = XW + \boldsymbol{\epsilon} \sim N(0, UGU^T + I\sigma^2), \end{aligned} \tag{14.3}$$

where X are the fixed factors, eg, age and gender; w are coefficients to be estimated, which indicate the fixed effects; U are the random factors; and γ indicates the corresponding random effects. Instead of estimating γ as a fixed vector in linear regression, LMM assumes an underlying distribution for γ. Since the random effect is modeled as the deviation of a fixed effect, the mean value of γ is set to zero and is usually assumed to have the distribution $\beta \sim N(0, G)$. Here, G works as a prior that explicitly models the variance-covariance structure of random factors, and is the key of LMM for handling correlated genotype data. Note that in LMM only the coefficients of fixed factors are shared in all individuals, representing the population level significance, and those of random factors are different for each subject, indicating individual-level variations. GCTA (Yang et al., 2011, 2013) is one example tool implementing this model for QT heritability analysis. It has been successfully applied to various phenotypes in many disease studies, such as schizophrenia (Ripke et al., 2013), Alzheimer's disease (Ridge et al., 2013), and Parkinson's disease (Do et al., 2011).

Another group of regression models is known to excel in sparsity, which conforms to our assumption that only a few rather than a large portion of SNPs are responsible for the changes of a specific phenotype. The least absolute shrinkage and selection operator (LASSO or lasso) (Eq. 14.4) is a typical example, which achieves the sparsity goal by penalizing the sum of absolute values of all coefficients, also called ℓ_1-norm. This penalized regression problem can be easily solved using a soft-thresholding technique as proposed in Tibshirani (1996), or through an iterative procedure with smooth approximation of ℓ_1-norm (Lee et al., 2006; Schmidt et al., 2007)

$$Y = Xw + \gamma ||w||_1. \tag{14.4}$$

While lasso has shown excellent performances in many studies to identify a small set of SNPs associated with imaging QTs, the sparsity constraint, on the other hand, also leads to competition between correlated SNPs and ultimately yields possibly unstable sparse patterns (Bach, 2008). To address this random selection problem, the elastic net (Zou and Hastie, 2005) was proposed via introducing an extra ℓ_2 penalty term:

$$Y = Xw + \gamma_1 ||w||_1 + \gamma_2 ||w||_2. \tag{14.5}$$

With these two penalty terms, it is capable of seeking a balance point between sparsity and grouping correlated SNPs. Wan et al. (2011) used this model to explore

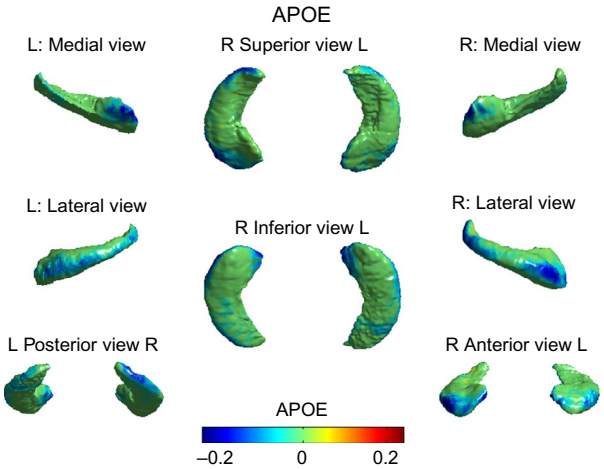

FIG. 14.2

Hippocampal surface map of genetic effects of the *APOE* SNP rs429358 estimated by an elastic net; see Wan et al. (2011) for details.

the genetic risk factors affecting the hippocampal surface and found that *APOE* and *TOMM40* were associated with hippocampal surface changes in anterior and middle regions (see Fig. 14.2 for the *APOE* result).

Other similar approaches include group lasso (Eq. 14.6), fused lasso (Eq. 14.7) and graphical fused lasso, all of which explicitly model the data structure as a prior in the penalty term so that grouped or correlated genetic factors can obtain similar weights in the training procedure:

$$Y = Xw + \gamma \sum_{i=1}^{g} \sqrt{\sum_{j \in G(i)} w_j^2}, \tag{14.6}$$

$$Y = Xw + \gamma \sum_{i<j} |w_i - w_j| . \tag{14.7}$$

Graphical fused lasso is simply an extension of fused lasso that makes constraints only on paired nodes that are connected in a prior graph. Silver et al. (2012b) discussed the group lasso with overlapping problem so that pathway information can be incorporated as a prior for SNP data. Dummy variables were proposed to be added into the weight matrix so that it reverts to a normal group lasso problem. Later, in another work (Silver et al., 2012a), they examined the same term in a multivariate model, which successfully identified several causal pathways associated with longitudinal structural change in the brains of patients with Alzheimer's disease.

Note that in these models estimating the multilocus effect on multiple imaging QTs is equivalent to performing separate analysis for each of them, and does not make use of the fact that multiple imaging QTs are usually highly correlated. This is possibly attributed to partial common underlying genetic architecture. Next, we will introduce several state-of-the-art methods utilizing advanced multivariate regression models to tackle this problem.

L1/L2, also referred to as $L_{2,1}$ (Eq. 14.8), is a typical multivariate model commonly used for identifying the multilocus effects shared across multiple related phenotypes:

$$Y = XW + \gamma ||W||_{2,1} = XW + \gamma \sum_i \sqrt{\sum_j w_{ij}^2}. \tag{14.8}$$

Like group lasso, it also takes advantage of ℓ_1-norm and ℓ_2-norm but distributes the constraints differently. Instead of imposing the ℓ_2-norm constraints within prior groups, here ℓ_2-norm is firstly applied to each row of W, which is the coefficients of one SNP across all QTs, so that they can be pulled together for similarity. The ℓ_1-norm is then applied across rows of W on top of their ℓ_2-norm to guarantee the global sparse patterns of the ultimate result. In this way, it is capable of capturing the multilocus patterns shared across correlated imaging QTs. Inspired by group lasso and this L1/L2 model, Wang et al. (2012a) later proposed a more advanced model, group-sparse multi-task regression and feature selection (G-SMuRFS), in an attempt to find a tradeoff between global sparsity and group level sparsity. As shown in Eq. (14.9),

$$Y = XW + \gamma_1 ||W||_{G2,1} + \gamma_2 ||W||_{2,1}. \tag{14.9}$$

G-SMuRFS has an extra penalty term $||W||_{G2,1} = \sum_{k=1}^{g} \sqrt{\sum_{i \in G(k)} \sum_j w_{ij}^2}$, which can explicitly model the genetic linkage disequilibrium (LD) structure so that SNPs within an LD block tend to be extracted together. On the other hand, $||W||_{2,1}$ helps guarantee the global sparsity so that insignificant SNPs within a generally significant LD block can be removed. Their group also developed another joint classification and regression model with the same penalty terms (Wang et al., 2012b). Both models have been proved to be successful in an AD study, in which a compact set of correlated SNPs was identified with much less root mean square error.

14.2.3 MULTI-SNP-MULTI-QT ASSOCIATIONS

In the brain imaging genomics area, we have both the genetics data and the imaging data at hand. The importance is that they are collected from the same population. Therefore there are essential demands to identify the associations between two multidimensional data, that is, the imaging data and genetics data that come from the same population. Regression techniques usually address problems with only one or very few responses, indicating that they identify the associations between multiple

predictors and one or few responses. Therefore they cannot fully mine the knowledge behind the complex brain imaging genomic data. Although we could regress multiple times for different responses using the same predictors, we still lose the relationship among those responses, especially for the brain structure which is of great interest and importance.

Bi-multivariate correlation models may be an appropriate fit to address this issue. We are aware of many models that have employed the bi-multivariate strategy to bioinformatics studies. Among those techniques, CCA (Hotelling, 1935) is a classical statistical technique that can find the associations between two sets of multidimensional variables. The CCA model is formally defined as follows:

$$\max_{u,v} u^{\mathrm{T}} X^{\mathrm{T}} Y v$$

subject to $u^{\mathrm{T}} X^{\mathrm{T}} X u = 1$ and $v^{\mathrm{T}} Y^{\mathrm{T}} Y v = 1$, where u and v are canonical loadings (or canonical weights) for genetic markers and imaging QTs, respectively, in this chapter.

The standard CCA does not perform feature selection. However, feature selection is an important concern in the brain imaging genomics area, since the data is usually of quite high dimension. A model lacking the ability of feature selection may be inadequate. In order to make use of the advantage of CCA and overcome its weakness, Witten et al. (2009) introduce the sparse CCA (SCCA) using a regularization technique to induce sparsity into the model. They propose the penalized SCCA method using the ℓ_1-norm (lasso) to penalize two canonical loadings. For ease of description, here we only present the penalty function used by the SCCA model. The ℓ_1-norm penalty is defined as

$$\Omega_{\mathrm{Lasso}}(u) = \sum |u_i| \leq c_1,$$

where c_1 is a parameter which can control the sparsity of the corresponding canonical loading. The lasso penalty can result in a larger number of covariates being zero if given suitable parameters. From the point of view of interpretation, this sparsity is desirable because it only captures the most important covariates among a huge amount (Zou and Hastie, 2005).

The SNPs have been known to be correlated other than be perfectly independent from each other. The LD, that is, nonrandom association of alleles at different loci, are common within the gene. This is also the case for the imaging data. The covariates within an imaging modality are not independent, implying that they are correlated. L1-SCCA (lasso-based SCCA) mainly focuses on feature selection, which is insufficient for the association study in brain imaging genomics. In the situation where those covariates are dependent, the L1-SCCA tends to randomly select one covariate from them, and discard the rest. This incurs an unstable solution and makes the interpretation hard, especially on the smoothed data (Lu, 2010). Therefore the models that can capture the entire group of covariates rather than one of them are of great interest.

On the contrary, we can make use of the unstable selection of L1-SCCA. Yan et al. (2015a) propose the bootstrap L1-SCCA (BoSCCA) by repeatedly running the L1-SCCA on different data partitions. Since L1-SCCA randomly selects one of the correlated covariates, it has a high probability of selecting all the correlated covariates as long as the L1-SCCA is run for a sufficient number of times. The BoSCCA is validated on the SNPs of the top 22 AD risk genes and the amyloid imaging measures. The top three components with the highest probability are the frontal medial orbital gyrus, anterior cingulate gyrus, and posterior cingulate gyrus. For the genotypes, *APOE* gene and its neighbors *APOC1* and *TOMM40* are selected with high probability. All of the genotypes and imaging phenotypes are associated with AD.

In order to accommodate other types of structures in the data, several structured SCCA methods (Du et al., 2014, 2015a; Yan et al., 2014a; Witten et al., 2009) have been used. We classify these SCCA methods into two types according to their distinct regularization terms. One type used the group lasso penalty, and the other used the graphical fused lasso penalty to conduct feature selection and feature grouping. The first type of structured SCCA methods, that is, the group lasso-based SCCA, required prior knowledge to define the group structure. The group lasso penalty is defined as follows:

$$\Omega_{\mathrm{GL}}(u) = \sum_{i=1}^{g} \sqrt{\sum_{j \in G_i} u_j^2} \leq c_1$$

This penalty function has two aspects: (1) the intragroup ℓ_2-norm constraint, which enables covariates in the same group to have equal weights; (2) the intergroup ℓ_1-norm constraint, which assures sparsity in terms of groups. That is, for a group of covariates, all of them will be selected or discarded together; and for covariates of different groups, they are independent of each other.

Du et al. (2014) proposed structure-aware SCCA (S2CCA) using group lasso, and incorporated both the covariance matrix information and the prior knowledge information to recover group-level bi-multivariate associations. The authors confirmed that a strong association exists between the *APOE* gene and a VBM imaging marker (hippocampus).

The second kind of structured SCCA methods use graph/network-guided fused lasso penalties. The SCCA methods above will lose efficacy because they are prior knowledge dependent. These methods can perform well on any given prior knowledge. If the prior knowledge is not available, these methods can also work by using the sample correlation to define the graph/network constraint. Apart from the L1-SCCA, Witten et al. (2009) also introduced a smoothed penalized SCCA which utilized the fused lasso as penalty, that is,

$$\Omega_{\mathrm{FL}}(u) = \sum |u_i| + \lambda \sum_{i \neq j} |u_i - u_j| \leq c_1$$

Here u_i and u_j are two neighboring covariates after ordering the covariates. The fused lasso-based SCCA (FL-SCCA) can be viewed as a degenerate network-guided

method since it imposes pairwise constraint to ensure smoothness. However, it requires these covariates to be ordered before running it, which limits its performance greatly. This is because it loses the relationship for those covariates that are not adjacent. Du et al. (2015a) have tested the FL-SCCA on the *APOE* gene and the amyloid burden measure of the AD database, and its results are acceptable.

Yan et al. (2014a) find that the brain structural constrait could be made up of a network instead of several nonoverlapping groups. However, the S2CCA can only identify the nonoverlapping group structure, indicating that it cannot deal with a more complex group or graph/network structure. Yan et al. introduce the knowledge-guided SCCA (KG-SCCA) where the network-guided penalty is defined as

$$\Omega_{\mathrm{NG}}(u) = \sum |u_i| + \lambda \sum_{(i,j)\in E, i<j} \tau(w_{ij}) ||u_i - \mathrm{sign}(w_{ij}) u_j||_2 \leq c_1$$

where u_i and u_j are two connected nodes on a prior network; $\mathrm{sign}(w_{ij})$ indicates the sign of their correlation and $\tau(w_{ij})$ stands for the strength of their connection. This penalty is an extension to the fused lasso; however, it does not require the covariates to be ordered first. After using the AD-related genetic markers and imaging markers, they validate that the SNPs in the *APOE* gene are strongly associated with the amyloid burden measure.

Du et al. (2015b) proposed another network-guided SCCA which utilized the GraphNet regularization term of Grosenick et al. (2013). The GraphNet-based SCCA (GN-SCCA) employs the Laplacian matrix-based GraphNet, which can be defined as

$$\Omega_{\mathrm{GraphNet}}(u) = \sum |u_i| + \lambda \sum_{(i,j)\in E, i<j} (u_i - u_j)^2 \leq c_1$$

GraphNet is an extension to the elastic net penalty, which has wider applications. The GN-SCCA can identify structure associations even if we have little prior knowledge or we do not have it. Based on the correlation matrix of either modality, GN-SCCA can incorporate the correlation matrices and use them to guide the structure identification. They apply the SCCA on the data set with respect to the AD-related genetic markers and the human brain amyloid burden. The result confirms that the marker *APOE* gene has a strong relationship between the frontal measurements in the brain.

Recently, Du et al. (2015a) extend the network-guided SCCA to a more robust one. Their SCCA utilizes the graph OSCAR penalty (Yang et al., 2012):

$$\Omega_{\mathrm{GOSCAR}}(u) = \sum |u_i| + \lambda \sum_{(i,j)\in E, i\neq j} \max(|u_i|, |u_j|) \leq c_1$$

The GOSCAR penalty employs the pairwise ℓ_∞-norm to constrain every pair of covariates, and expects them to be equal if they are highly correlated or dissimilar if the correlation between them is quite low. The importance of this penalty is that it is sign independent, and thus it only focuses on whether $|u_i|$ equals $|u_j|$. Therefore the GOSC-SCCA will not suffer from the sign directionality issue and is more robust

than those SCCAs above. They also assure an association between the *APOE* gene and the brain amyloid burden measurement.

Another way to handle the correlation between both multiple genetic markers and multiple imaging markers is to take advantage of the RRR technique. The RRR can be viewed as an extension to traditional regression techniques, which is capable of predicting multiple responses simultaneously. In traditional RRR, for the weight matrix $W \in \mathbb{R}^{p*q}$ satisfying rank $(W) \leq \min(p, q)$, it can be written as a product of two full rank r matrices $W = BA^{\mathrm{T}}$, where $B \in \mathbb{R}^{p*r}$ and $A \in \mathbb{R}^{q*r}$. The RRR can then be reformulated to solve the minimization problem:

$$\min_{A,B} \mathrm{Tr}\{(Y - XBA^{\mathrm{T}})\Gamma(Y - XBA^{\mathrm{T}})^{\mathrm{T}}\}$$

This factorization of W not only helps decrease the parameters to be estimated, but also enables the sparsity constraints on both SNPs and imaging QTs, respectively. Sparse reduced rank regression (sRRR) (Vounou et al., 2010) is a method in which the ℓ_1-norm is utilized to penalize the coefficients A and B, so that sparse variable selection can be achieved simultaneously on both sides. By assuming that $X^{\mathrm{T}}X = I$ and $\Gamma = I$, the minimization problem of sRRR can be formulated as

$$\min_{A,B} -2\mathrm{Tr}\{A^{\mathrm{T}}Y^{\mathrm{T}}XB\} + \mathrm{Tr}\{A^{\mathrm{T}}AB^{\mathrm{T}}B\} + \gamma_1||A||_1 + \gamma_2||B||_1$$

Like the SCCA, the sRRR is also a biconvex problem, and thus can be solved by an iterative procedure by alternately fixing one coefficient and updating another using soft-thresholding in each step (Tibshirani, 1996). The sRRR method uses multiple phenotypes as responses and multiple genotypes as predictors. They successfully detect the most important variables in both the genetic and imaging domains.

The generalized low rank regression (GLRR) model proposed recently by Zhu et al. (2014) is another approach making use of the matrix rank to induce sparsity, but under Bayesian infrastructure. Although both SCCA and sRRR can handle multiple genotypes and imaging phenotypes, they are different in terms of recovering the association. SCCA works symmetrically and there is no difference between the genotypes and the imaging phenotypes; however, the sRRR performs more like traditional regression and we cannot swap the genotypes and the imaging phenotypes to obtain the same findings.

14.3 MINING HIGHER LEVEL IMAGING GENOMIC ASSOCIATIONS VIA SET-BASED ANALYSIS

Higher level association analysis has been demonstrated that it can yield biologically meaningful findings by integrating prior knowledge (eg, pathway information) into a set of genetic findings. The prior knowledge could be from gene ontology (GO), functional annotation databases and pathway analysis systems. Various gene-based

association tests have been proposed to construct genetic candidates based on phenotype-associated variants from GWAS. Recently, it has been applied to brain imaging genomic applications, by adopting prior knowledge from both genetic and imaging domains, to explore the Gene Set (GS) effects on specific imaging QT, or a predefined BC containing multiple imaging QTs.

In this section, we discuss two classes of higher level imaging genomic association tests: context-based test and context-free test, categorized by whether they use background information or not; and further discuss a two-dimensional enrichment analysis paradigm that jointly explores meaningful GS-BC modules.

14.3.1 CONTEXT-BASED TEST

Given a set of candidate genes, a context-based method tests if there is a trait-association difference between this GS and a random GS of the same size. By comparing the proportion of trait-associated signals from candidate and random GSs, a context-based test tells how important the set of candidate genes is compared to random GSs. Because of this, methods applying the context-based test require the significance results (eg, p-values) of not only the candidate genes but also all the other genes in the relevant context. The most common test is the pathway enrichment test, which can be classified into two types: over-representation analysis and rank-based analysis. Below we briefly discuss both types, including their applications in brain imaging genomics.

14.3.1.1 Over-representation analysis

In the over-representation test, a threshold is used to determine the set of candidate genes L (ie, genes with QT-associated p-values exceeding the threshold), such that all genes in L are significantly associated with specific imaging QT. We test if a predefined GS S (eg, a pathway) is over-represented in L. This can be formulated as an independence test problem. The most commonly used over-representation tests are based on hypergeometric (Fisher's exact test), binomial, and/or χ^2 distribution (Draghici et al., 2003; Goeman and Buhlmann, 2007).

Assume we have all N genes in the analysis. Of these, $n = |L|$ genes in the set L are significant ones (ie, p-value exceeds a threshold), $m = |T|$ genes are from a given GS T of interest (eg, a pathway), and k out of n significant genes are from T. Using a hypergeometric test (Fisher's exact test) for illustration, the over-represented p-value of having k or more genes from T in L can be calculated by summing the probabilities of a random set of n genes having $k, k+1, \ldots, n$ genes from T:

$$p = \Pr\left(|L \cap T| \geq k\right) = \sum_{i \geq k} \left(\binom{m}{i} \binom{N-m}{n-i} \Big/ \binom{N}{n} \right). \tag{14.10}$$

The hypergeometric distribution is rather difficult to calculate when the number of genes involved is large. However, it tends to be a binomial distribution when N is large.

There are a number of studies that have performed over-representation analyses and identified meaningful functional GSs with significant associations to the relevant phenotype or disease. For example, Perez-Palma et al. (2014) applied hypergeometric test to construct a network-based pathway enrichment using meta-analysis statistics of GWAS, and identified the over-representation of the glutamate signaling pathway in Alzheimer's disease.

There are several issues in the over-representation test. First, the enrichment statistic is based on an arbitrarily selected threshold (eg, $p \leq 0.05$) that is used for determining the significant candidate set. Interesting signals might be missed when there are many modest trait-associated genes that do not pass the threshold. Second, the over-representation tests consider only the number of significant genes but ignores their strength of associations. Third, the genes are not independent from the others. Over-representation tests treat each gene as an independent unit, and ignore their correlation structure, which may yield a biased enrichment estimation.

14.3.1.2 Rank-based enrichment analysis

To overcome the limitations of over-representation analysis, rank-based enrichment analysis has been developed to include all gene-level p-values in the analysis. One widely used method is Gene Set enrichment analysis (GSEA), which was originally devised for gene expression data analyses and then extended to GWAS analyses. The GSEA tests whether genes from a predefined set S (eg, a pathway) are distributed in the top (or bottom) of a ranked gene list L ordered by gene-level p-values, and thus is significantly associated with the GWAS trait. The implementation of the GSEA algorithm is briefly described below.

Given a predefined GS S and a sorted list of genes L with gene-level statistics (eg, gene-level p-values from the GWAS of a specific imaging trait), an enrichment score (ES) of S is calculated using a Kolmogorov-Smirnov (K-S) like statistic with weight 1. That is, by walking down the list L, a running-sum statistic is increased when encountering a gene in S, and is decreased when encountering a gene not in S. The ES is then provided by the maximum deviation from zero of the running sum. Statistical significance of ES (empirical p-value) is then estimated by performing phenotype-based permutation.

In brain imaging genomics, original GSEA and its modifications have been widely applied in pathway or network enrichment analysis. Ramanan et al. (2012) performed genome-wide pathway analysis of memory impairment in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort using GSA-SNP, a GSEA-based GS analysis software. They identified 27 pathways with significant ESs against the composite memory score, of which most are involved in memory consolidation. One of the pathways displays colocalized expression in normal brain tissue along with known AD risk genes. Younesi and Hofmann-Apitius (2013), adopted GSEA to validate the functional association with a brain region-specific protein-protein interaction subnetwork extracted by text mining.

Rank-based methods overcome some limitations of several over-representation tests by using all gene-level statistics, without requiring a user-specified threshold. Furthermore, a phenotype-based permutation can keep the correlation structure among genes, and thus provide a more reasonable assessment of significance than permuting genes. However, rank-based methods also have several limitations. First, like over-representation methods, rank-based tests consider pathways independently, which often overlap with one another. Because of this, a pathway may be significantly enriched due to the common genes it shares with a real enriched pathway. Second, rank-based methods take into account the ranks of genes but ignore the strength of associations between genes and phenotypes. Some modifications have been proposed to improve this problem by adding weights to ranked genes based on their association strengths (Mooney and Wilmot, 2015).

14.3.2 CONTEXT-FREE TEST

In contrast to context-based tests, context-free tests use another strategy to formulate a null hypothesis. They have been applied in gene-based association analysis. Gene-based association analysis tests the association between a gene and a phenotype, based on the statistics of SNPs within this gene and without needing information from outside of the gene. Below we briefly discuss two groups of gene-based association analyses.

Genome-wide analysis has been employed in brain imaging genomics to identify individual susceptibility loci of neurodegenerative diseases, which however explain only a modest proportion of the total variance in liability to imaging trait or disease. Gene-based analysis jointly considers all variants within a gene to obtain a single statistic representing the association significance of the entire gene, and thus has three advantages. First, the gene is the functional unit of the genome and highly consistent across individuals. Second, compared to SNP-level association tests, gene-based analysis reduces the number of multiple corrections (from ∼0.5 million to 20,000–30,000). Third, the findings from gene-based analysis can be directly adopted by further analysis such as protein-protein interaction, pathway enrichment analysis, and so on.

A number of gene-based association tests have been proposed and can be categorized into two groups. One group incorporates the full set of SNPs within the gene, to test their association with a specified phenotype. The other tests the null hypothesis that no SNPs within the given gene show association with phenotype, whereas the alternative is that at least one SNP in the gene is associated with the phenotype. In other words, the first group simultaneously consider all SNPs and the second one focuses on the best SNP. Typical methods of the first group include regression, PLINK set-test (Purcell et al., 2007), Fisher's combination (Curtis et al., 2008), and VEGAS test (Liu et al., 2010b). The second group includes the VEGAS-Max test (Liu et al., 2010b), GATES (LI et al., 2011), and so on.

Linear regression (for QTs) and logistic regression (for binary traits) are straightforward methods to evaluate the overall association between a gene and a trait. In

regression, all SNPs from the gene are treated as predictors simultaneously, with the phenotype as response. The statistical power of regression might be decreased when a lot of SNPs are involved, due to high degrees of freedom. Various strategies have been proposed to reduce the dimensionality by collapsing correlated SNPs, like by Fourier transformation (Wang and Elston, 2007), principal component analysis (PCA) (Gauderman et al., 2007; Wang and Abbott, 2008), and clustering analysis (Buil et al., 2009). Raw genotyping and phenotyping data are required for all regression methods.

Set-based tests implemented in PLINK use an LD structure within the SNPs to select a subset of representative ones. Association significance of the gene is then calculated as the mean statistic of the subset. The empirical p-value is obtained by repeating the permutation procedure many times. It should be noted that the LD structure among genotype variants should be kept in each permutation, such that the phenotype label can be changed. PLINK set-based tests are time consuming when applied to genome-wide results due to the permutations.

Another strategy is Fisher's combination method that combines all SNP-level p-values using the following formula:

$$X^2 \sim -2\sum_{i=1}^{n} \ln(p_i). \tag{14.11}$$

From Fisher's combination, the combination of n independent tests under the same null hypothesis will follow a χ^2 distribution with $2n$ degrees of freedom (n is the number of the SNP p-value). However, in gene-based analysis, the SNP association tests are not independent due to their LD. Permutation is still required to gain empirical significance, which takes computational intensity. Fisher's combination test only requires SNP-level p-values but not raw genotyping and phenotyping data.

Versatile gene-based association study (VEGAS), developed by Liu et al. (2010b), replaces permutation with simulation based on Fisher's combination, to reduce computation time. In VEGAS, all SNP p-values from a given gene are firstly transformed to upper-tail χ^2 statistics with one degree of freedom (df). The gene-based statistic is then calculated as the sum of all (or a predefined subset) of these χ^2 1 df statistics according to Fisher's combination. As mentioned before, the statistic will follow a χ^2 distribution with n df if all SNPs are independent. However, this is unlikely to be the case in real genetic data.

Instead of performing permutations many times to gain an empirical p-value, VEGAS adopts the Monte Carlo approach that makes use of simulations from the multivariate normal distribution to do this. The details are as follows. Given a gene with n SNPs, Σ is an $n \times n$ matrix of pairwise LD (r). An n-element multivariate normally distributed vector with mean 0 and variance Σ is simulated in the next two steps: (1) generate n independent, standard, normally distributed random variables, (2) multiply n variables from (1) by the Cholesky decomposition matrix of Σ (the lower triangular matrix C such that $CC^{\mathrm{T}} = \Sigma$). The new random vector $Z = (z_1, z_2, \ldots, z_n)$ will have a multivariate normal distribution $Z \sim N_n(0, \Sigma)$. It then

transforms Z into a vector of correlated χ^2 1 df variables, $Q = (q_1, q_2, \ldots, q_n), q_i = z_i^2$. Simulated gene-based test statistics are then calculated as $\sum_{i=1}^{n} q_i$, which will have the same approximate distribution as the observed statistics. The empirical p-value can be calculated by repeating this simulation procedure many times.

GATES is one of the most common methods in the second group. Instead of considering all SNPs within the gene, GATES extends the Simes test based on the null hypothesis that no SNPs within the gene are associated with the disease. The gene-based p-value is calculated using

$$p_{\text{value}} = m_{\text{e}} * \text{Min}\left(\frac{p_{(j)}}{m_{\text{e}(j)}}\right), \tag{14.12}$$

where m_{e} is the effective number of independent p-values among the m SNPs and $m_{\text{e}(j)}$ is the effective number of independent p-values among the top j SNPs.

In VEGAS, one can also only use the best SNP (ie, the most significant SNP) within the gene to calculate gene-level statistics, called VEGAS-Max and belonging to the second gene-based analysis group.

In imaging genomics, GWAS have been widely performed on imaging QTs and identified a number of risk genetic variants (Saykin et al., 2010; Shen et al., 2010). While both groups of gene-based analysis have been widely used to measure the associations between gene and imaging, Perez-Palma et al. (2014) employed GATES to calculate gene-level brain imaging associations extracted from meta-analysis. Liang et al. (2015) used VEGAS to assign SNP p-values to respective genes in their network-based high-level imaging genomic association study, where 200 genes were identified to be associated with all 14 subcortical imaging measures, including left hippocampus and right amygdala, which were previously demonstrated to be related to Alzheimer's disease.

14.3.3 TWO-DIMENSIONAL IMAGING GENOMIC ENRICHMENT ANALYSIS

Brain imaging genomics is an emerging field that studies how genetic variation influences brain structure and function. In the genetics domain, GWAS have been performed to identify genetic markers such as SNPs that are associated with brain imaging QTs (Saykin et al., 2010; Shen et al., 2010). From Sections 14.3.1 and 14.3.2, using biological pathways and networks as prior knowledge, enrichment analysis has been performed to discover pathways or network modules enriched by GWAS findings to enhance statistical power and aid biological interpretation.

However, present analytic methods used in GWAS of imaging QTs typically ignore either the interrelated structure between genes or the correlation between imaging QTs, and are insufficient to provide insight into the mechanisms of complex diseases that could involve multiple genes and multiple QTs.

Recently, IGEA (Yao et al., 2015), a new enrichment analysis paradigm, has been proposed to jointly consider sets of interest (ie, GS and BC) in both genetic and

imaging domains and examine whether any given GS-BC module is enriched in a list of gene-QT findings. IGEA aims to discover higher level associations between meaningful GS and BC, which typically include multiple genes and multiple QTs. Fig. 14.3 shows a brief overview of the IGEA framework.

In IGEA, there are many types of prior knowledge that can be used to determine meaningful GS and BC entities. In the genomic domain, the prior knowledge could be based on GO or functional annotation databases; in the imaging domain, the prior knowledge could be neuroanatomic ontology or brain databases. Gene-QT association can be identified from brain imaging GWAS results, or existing imaging genomic findings. Yao et al. (2015) in their study used brain-wide expression data from the Allen Human Brain Atlas (AHBA, Allen Institute for Brain Science, Seattle, WA; available from http://www.brain-map.org/) to extract GS and BC modules such that genes within a GS share similar expression profiles and so do ROIs

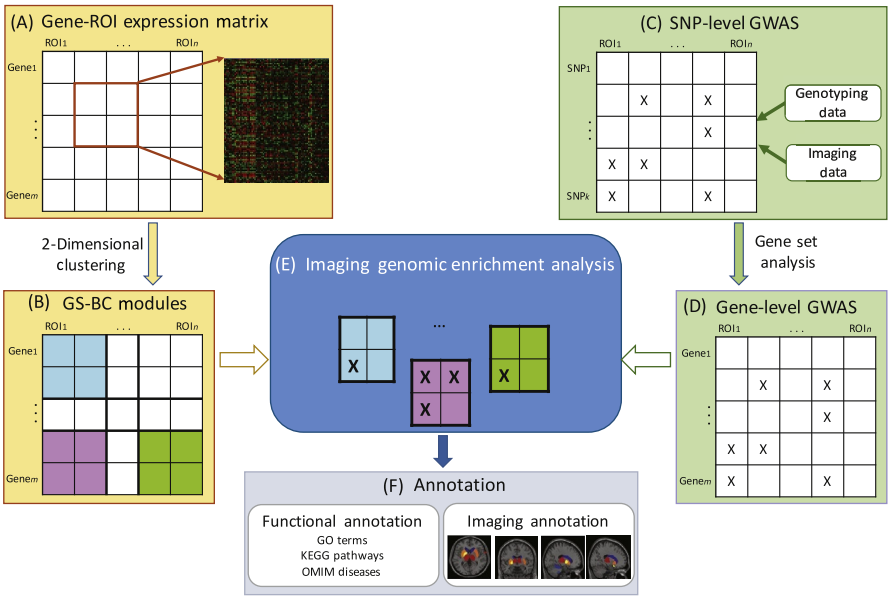

FIG. 14.3

Overview of the IGEA framework proposed in Yao et al. (2015). (A) Construct gene-ROI expression matrix from the Allen Human Brain Atlas (AHBA data). (B) Construct GS-BC modules by performing 2D hierarchical clustering, and then filter out nonsignificant 2D clusters. (C) Perform SNP-level GWAS of brain-wide imaging measures. (D) Map SNP-level GWAS findings to gene-based results. (E) Perform IGEA by mapping gene-based GWAS findings to the identified GS-BC modules. (F) For each enriched GS-BC module, examine the GS using GO terms, KEGG pathways, and OMIM disease databases, and map the BC to the brain.

within a BC, and extracted gene-QT associations from brain-wide genome-wide association analysis.

The enrichment test can be performed by adopting either over-representation or rank-based test described in Section 14.3.1. Yao et al. (2015) used the imaging genomic data from the ADNI as test beds, and identified 12 significant GS-BC modules. From their results, most identified GSs had significant functional enrichment, and several could be related to the neurodegenerative disease and its development; identified BCs also involved structures responsible for neurodegenerations, including motivated behaviors, sensory information processing, executive functions, major spots for amyloid accumulation in AD, and so on. By jointly considering the complex relationships between interlinked genetic markers and correlated brain imaging phenotypes, higher level imaging genomic association analysis can provide additional power for extracting biological insights on neurogenomic associations at a systems biology level.

14.4 DISCUSSION

14.4.1 PROMINENT FINDINGS

In the last decade, statistical and machine learning has been playing an essential role in imaging genomic studies and has successfully promoted the discoveries of biologically meaningful biomarkers as well as underlying association patterns. Potkin et al. used PLINK to examine the genetic basis of hippocampal atrophy in Alzheimer patients and successfully identified five risk genes involved in the regulation of protein degradation, apoptosis, neuronal loss, and neurodevelopment. Instead of structural changes, another group recently reported a GWAS result between brain connectivity and genetic variants, in which one risk gene *SPON1* was identified and further replicated in an independent cohort (Jahanshad et al., 2013). In Kohannim et al. (2012), multilocus effects were investigated against the temporal lobe volume using the elastic net method, where two SNPs in genes *RBFOX1* and *GRIN2B* were, respectively, found to be highly contributory genotypes. By applying a variant of PCR across all genes and a large database of voxel-wise imaging data, Hibar et al. (2011) identified 10 significant SNPs in *GRB-associated binding protein 2 gene* (*GAB2*), to significantly associate with all voxels. Similarly, Vounou et al. (2012) used sRRR, after validating its superior performance on a large-scale dataset, to examine the multilocus effects over the voxel-wise imaging data but using their longitudinal changes. Their findings confirmed the key role of *APOE* and *TOMM40* in AD and highlighted some other potential associations as well. In addition to these variant level findings, prior knowledge guided methods also help to reveal candidate pathways, whose perturbation may possibly lead to changes in imaging phenotypes. For example, in Silver et al. (2012a), an overlapping group lasso penalty added to RRR was used to model the pathway belonging to SNPs and several pathways were reported to be associated with longitudinal structure

changes in brain, such as the insulin signaling pathway, Chemokine signaling pathway, and Alzheimer's disease pathway. Recently, with the help of S2CCA, Du et al. (2014, 2015a) examined the multiple-to-multiple relationship between *APOE* and brain-wide amyloid accumulations, and reported localized amyloid patterns affected by the joint effect of *APOE* SNPs. At the same time, in Yan et al. (2014a) the same experiment was performed using KG-SCCA, where a transcriptome co-expression network was applied as a prior, and similar amyloid accumulation patterns were identified.

14.4.2 FUTURE DIRECTIONS

The advent of brain imaging and genotyping techniques has brought unprecedented opportunities for discoveries of the underlying disease mechanisms. Whole genome sequencing, longitudinal imaging, and brain connectome data are now widely accessible, requiring more complicated models to capture the reality hidden behind the data. Existing methods, though successful, still work on a relatively small scale and will have limited power as the number of datasets becomes ever larger. Regression or association models capable of dealing with superdimensionality are highly desired in the near future.

Big data is another promising future direction. Plenty of efforts have been recently made to introduce big data science into the brain imaging genomics field, which is believed to hold great promise for overcoming the computation bottlenecks. Most attempts now are focusing on boosting the performance of traditional GWAS by taking advantage of supercomputing techniques. Example high-performance software tools include FaST-LMM (Lippert et al., 2013), EpiGPU (Hemani et al., 2011), and GBOOST (Yung et al., 2011). Yan et al. (2014b) made an initial attempt to accelerate the SCCA implementation by combining Intel Math Kernel Library (MKL) and the offload model for Intel Many Integrated Core (MIC), in which they observed consistent twofold speedup without any code modification. Another recent study (Wang et al., 2013) coupled the Map/Reduce framework with Random Forest for associating SNPs and imaging phenotypes, which achieved at most 10-fold improvement in running time. Despite little work for now, these initial efforts and their promising results show the great potential of big data techniques. More applications of machine learning in brain- and genome-wide studies are expected in future imaging genomics research.

REFERENCES

Bach, F.R., 2008. Consistency of the group lasso and multiple kernel learning. J. Mach. Learn. Res. 9, 1179–1225.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate—A practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57, 289–300.

Buil, A., Martinez-Perez, A., Perera-Lluna, A., Rib, L., Caminal, P., Soria, J.M., 2009. A new gene-based association test for genome-wide association studies. BMC Proc. 3 (7), S130.

Bush, W.S., Moore, J.H., 2012. Chapter 11: Genome-wide association studies. PLOS Comput. Biol. 8, e1002822.

Chi, E.C., Allen, G.I., Zhou, H., Kohannim, O., Lange, K., Thompson, P.M., 2013. Imaging genetics via sparse canonical correlation analysis. IEEE Int. Symp. Biomed. Imaging 2013, 740–743.

Curtis, D., Vine, A.E., Knight, J., 2008. A simple method for assessing the strength of evidence for association at the level of the whole gene. Adv. Appl. Bioinform. Chem. 1, 115–120.

Do, C.B., Tung, J.Y., Dorfman, E., Kiefer, A.K., Drabant, E.M., Francke, U., et al., 2011. Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. PLoS Genet. 7, e1002141.

Draghici, S., Khatri, P., Martins, R.P., Ostermeier, G.C., Krawetz, S.A., 2003. Global functional profiling of gene expression. Genomics 81, 98–104.

Du, L., Jingwen, Y., Kim, S., Risacher, S.L., Huang, H., Inlow, M., et al., 2014. A novel structure-aware sparse learning algorithm for brain imaging genetics. Med. Image Comput. Comput. Assis. Interven. 17, 329–336.

Du, L., Huang, H., Yan, J., Kim, S., Risacher, S.L., Inlow, M., et al., 2015a. Structured sparse CCA for brain imaging genetics via graph OSCAR. In: The International Conference on Intelligent Biology and Medicine, Indianapolis, USA.

Du, L., Yan, J., Kim, S., Risacher, S., Huang, H., Inlow, M., et al., 2015b. GN-SCCA: GraphNet Based Sparse Canonical Correlation Analysis for Brain Imaging Genetics. Springer, New York.

Gauderman, W.J., Murcray, C., Gilliland, F., Conti, D.V., 2007. Testing association between disease and multiple SNPs in a candidate gene. Genet. Epidemiol. 31, 450–450.

Goeman, J.J., Buhlmann, P., 2007. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 23, 980–987.

Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., Taylor, J.E., 2013. Interpretable whole-brain prediction analysis with GraphNet. NeuroImage 72, 304–321.

Hemani, G., Theocharidis, A., Wei, W., Haley, C., 2011. EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. Bioinformatics 27, 1462–1465.

Hibar, D.P., Stein, J.L., Kohannim, O., Jahanshad, N., Jack, C.R., Weiner, M.W., et al., 2011. Principal components regression: multivariate, gene-based tests in imaging genomics. In: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Chicago, IL, pp. 289–293.

Hotelling, H., 1935. The most predictable criterion. J. Edu. Psychol. 26, 139.

Jahanshad, N., Rajagopalan, P., Hua, X., Hibar, D.P., Nir, T.M., Toga, A.W., et al., 2013. Genome-wide scan of healthy human connectome discovers SPON1 gene variant influencing dementia severity. Proc. Natl. Acad. Sci. USA 110, 4768–4773.

Kohannim, O., Hibar, D.P., Jahanshad, N., Stein, J.L., Hua, X., Toga, A.W., et al., 2012. Predicting temporal lobe volume on MRI from genotypes using L(1)-L(2) regularized regression. In: Proceedings of the 9th International Symposium on Biomedical Imaging: ISBI 2012, Barcelona, Spain, 1160–1163.

Lee, S.I., Lee, H., Abbeel, P., Ng, A.Y., 2006. Efficient L1 regularized logistic regression. In: Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06).

Li, M.X., Gui, H.S., Kwan, J.S., Sham, P.C., 2011. GATES: a rapid and powerful gene-based association test using extended Simes procedure. Am. J. Hum. Genet. 88, 283–293.

Liang, H., Meng, X., Chen, F., Zhang, Q., Yan, J., Yao, X., et al., 2015. A network-based framework for mining high-level imaging genetic associations. In: MICCAI Workshop on Imaging Genetics, October 9, 2015.

Lin, D., Calhoun, V.D., Wang, Y.P., 2014. Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. Med. Image Anal. 18, 891–902.

Lippert, C., Listgarten, J., Davidson, R.I., Baxter, S., Poon, H., Kadie, C.M., et al., 2013. An exhaustive epistatic SNP association analysis on expanded Wellcome Trust data. Sci. Rep. 3, 1099.

Liu, J.Y., Hutchison, K., Perrone-Bizzozero, N., Morgan, M., Sui, J., Calhoun, V., 2010a. Identification of genetic and epigenetic marks involved in population structure. PLoS ONE 5 (10), e13209.

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.

Lu, Z.Q.J, 2010. The elements of statistical learning: data mining, inference, and prediction, second ed. J. R. Stat. Soc. A 173, 693–694.

Mooney, M.A., Wilmot, B., 2015. Gene set analysis: a step-by-step guide. Am. J. Med. Genet. B 168, 517–527.

Perez-Palma, E., Bustos, B.I., Villaman, C.F., Alarcon, M.A., Avila, M.E., Ugarte, G.D., et al., 2014. Overrepresentation of glutamate signaling in Alzheimer's disease: network-based pathway enrichment using meta-analysis of genome-wide association studies. PLoS ONE 9, e95413.

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.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D., 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38, 904–909.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., et al., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575.

Ramanan, V.K., Kim, S., Holohan, K., Shen, L., Nho, K., Risacher, S.L., et al., 2012. Genome-wide pathway analysis of memory impairment in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort implicates gene candidates, canonical pathways, and networks. Brain Imaging Behav. 6, 634–648.

Ridge, P.G., Mukherjee, S., Crane, P.K., Kauwe, J.S., 2013. Alzheimer's disease: analyzing the missing heritability. PLoS ONE 8, e79771.

Ripke, S., O'dushlaine, C., Chambert, K., Moran, J.L., Kahler, A.K., Akterin, S., et al., 2013. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat. Genet. 45, 1150–1159.

Saykin, A.J., Shen, L., Foroud, T.M., Potkin, S.G., Swaminathan, S., Kim, S., et al., 2010. Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: genetics core aims, progress, and plans. Alzheimer's Dement. 6, 265–273.

Schmidt, M., Fung, G., Rosales, R., 2007. Fast optimization methods for l1 regularization: a comparative study and two new approaches. In: Proceedings of the 18th European Conference on Machine Learning: ECML 2007, Warsaw, Poland, vol. 4701, pp. 286–297.

Shen, L., Kim, S., Risacher, S.L., Nho, K., Swaminathan, S., West, J.D., et al., 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, 1051–1063.

Silver, M., Janousova, E., Hua, X., Thompson, P.M., Montana, G., Alzheimer's Disease Neuroimaging Initiative, 2012a. Identification of gene pathways implicated in Alzheimer's

disease using longitudinal imaging phenotypes with sparse regression. NeuroImage 63, 1681–1694.
Silver, M., Montana, G., Alzheimer's Disease Neuroimaging Initiative, 2012b. Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps. Stat. Appl. Genet. Mol. Biol. 11, 7.
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.
Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B 58, 267–288.
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, 1147–1159.
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.
Wan, J., Kim, S., Inlow, M., Nho, K., Swaminathan, S., Risacheri, S.L., et al., 2011. Hippocampal surface mapping of genetic risk factors in AD via sparse learning models. Med. Image Comput. Comput. Assis. Interven. 14, 376–383.
Wang, T., Elston, R.C., 2007. Improved power by use of a weighted score test for linkage disequilibrium mapping. Am. J. Hum. Genet. 80, 353–360.
Wang, K., Abbott, D., 2008. A principal components regression approach to multilocus genetic association studies. Genet. Epidemiol. 32, 108–118.
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.
Wang, H., Nie, F., Huang, H., Risacher, S.L., Saykin, A.J., 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.
Wang, Y., Goh, W., Wong, L., Montana, G., Alzheimer's Disease Neuroimaging Initiative, 2013. Random forests on Hadoop for genome-wide association studies of multivariate neuroimaging phenotypes. BMC Bioinformatics 14 (Suppl. 16), S6.
Witten, D.M., Tibshirani, R., Hastie, T., 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10, 515–534.
Yan, J., Du, L., Kim, S., Risacher, S.L., Huang, H., Moore, J.H., et al., 2014a. Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm. Bioinformatics 30, i564–i571.
Yan, J., Zhang, H., Du, L., Wernert, E., Saykin, A.J., Shen, L., 2014b. Accelerating sparse canonical correlation analysis for large brain imaging genetics data. In: Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment, Atlanta, GA, USA, 2616515: ACM, pp. 1–7.
Yan, J., Du, L., Kim, S., Risacher, S.L., Huang, H., Moore, J.H., et al., 2015a. BoSCCA: mining stable imaging and genetic associations with implicit structure learning. In: MICGen 2015: MICCAI Workshop on Imaging Genetics.
Yan, J., Kim, S., Nho, K., Chen, R., Risacher, S.L., Moore, J.H., et al., 2015b. Hippocampal transcriptome-guided genetic analysis of correlated episodic memory phenotypes in Alzheimer's disease. Front. Genet. 6, 117.

Yang, J., Lee, S.H., Goddard, M.E., Visscher, P.M., 2011. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88, 76–82.

Yang, S., Yuan, L., Lai, Y.C., Shen, X., Wonka, P., Ye, J., 2012. Feature grouping and selection over an undirected graph. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: KDD 2012, Beijing, China, pp. 922–930.

Yang, J., Lee, S.H., Goddard, M.E., Visscher, P.M., 2013. Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. Meth. Mol. Biol. 1019, 215–236.

Yao, X., Yan, J., Kim, S., Nho, K., Risacher, S.L., Inlow, M., et al., 2015. Two-dimensional enrichment analysis for mining high-level imaging genetic associations. In: Proceedings of the 8th International Conference on Brain Informatics & Health: BIH 2015, London, UK, Lecture Notes in Artificial Intelligence, 9250, 115–124.

Younesi, E., Hofmann-Apitius, M., 2013. Biomarker-guided translation of brain imaging into disease pathway models. Sci. Rep. 3, 3375.

Yung, L.S., Yang, C., Wan, X., Yu, W., 2011. GBOOST: a GPU-based tool for detecting gene-gene interactions in genome-wide case control studies. Bioinformatics 27, 1309–1310.

Zhu, H., Khondker, Z., Lu, Z., Ibrahim, J.G., Alzheimer's Disease Neuroimaging Initiative, 2014. Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. J. Am. Stat. Assoc. 109, 997–990.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. B 67, 301–320.