INVITED PAPER

Brain Imaging Genomics: Integrated Analysis and Machine Learning

This article describes applications of novel and traditional data-science methods to the study of brain imaging genomics. There is a discussion as to how researchers combine diverse types of high-volume data sets, which include multimodal and longitudinal neuroimaging data and high-throughput genomic data with clinical information and patient history, to develop a phenotypic and environmental basis for predicting human brain function and behavior.

By Li Shen, *Senior Member IEEE*, and Paul M. Thompson

ABSTRACT | Brain imaging genomics is an emerging data science field, where integrated analysis of brain imaging and genomics data, often combined with other biomarker, clinical, and environmental data, is performed to gain new insights into the phenotypic, genetic, and molecular characteristics of the brain as well as their impact on normal and disordered brain function and behavior. It has enormous potential to contribute significantly to biomedical discoveries in brain science. Given the increasingly important role of statistical and machine learning in biomedicine and rapidly growing literature in brain imaging genomics, we provide an up-to-date and comprehensive review of statistical and machine learning methods for brain imaging genomics, as well as a practical discussion on method selection for various biomedical applications.

Manuscript received June 1, 2019; revised August 28, 2019; accepted October 8, 2019. Date of publication October 29, 2019; date of current version December 26, 2019. This work was supported under Grant P41 EB015922, Grant RF1 AG041915, Grant U01 AG024904, Grant RF1 AG051710, Grant R21 AG056782, and Grant P01 AG026572. The work of L. Shen was supported in part by NIH and NSF under Grant R01 EB022574, Grant R01 LM011360, Grant RF1 AG063481, and Grant IIS 1837964. The work of P. M. Thompson was supported in part by the NIH Big Data to Knowledge (BD2K) Program under Consortium Grant U54 EB020403, in part by the ENIGMA World Aging Center under Grant NIA R56 AG058854, in part by the ENIGMA Sex Differences Initiative under Grant R01 MH116147, and a grant to the ENIGMA-PGC PTSD Working Group under Grant R01 MH111671. *(Corresponding author: Li Shen.)*

L. Shen is with the Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: li.shen@pennmedicine.upenn.edu).

P. M. Thompson is with the Imaging Genetics Center, Mark and Mary Stevens Institute for Neuroimaging and Informatics, Keck School of Medicine, University of Southern California, Los Angeles, CA 90232 USA (e-mail: pthomp@usc.edu).

Digital Object Identifier 10.1109/JPROC.2019.2947272

KEYWORDS | Big data; brain imaging; genomics; machine learning; statistics.

I. INTRODUCTION

With recent technological advances in acquiring multimodal brain imaging data and high-throughput genomics data, brain imaging genomics is emerging as a rapidly growing research field. It performs integrative studies that analyze genetic variations, such as single nucleotide polymorphisms (SNPs), as well as epigenetic and copy number variations (CNVs), molecular features captured by various omics data, and brain imaging quantitative traits (QTs), coupled with other biomarker, clinical, and environmental data. The goal of imaging genomics is to gain new insights into the phenotypic characteristics and the genetic and molecular mechanisms of the brain, as well as their impact on normal and disordered brain function and behavior. Given the unprecedented scale and complexity of the brain imaging genomics data sets, major computational and statistical challenges have to be met to realize the full potential of these valuable data. Overcoming these challenges has become a major and active research topic in the field of statistical and machine learning, where effective and efficient data analytic methods have been developed to reveal the genetic and molecular underpinnings of neurobiological systems, which can impact the development of diagnostic, therapeutic, and preventative approaches for complex brain disorders.

Many advances in brain imaging genomics are attributed to large-scale landmark studies, such as the Alzheimer's

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

Disease Neuroimaging Initiative (ADNI) [1], the Enhancing Neuro Imaging Genetics through Meta Analysis (ENIGMA) Consortium [2], and the UK Biobank [3]. These studies facilitate the availability of big brain imaging genomics data to the worldwide research community, which contributes to the generation of a large body of literature concerning methodological developments and biomedical applications in brain imaging genomics, including a number of review articles summarizing relevant advances from multiple different perspectives.

For example, ADNI is a landmark Alzheimer's disease (AD) biomarker study. The ADNI cohort constitutes a very rich repository of multimodal data such as genome-wide genotyping, whole genome sequencing, blood transcriptome, blood epigenome, plasma/serum /cerebrospinal-fluid proteome, plasma/serum metabolome, neuroimaging such as multimodal magnetic resonance imaging (MRI) and positron emission tomography (PET), cognitive, behavioral, and clinical data. Due to its open-science nature, data from ADNI have been widely used by the research community around the world to produce hundreds of publications in brain imaging genomics. These advances were periodically reviewed by the ADNI Genetics Core [4], [5] and the entire ADNI team [1], [6].

ENIGMA is another major initiative that contributes significantly to the field of brain imaging genomics. The ENIGMA Consortium is a global team science effort with the shared goal of understanding disease and genetic influences on the brain. The progress of the ENIGMA Consortium has been regularly summarized in several review articles over the years (see [2] and [7]–[9]). Thompson *et al.* [2] provided the most recent update of the ENIGMA Consortium, which included over 1400 scientists from 43 countries studying the human brain using imaging, genomics, and other brain metrics.

The UK Biobank [3], a prospective epidemiological cohort of over 500 000 individuals, is another prominent study that offers an enormous amount of brain imaging genomics data. It has a full genetic data release for $\sim$500 000 samples [10] and full brain imaging data release for $\sim$15 000 samples in six modalities [11]. The team completed large-scale genome-wide association studies of brain imaging QTs recently, which examined $>$11 million SNPs on 3144 imaging QTs in 8428 samples for discovery and two additional sets of 930 and 3456 samples for replication [12]. This article represents the current frontiers in large-scale brain imaging genomics, yielding invaluable insights into the genetic architecture of the brain.

In addition to ADNI, ENIGMA, and UK Biobank, there are many other research activities in brain imaging genomics, which have yielded various review articles. For example, Liu and Calhoun [13] reviewed multivariate methods for analyzing and integrating imaging and genetics data. Yan *et al.* [14] reviewed regression and correlation methods for brain imaging genomics as well as set-based methods for mining high-level imaging genomics associations. Mufford *et al.* [15] reviewed methods and topics of brain imaging genomics in psychiatry. Liu *et al.* [16] reviewed multimodal analysis strategies for analyzing and integrating multiomics data and brain imaging data in the context of schizophrenia studies.

In short, the comprehensive reviews discussed earlier cover topics in brain imaging genomics from different perspectives. Some of them focus on reviewing data, methods, analyses, and/or results from a specific study, such as ADNI [1], [4]–[6] or ENIGMA [2], [7]–[9]. Some reviews examine the research activities and progress in the context of a specific discipline (i.e., psychiatry in [15]) or disorder (i.e., schizophrenia in [16]). Others provide methodology-oriented reviews on multivariate analyses [13] and machine learning [14]. Given that statistical and machine learning is playing increasingly important roles in biomedical research and new methods are emerging in the literature at a rapid pace [17], we feel that it will be valuable to provide an updated review on the topic of statistical and machine learning in brain imaging genomics. Thus, the goal of this article is to provide an up-to-date and comprehensive coverage of statistical and machine learning methods for solving problems in brain imaging genomics as well as practical discussion on method selection for various biomedical applications.

Fig. 1 shows the schematic of the topics that we will cover in this review. The major part of this article will be devoted to the discussion of methods for solving the following three types of learning problems in brain imaging genomics [see Fig. 1(a)].

1) First, we will examine the problem of heritability estimation of brain imaging phenotypes in Section II, where the goal is to determine how much phenotypic variation is determined by genetics.
2) Second, we will explore the problem of learning imaging genomics associations. Since a majority of articles reviewed here belong to this category, we will devote Sections III–VI to this topic. We will review a few fundamental strategies in Section III, including SNP-based methods, polygenic risk scores (PRSs), multi-SNP methods, multitrait methods, pathway and network enrichment methods, and interaction methods. We will discuss meta-analysis strategies in Section IV. We will review multivariate regression models in Section V and bimultivariate correlation models in Section VI to identify complex multi-SNP-multitrait associations.
3) Third, in Section VII, we will review methods for predicting an outcome of interest by integrating imaging and genomics data, as well as methods for joint association learning and outcome prediction.

Finally, in Section VIII, we will provide: 1) a discussion of principles of method selection based on biomedical application considerations [see Fig. 1(b)] and statistical and machine learning considerations [see Fig. 1(c)]; 2) a

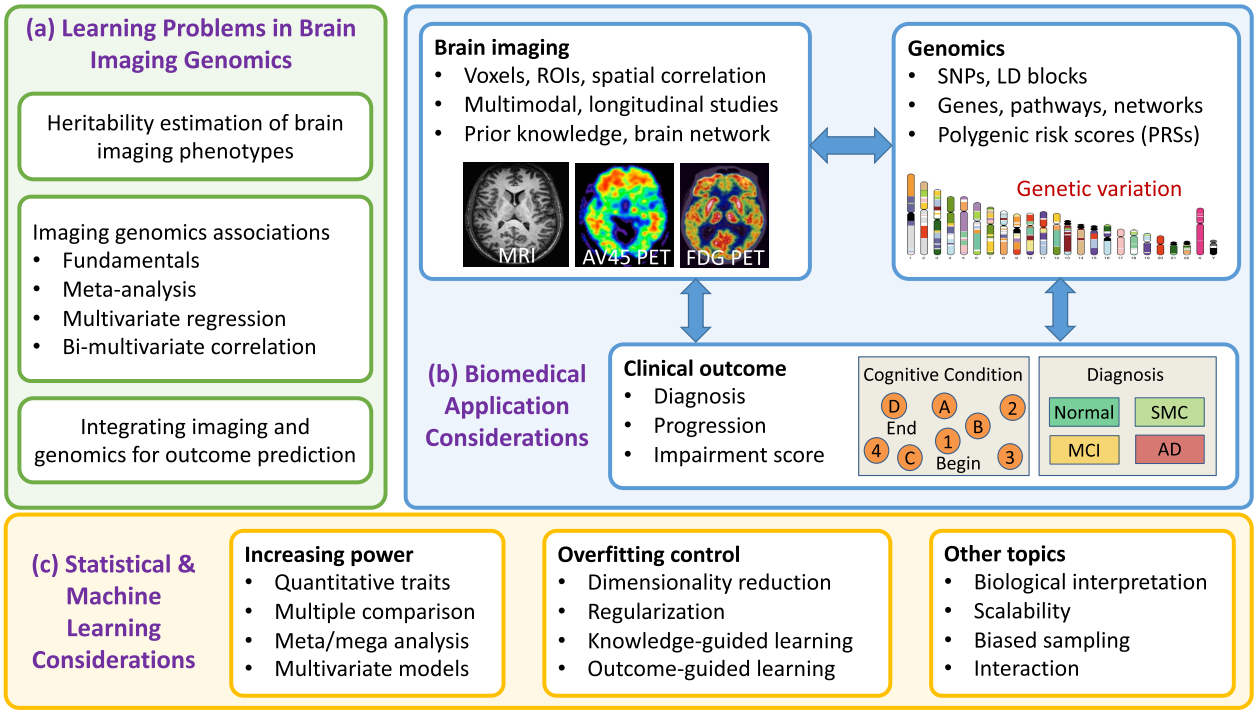


Fig. 1. ***Schematic of topics covered in this review. (a) Learning problems in brain imaging genomics: This review is organized by these topics. (b) Biomedical application considerations: These are example topics related to the studied brain imaging, genomics, and outcome data. (c) Statistical and machine learning considerations: These are example topics considered by the reviewed statistical and machine learning methods.***

discussion on scientific and clinical impact; and 3) a discussion on related work and future directions.

II. HERITABILITY ESTIMATION

Early genetic studies of the brain largely focused on estimating heritability—the proportion of the observed variance in a trait that is explained by additive genetic factors [18]. Well before quantitative genetics was applied to neuroimaging data, classical genetic methods were developed to estimate the proportion of variance in a trait that was due to genetic and environmental factors—as well as random variation, such as measurement errors. The motivation to estimate heritability was that a highly heritable trait might be an attractive target for in-depth genetic analyses compared with a trait with little or no genetic variance. In the following, we cover methods to estimate heritability based on genome-wide genotyping data. First, we note that heritability can be estimated based on data collected using twin or family designs, where the degree of genetic influence is estimated from trait correlations in relative to different degrees of genetic overlap.

A. Twin and Pedigree Methods

Around 2001, neuroimaging studies of twins began to report correlations in regional brain measures in identical and fraternal twins, whereby identical twins had more similar brain structure than randomly selected pairs of individuals of the same age and sex. According to the classical quantitative genetics, if the intraclass correlation is higher in monozygotic (MZ) than dizygotic (DZ) twins, then a trait is heritable. Falconers heritability statistic, h^2, is defined as twice the difference between the MZ and DZ intraclass correlations. Thompson *et al.* [19] reported the statistical maps of Falconers h^2 statistics, for measures of gray matter density across the cortex, showing significant heritability, in a small MRI study of 80 young adult twins. Later studies built on this approach to fit structural equation models (SEMs) to quantify both genetic and environmental components of variance, for brain measures derived from MRI, diffusion tensor imaging (DTI), electroencephalogram (EEG), and functional MRI (fMRI), also using twin or family designs. A common model used for these studies was the ACE model, which estimates additive genetic (A), common (C), and unique (E) environmental contributions to trait variance (see [20] for a review of early neuroimaging studies using the ACE model).

Brun *et al.* [21], for example, used a general MRI analysis method called tensor-based morphometry (TBM) to map the heritability of brain morphology in MRI scans from 23 monozygotic and 23 dizygotic twin pairs using the ACE genetic model. Significance was tested using voxelwise permutation methods. A similar work with other computational anatomy approaches extended the ACE model to scalar maps defined on the vertices of 3-D surface models of brain structures, such as the ventri-

cles [22]. In that study, path coefficients for the ACE model that best fit the data indicated significant contributions from genetic factors (A = 7.3%), common environment (C = 38.9%), and unique environment (E = 53.8%) to lateral ventricular volume.

Extending the ACE model to diffusion MRI, to assess the genetics of brain white matter microstructure, Shen *et al.* [23] confirmed the overall heritability of the major white matter tract metrics but also identified differences in heritability. Highly heritable measures were found for tracts connecting particular cortical regions, such as medial frontal cortices, postcentral, paracentral gyri, and the right hippocampus. Later studies reported genetic correlations between the measures of cortical gray matter thickness and DTI-derived white matter measures [24]. Comparable methods applied to fMRI revealed significant heritability for the measures of functional synchrony in the brains resting-state networks (RSNs). Fu *et al.* [25] estimated both genetic and environmental effects on eight well-characterized RSNs. To do so, they fit the classical ACE twin model to the functional connectivity covariance at each voxel in the RSN. Although environmental effects accounted for the majority of variance in widespread areas, specific brain regions showed significant genetic control within individual RSNs.

Methods to estimate heritability were advanced as well. Open-source tools, such as OpenMx and SOLAR, were adapted to handle brain-derived phenotypes, including entire images. Kochunov *et al.* [26] examined the agreement in the heritability estimates, across a variety of data sets, for four different methods for heritability estimation which have been applied to neuroimaging data. SOLAR-Eclipse (www.solar-eclipse-genetics.org) and OpenMx (openmx.ssri.psu.edu) use iterative maximum-likelihood estimation (MLE) methods. Accelerated permutation inference for ACE (APACE) [27] and fast permutation heritability inference (FPHI) [28] use fast, noniterative approximation-based methods. Heritability estimates from the two MLE approaches closely agreed on both simulated and imaging data, but the two approximation approaches showed lower heritability estimates when running on data that deviated from normality. The authors advocated a data homogenization approach that improved agreement across packages using inverse Gaussian transformation to enforce normality on the input trait data.

B. GWAS Methods for SNP Heritability

As soon as genome-wide genotyping became cheaper and more common, methods were developed to estimate heritability from all genome-wide SNPs. The GCTA method (genome-wide complex trait analysis [29]; https://cnsgenomics.com/software/gcta/), for example, estimates heritability from general population data, and rather than requiring twins or pedigrees, it can be applied to data from individuals who are typically regarded as unrelated. GCTA computes both genetic and phenotypic covariance matrices from trait data and high-density SNP data, after calculating a kinship matrix and a genotypic relatedness matrix (GRM). Based on singular values of the GRM, GCTA estimates the percentage of phenotypic variance explained by all common SNPs (i.e., the SNP heritability of a trait), with a restricted maximum-likelihood linear mixed model (GREML). GCTA has been used to estimate "missing" heritability—the genetic contribution from all SNPs in aggregate—without needing to know exactly which SNPs are contributing to the variance.

Direct application of GCTA to the heritability analysis of high-dimensional brain imaging QTs is computationally intractable. To overcome this limitation, Ge *et al.* [30] proposed a massively expedited genome-wide heritability analysis (MEGHA) method, which approximates GCTA and is suitable for analyzing a large number of phenotypes efficiently. It was successfully used to create vertexwise heritability mapping of nearly 300 000 cortical thickness QTs. Ge *et al.* [31] proposed a moment matching method for SNP-based heritability estimation (MMHE) and further extended the GWAS-based heritabilty analysis to handle multidimensional traits (e.g., shape). It was successfully applied to the heritability estimation of the shape of a set of brain structures. In a subsequent study [32], MMHE was used to complete a phenome-wide heritability analysis of the UK Biobank [3].

A related method—linkage disequilibrium score regression (LDSC) [33]—was also developed to estimate heritability due to all SNPs. Remarkably, it does not require individual genotypes at all, but it only uses the summary statistics from a genome-wide association study. The approach exploits a feature of the genome called LD—the fact is that statistical correlations are found in a series of adjacent SNPs. Let N be the sample size, M be the number of all SNPs, and h^2 be the heritability of a phenotype due to all SNPs. Given an SNP j, its LD Score l_j is defined as $l_j = \Sigma_{k=1}^{M} r_{jk}^2$, where r_{jk}^2 is the LD between SNPs j and k measured by the squared correlation coefficient. The LD Score l_j measures the amount of genetic variation tagged by j. Bulik-Sullivan *et al.* [33] noted that under a polygenic model, the expected χ^2 association statistics for SNP j are

$$E[\chi^2|l_j] = Nh^2\, l_j/M + Na + 1$$

where h^2/M is the average heritability explained per SNP and a measures the contribution of confounding biases, such as cryptic relatedness and population stratification. Based on this, if one regresses the χ^2 statistics from GWAS against LD Score (i.e., LDSC), the resulting intercept minus one can serve as an estimator of the mean contribution of confounding bias to the inflated test statistics. Consequently, LDSC can also be used to produce SNP-based heritability estimates for any phenotypes, including voxel- or region-based imaging QTs, partition

this heritability into separate categories (based on regions of the genome, such as specific chromosomes or types of genetic variant), and to calculate genetic correlations between separate phenotypes.

When applied to imaging GWAS (explained next), the LDSC method has revealed patterns of genetic correlations across brain regions, leading to the notion that the brain may be partitioned into genetic modules or sets of regions with overlapping genetic determinants. Classical multivariate twin models had also reported evidence for such genetic clusters [34]. In [34], a multivariate model in 1038 twins identified a common genetic factor that accounted for almost all the heritability of intracranial volume (0.88) and a substantial proportion of the heritability of all subcortical structures, particularly those of the thalamus (0.71 out of 0.88), pallidum (0.52 out of 0.75), and putamen (0.43 out of 0.89). LDSC has also been used to reveal an overlap between the genetic loci associated with brain structure and with schizophrenia based on the summary statistics from various published GWAS [35]. Similar multivariate genetic models show that genetic influences on longitudinal growth or loss rates over time significantly overlap with genetic loci associated with baseline volumes for many structures. This may be an important observation in the quest to identify loci that influence rates of brain development and degeneration [36].

III. IMAGING GENOMICS ASSOCIATIONS: FUNDAMENTALS

Given an imaging phenotype, heritability analysis estimates how much of its variance is explained by the entire genome or all the SNPs on one or more chromosomes. In order to locate specific genetic variants that contribute to the phenotypic change, a genetic association analysis needs to be performed. Thus, a major research theme in brain imaging genomics is how to effectively identify interesting imaging genomics associations, which is the topic to be covered in Sections III–VI. In some cases, heritability analysis can be used as a prescreening step to identify imaging QTs with moderate to high heritability, and subsequent genetic association studies can then be applied only to those heritable QTs (e.g., in [38]).

A major challenge in brain imaging genomics is that both imaging and genomics data are high dimensional. The ability to test over a million SNPs in the genome for associations with hundreds, thousands, or even more imaging traits in the brain induces a huge burden for multiple comparison correction. While failure to properly correct for multiple comparisons leads to a high risk for false discoveries, excessive corrections greatly reduce the power to detect true signals. Thus, multiple comparisons and detection power are two important topics relevant to most association studies reviewed in this article.

Lindquist and Mejia [39] provided an excellent review of a few major statistical approaches to address the problem of multiple comparisons using neuroimaging studies as an example. The goal is to choose an appropriate threshold to balance between sensitivity (true positive rate) and specificity (true negative rate). Two metrics to quantify the likelihood of obtaining false positives are often used: 1) the familywise error rate (FWER; the probability of obtaining at least one false positive in a family of tests) and 2) the false discovery rate (FDR; the proportion of false positives among all rejected tests). Bonferroni correction [40], aiming to control the FWER at a user-specified level, is the most common approach for multiple comparison correction. Despite being simple to use, it is very conservative and often reduces detection power. Random field theory (RFT) [41]—a popular approach for controlling the FWER in fMRI studies—considers the spatial correlation in the images and appears to be less conservative than the Bonferroni method. Permutation methods are nonparametric methods that do not make assumptions on the data distribution for controlling the FWER. While they offer substantial improvements in detection power, especially in small sample sizes, they are very computationally expensive; some recent innovations have been used to accelerate permutation testing [42]. The FDR [43] is a newer approach that controls false positives. It is less stringent than FWER methods and thus has an increased detection power.

While some imaging genomics studies reviewed here employ the above-mentioned methods for multiple comparison correction, others develop their own strategies for handling the issues of multiple comparisons and detection power. For example, Hua *et al.* [44] proposed two strategies to handle multiple comparisons and increase the power of detecting imaging genomics associations. On one hand, they treated the imaging QTs of the entire brain as a single multivariate response and used distance covariance to capture the association between all the QTs and each SNP, which greatly reduced the number of statistical tests. On the other hand, they proposed a new FDR-based algorithm that demonstrated an increased detection power compared with two existing FDR methods.

Another critical challenge in brain imaging genomics is the relatively small effect size of SNPs on the brain. Most SNPs account for under 1% of the variance in a brain QT when considered individually. Thus, the studies reviewed here all need to address this challenge, and many of these studies have aimed to develop effective strategies with increased detection power to capture interesting imaging genomics associations. For example, one strategy is to reduce the effective number of tests to alleviate the burden of multiple comparison correction (see targeted SNP/QT studies discussed in Section III-A). The second strategy is to measure combined or collective effects of multiple markers together to increase the detection power (see studies discussed in Sections III-B–III-E). The third strategy is to increase the sample size to enable the discovery of individual SNPs with small effect sizes (see studies discussed in Section IV). The fourth strategy is to apply a single multivariate model involving all the studied SNPs

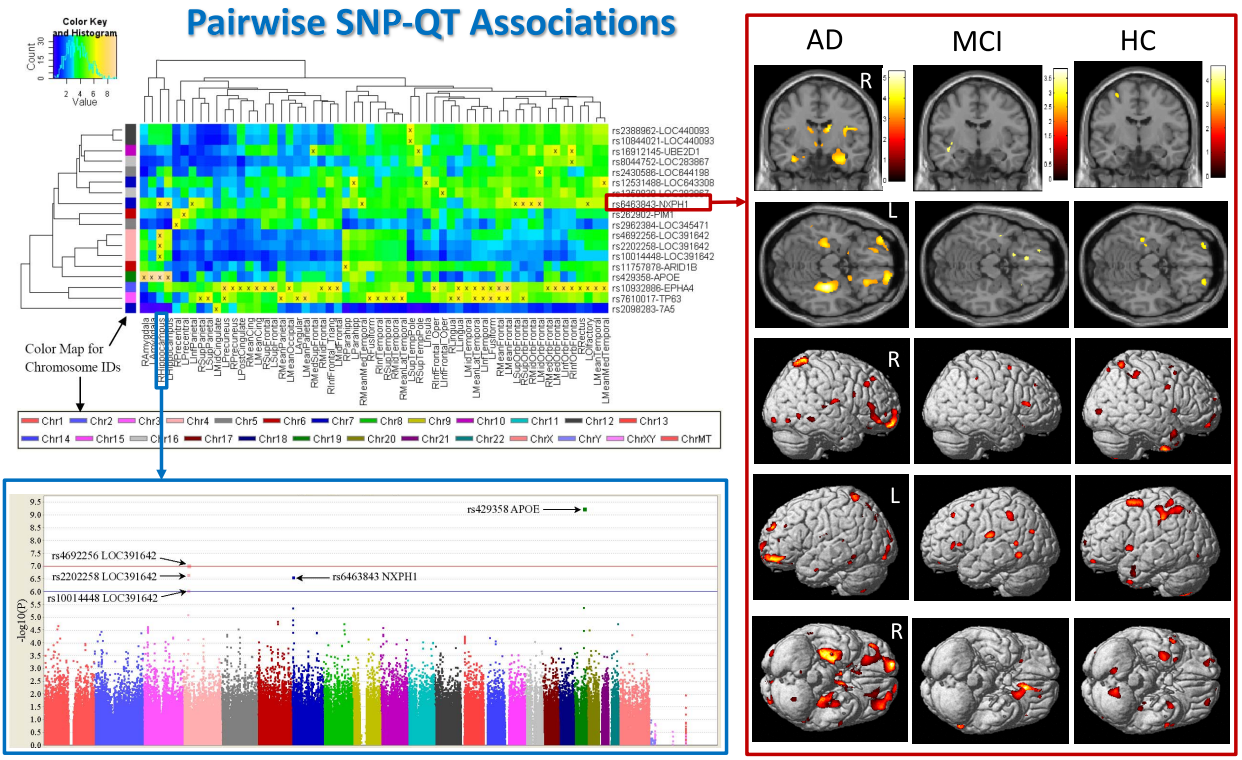


Fig. 2. ***Example pairwise SNP–QT Associations [37]. Top left: All the pairwise SNP–QT association findings, where blocks labeled with "x" reach the level of $p < 10^{-6}$. Bottom left: Manhattan plot for the GWAS results of gray matter density of the right hippocampus (i.e., blue box). Right: Voxel-based morphometry result of mapping the genetic effect of rs6463843 (in the flanking region of the NXPH1 gene) to the brain (i.e., red box). Images are reproduced here with permission from Elsevier [37].***

and QTs without needing to adjust for multiple testing (see studies discussed in Sections V and VI).

Before covering more advanced statistical and machine learning strategies for mining brain imaging genomic associations in Sections IV–VI, we first review a few fundamental methods in this section. We start from the simplest single-SNP–single-QT approaches that search for pairwise imaging genomics associations on an SNP-by-SNP and QT-by-QT basis. Next, we discuss strategies using PRSs, which examine the aggregated effect from a set of disease-related SNPs on an imaging QT. Then, we go over basic multi-SNP or multitrait methods, which aims to learn imaging genomics associations involving either multiple SNPs or multiple traits. After that, we review enrichment analysis methods that intend to discover high-level imaging genomics associations related to biological entities, such as biological pathways, functional interaction networks, and/or brain circuits (BCs). Finally, we briefly discuss interaction methods that focus on the exploration of epistatic effects instead of main effects.

A. Single-SNP–Single-QT Methods

Given a set of genetic markers such as SNPs and a set of imaging QTs, the simplest and most commonly used analytical strategy is to perform a pairwise analysis between each SNP and each QT at the individual marker level. An SNP takes a value of 0, 1, or 2 (i.e., the genotype value), indicating the number of minor alleles at the corresponding chromosome location. An imaging QT typically takes a continuous value. A simple linear regression model can be used to examine the additive effect of the SNP on the imaging QT. An alternative strategy is to use analysis of variance (ANOVA), which is similar to linear regression but ignores the ordering of the genotype values. It examines the trait mean differences among three genotype groups. Both the strategies can be used together with hypothesis testing to obtain a p-value. If multiple pairwise SNP–QT associations are examined, multiple comparison correction needs to be performed to identify significant findings.

Fig. 2 shows three major types of SNP–QT analyses.

1) *Targeted QT Analyses:* The first type is to perform genetic analysis on one or more targeted imaging QTs. For example, in Fig. 2, the bottom left (i.e., blue box) shows the Manhattan plot for the GWAS results of gray matter density of the right hippocampus.
2) *Targeted SNP Analyses:* The second type is to examine the genetic effects of one or more SNPs on all the imaging QTs across the brain. For example, in Fig. 2, the right (i.e., red box) shows the voxel-based morphometry (VBM) result of mapping the genetic effect of rs6463843 (in the flanking region of the NXPH1 gene) to the brain.

3) *Brain-Wide Genome-Wide (BWGW) Analyses:* The third type is to perform massive univariate analyses for all the possible SNP–QT pairs across the entire brain and the entire genome. For example, in Fig. 2, the top left summarizes all the pairwise SNP–QT association findings (only top findings are shown), where blocks labeled with "x" reach the level of $p < 10^{-6}$. Note that, in [37], $p < 10^{-6}$ was explored as a somewhat less stringent threshold to identify imaging genomics associations showing a trend toward significance as well as examine clustering patterns of the corresponding SNP and imaging QT findings.

In the following, we discuss a few example studies in each of these three categories.

In one targeted QT study, Stein *et al.* [45] performed a genome-wide association study of the bilateral temporal lobe volume (TLV) as the QT. A linear regression analysis was conducted at each SNP to examine its genetic effect on the QT and covaried for age and sex. In another targeted QT study, Scelsi *et al.* [46] computed a novel disease progression score (DPS) from multimodal neuroimaging data and performed GWAS on it. The DPS was generated by the GRACE algorithm [47] from the longitudinal cortical amyloid burden and bilateral hippocampal volume, providing an estimate of how advanced an individual's disease progression is in comparison with the cohort average. A linear regression analysis was conducted at each SNP to examine its genetic effect on the DPS and covaried for sex, age at first amyloid scan, education, two principal components of population structure, and number of APOE e4 alleles.

In one targeted SNP study, Risacher *et al.* [48] examined the effect of the APOE e4 SNP rs429358 on several MRI and PET imaging QTs. Specifically, the effects of diagnosis, APOE e4 carrier status, and their interaction on regional amyloid deposition, regional glucose metabolism, hippocampal volume, and entorhinal cortex thickness were examined using a two-way analysis of covariance (ANCOVA) and covaried for age and gender. In another targeted SNP study, Ho *et al.* [49] examined the effect of a commonly carried allele of the obesity-related FTO gene on regional brain volume measures captured by MRI. Specifically, the general linear model was used to evaluate the relation of the imaging QT at each voxel to the SNP rs3751812 controlling for age and sex.

In one BWGW study, Shen *et al.* [37] used a BWGW approach to investigate genetic effects on imaging QTs. The studied QTs included 56 volumetric and cortical thickness measures and 86 local gray matter density values for regions of interests (ROIs) across the entire brain. These imaging QTs were preadjusted to remove the effects of age, gender, education, handedness, and incracranial volume (ICV). A linear regression analysis was conducted at each SNP to examine its genetic effect on each QT. In another BWGW study, Stein *et al.* [57] performed the first voxel-based GWAS analysis. Using TBM to define imaging QTs, they examined genome-wide association at each voxel. A linear regression analysis was conducted at each SNP-by-voxel pair to examine the SNP genetic effect on each voxelwise QT and covaried for age and sex.

Although a voxelwise GWAS enables the examination of imaging genomics associations at the finest resolution, it is facing a major computational challenge, given the huge number of univariate SNP–QT associations to test. To overcome this challenge, Huang *et al.* [50] proposed a fast voxelwise GWAS (FVGWAS) framework to facilitate efficient BWGW study at the voxel level. FVGWAS employs three components to achieve this goal. The first component is a heteroscedastic linear model that allows a very flexible covariance structure suitable for voxelwise imaging QTs. The second component is a global sure independence screening (GSIS) procedure [51] that can greatly reduce the search space size from $N_s N_v$ to $\sim N_0 N_v$ for $N_0 \ll N_s$. Here, N_s is the number of SNPs and N_v is the number of voxels. The third component is a detection procedure based on wild bootstrap methods which is computationally cheap due to no involvement of repeated analyses of simulated data sets. As a result, for standard linear association, the computational complexity of FVGWAS is $O((N_s + N_v)n^2)$, outperforming $O(nN_vN_s)$ for standard voxelwise GWAS [45], where n is the number of subjects. FVGWAS is available at https://www.nitrc.org/projects/fvgwas/.

One issue related to imaging genomics is that most GWAS studies (e.g., ADNI) are based on a case-control design, and the data are typically a biased sample of the target population. Directly correlating imaging QTs (as secondary traits) with genotype may lead to biased inference generating misleading results. Kim *et al.* [52] compared the standard linear regression model and disease status adjusted linear model with two models adjusting for biased case-control sample (i.e., inverse probability weighted regression [53] and retrospective likelihood [54]) on the analysis of ADNI data. Zhu *et al.* [55] completed a similar systematic evaluation of the biased sampling issue using both simulation and ADNI data. They compared the standard linear regression model and disease status adjusted linear model with two models adjusting for biased case-control sample (i.e., retrospective likelihood [54] and reparameterization of conditional model in [56]). Although the standard linear analysis was found to be generally valid on the ADNI data in [52], simulation studies in [55] showed that linear regression models without adjusting for biased sampling demonstrated severely inflated Type I error rates in some cases. In general, caution should be taken while analyzing imaging QT data as secondary phenotypes in case-control studies.

Table 1 summarizes the studies discussed earlier, where pairwise SNP–QT associations are examined on an SNP-by-SNP and QT-by-QT basis. These single-SNP–single-QT methods are simple and straightforward. The findings discovered by these methods are easy to interpret since each resulting association involves only one SNP and one

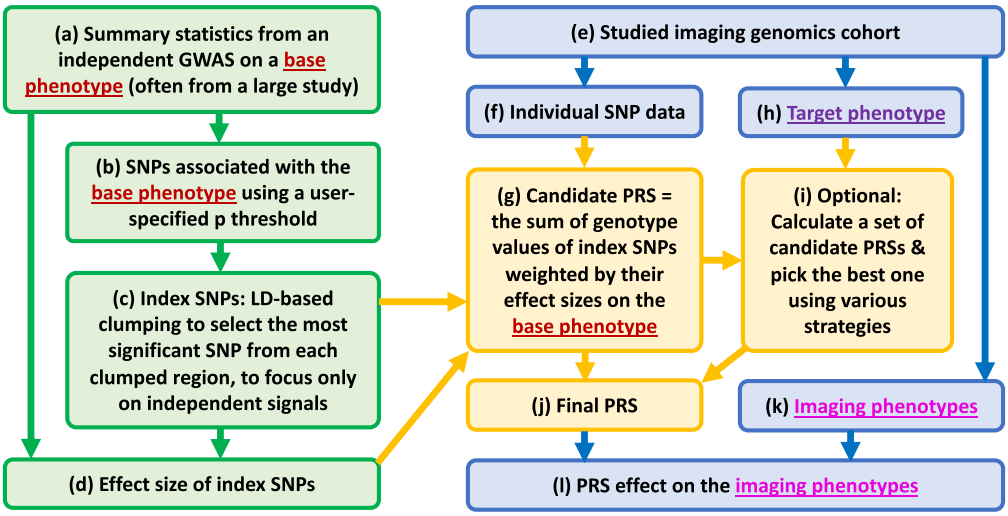


Fig. 3. ***Example flowchart to calculate a PRS and apply it to brain imaging genomics studies. Step (i) is optional, where various strategies can be used to calculate a set of candidate PRSs (e.g., by exploring a few p thresholds [61], [62]) and pick the PRS best associated with the target phenotype [see (h)] as the final PRS [see (j)]. See the main text for more details.***

Table 1 Example Studies Using Single-SNP–Single-QT Methods, Where Pairwise SNP–QT Associations Are Examined on an SNP-by-SNP and QT-by-QT Basis

Ref	Notes
[45]	GWAS, targeted QT, linear regression
[46]	GWAS, targeted QT, linear regression
[48]	Targeted SNP, targeted QTs, two-way ANCOVA
[49]	Targeted SNP, voxelwise QTs across brain, general linear model
[37]	GWAS, ROI-based QTs across brain, linear regression
[57]	GWAS, voxelwise QTs across brain, linear regression
[50]	Fast voxelwise GWAS (FVGWAS): heteroscedastic linear model, global sure independence screening, wild bootstrap
[52]	Regression models for analyzing secondary phenotypes
[55]	Regression models for analyzing secondary phenotypes

QT. Given the high dimensionality of both imaging and genomic data, studies examining a massive number of SNP–QT associations may face major computational and statistical challenges. In addition, multivariate associations involving multiple SNPs or multiple QTs will not be able to be identified by these methods.

B. Polygenic Risk Scores

One approach to identify imaging genomics associations involving multiple SNPs is to use a PRS [58]. A PRS captures the aggregate genetic effect from a set of trait-related SNPs that may not achieve significance at the individual level but collectively may explain a substantial portion of the trait variance. It is often calculated as the sum of their genotype values weighted by their effect sizes on a base phenotype (e.g., case-control status). Dima and Breen [59] reviewed the usefulness and applications of PRSs in imaging genetics. Chasioti *et al.* [60] reviewed recent progress in PRS in AD and other complex disorders. The cohorts with both brain imaging and genetics data are often much smaller than those designed for large GWAS. A PRS can typically be calculated based on using the SNP-based effect sizes from large GWAS on a base diagnostic phenotype to make full use of the power of the large sample. After that, it can be applied to small samples with imaging data to examine its association with interesting imaging QTs.

Fig. 3 shows an example flowchart to calculate a PRS and apply it to brain imaging genomics studies. First, using the summary statistics from an independent GWAS (often a large-scale landmark study) on a base phenotype [see Fig. 3(a)], a set of SNPs associated with the base phenotype can be obtained using a user-specified p threshold [see Fig. 3(b)]. Second, LD clumping is often performed to select the most significant SNP from each clumped region to form a set of independent loci named index SNPs [see Fig. 3(c)]. Third, using the effect sizes of index SNPs from the summary statistics data [Fig. 3(d)] and individual SNP data [see Fig. 3(f)] from the studied imaging genomics cohort [see Fig. 3(e)], one can calculate a PRS that is the sum of genotype values of index SNPs weighted by their effect sizes on the base phenotype [see Fig. 3(g)]. While this PRS can directly be used, some studies (see [61] and [62]) perform an optional step [see Fig. 3(i)] to calculate a set of candidate PRSs by exploring a few p thresholds and then pick the PRS best predicting the target phenotype [see Fig. 3(h)] as the final PRS using several strategies described next. Finally, the effect of the resulting PRS on interesting imaging phenotypes can be examined [see Fig. 3(l)].

Scelsi *et al.* [46] performed a PRS study on a novel image-based DPS discussed in Section III-A, using a workflow similar to that shown in Fig. 3. They obtained index SNPs and their effect sizes using the large AD GWAS conducted by the International Genomics of Alzheimer's Project (IGAP) [63]. Instead of computing one PRS, they calculated 15 PRSs by exploring 15 p thresholds in the

Table 2 Example Studies Using PRSs for Brain Imaging Genomics. A PRS Summarizes the Aggregate Effect From an Ensemble of SNPs Related to a Base Phenotype. The Effect of the PRS Is Examined on Interesting Imaging QTs

Ref	Notes
[58]	PRS: Power and predictive accuracy
[59]	PRS: Usefulness and applications in imaging genetics
[46]	Standard PRS workflow, image-based DPS
[61]	Standard PRS workflow, hippocampal volume
[62]	Standard PRS workflow, cortical thickness
[65]	PHS instead of PRS, amyloid and MR imaging QTs
[67]	PRSice: PRS software, http://prsice.info/

range of 0.95–10^{-5}. They identified only one PRS with p threshold of 10^{-4}, which is significantly associated with the image-based DPS.

Mormino *et al.* [61] performed a PRS study on MRI-derived hippocampal volume using the workflow shown in Fig. 3. They used the IGAP GWAS summary statistics to obtain the index SNPs and their effect sizes. They explored a dozen p thresholds ranging from 5×10^{-8} to 0.05 to generate multiple PRSs. The final PRS was selected as the one best differentiating clinically normal (CN) and AD participants in ADNI-1 sample. This PRS was found to be associated with hippocampal volume for ADNI-1 sample without dementia.

Sabuncu *et al.* [62] performed a PRS study on cortical thickness measures. They used the summary statistics from another large-scale GWAS in AD [64] to obtain the index SNPs and their effect sizes. They further screened these SNPs using five different thresholds based on the genetic association results on a subset of ADNI data containing only CN and AD participants to create five different PRSs. The PRS with the highest correlation with Mini-Mental State Examination (MMSE) score and Clinical Dementia Rating Sum of Box (CDR-SB) score and strongest association with AD diagnosis was used in the subsequent imaging genomic analyses on a nonoverlaping ADNI sample containing only CN subjects. This PRS was identified to be associated with AD-specific cortical thickness.

Tan *et al.* [65] studied a similar problem on developing a polygenic hazard score (PHS) instead of PRS. They used the IGAP GWAS summary statistics to identify a set of SNPs with $p < 10^{-5}$. They evaluated these SNPs using the Alzheimer's Disease Genetics Consortium (ADGC) Phase 1 data. Using a stepwise Cox proportional hazards model, they identified 31 top SNPs and formed a PHS [66]. This PHS was applied to the ADNI data and found to be associated with ADNI imaging phenotypes, such as regional amyloid burden using Amyloid-PET and regional volume loss using MRI.

Euesden *et al.* [67] presented PRSice, a software tool for generating PRS. It takes GWAS summary statistics on a base phenotype and genotype data on a target phenotype and returns a PRS for each individual. It calculates PRS at multiple p thresholds and can select the most predictive one. The software is available at http://prsice.info/.

Table 2 summarizes the studies described earlier. A PRS captures the aggregate effect from an ensemble of SNPs related to a base phenotype. In disease-relevant brain imaging genomics studies, examining the effect of a PRS instead of each individual SNP on imaging QTs has great potential to increase statistical power as well as gain meaningful insights into the biological mechanism from genetic determinants to brain endophenotypes and to disease status. However, there is also some discussion in recent literature regarding potential limitations in PRS-based analyses. For example, bias toward the reference population was observed in [68]. Specifically, the generalizability of a PRS across different populations appeared to be limited. Greater diversity should be prioritized to realize the full potential of PRS. In addition, statistical power differences across diseases and cohorts were also observed in [69]. Several factors could limit the power of a PRS. One factor could be the cohort difference between the base and the target GWAS. Another factor could be limited sample sizes of available data for certain diseases, in particular for heterogeneous disorders that can be stratified into different subtypes with even smaller sample size in each group.

C. Multi-SNP Methods

A single-SNP analysis is often limited by the modest SNP effect sizes. Multi-SNP methods examine a joint effect from a set of SNPs on a phenotypic trait. It has enormous potential to improve the power of genetic association studies and identify polygenic or multilocus mechanisms for complex diseases. There are several categories of multi-SNP analysis strategies. The first category focuses on the joint analysis of a set of targeted SNPs based on the prior knowledge. For example, one approach is to analyze a PRS involving top SNPs from an independent GWAS, as previously described in Section III-B. Another approach is to analyze a set of disease-related SNPs from the literature (see [70]). The second category is to perform GWAS at the gene level instead of the SNP level, where the aggregate effect of all the SNPs within each gene on the target phenotype is examined to increase statistical power (see [71]–[73]). The third category employs the data-driven strategies to automatically identify relevant SNPs from either the entire genome or a set of candidate SNPs [74]. In the following, we discuss a few example studies using these strategies. Section VI will cover additional studies using the third category of strategies.

Apostolova *et al.* [70] examined the top 20 AD SNPs and their joint effect with brain amyloidosis in an ADNI sample, including 322 CN, 496 mild cognitive impairments (MCIs), and 159 AD participants. Stepwise multivariate linear regression was used to examine the association between joint exposure of 20 AD risk alleles and mean amyloid burden from florbetapir PET scans while controlling for age, sex, and APOE e4 status. Voxelwise 3-D stepwise regression was also used to map the genetic effect onto the brain. The study identified an association between several AD SNPs and brain amyloidosis.

Hibar *et al.* [71] extended the SNP-based voxelwise GWAS (vGWAS) method [45] to a gene-based voxelwise

GWAS (vGeneWAS) method. It was demonstrated on a BWGW study using the same ADNI sample. The joint effect of SNPs within each gene on each voxel was examined using a multiple partial-F test while controlling for age and sex. To address the SNP colinearity issue, a principal component analysis (PCA) was performed on the SNPs within each gene. The "eigenSNPs" capturing the first 95% data variance were then used in the multiple partial-F tests. This method can be thought of as a variant of principal component regression (PCReg) [75].

Ge *et al.* [72] further extended vGWAS and vGeneWAS to a new SNP-based GWAS or vGeneWAS framework with increased power and demonstrated it on a BWGW study using the same ADNI sample. This method includes three new methodological contributions. The first one is a fast implementation of voxelwise and clusterwise inferences using an RFT to improve statistical power via embracing the spatial correlation in the images. The second one is a multilocus model based on least-square kernel machines (LSKMs) to evaluate the joint effect of multiple SNPs within each gene on each voxelwise QT. The multilocus method employs a semiparametric regression model [76], where the covariate effects on the QT are modeled linearly and parametrically and the SNP effects on the QT are modeled nonparametrically using the LSKM approach. This method allows for revealing nonlinear effects introduced by the interaction among SNPs. The third contribution is a fast permutation procedure that uses a parametric tail approximation to provide accurate p estimations in an efficient manner.

Xu *et al.* [73] proposed a new method called imaging-wide association study (IWAS), which was inspired by transcriptome-wide association study (TWAS) [77]. It aims to integrate imaging QTs with GWAS to improve statistical power and biological interpretation. It is a gene-based approach and has two steps involving two sets of GWAS data, respectively: 1) the reference GWAS data containing imaging QTs and 2) the main GWAS data containing target phenotype such as disease status. In the first step, which analyzes the reference GWAS data, for each gene, IWAS estimates a set of SNP weights via regressing an imaging QT on all the SNPs. In other words, it builds a prediction model for the genetic component of the imaging QT. In the second step, which analyzes the main GWAS data, IWAS uses the weights learned in the first step to calculate a weighted genotype score for each gene and examines its association with the target phenotype. Using the strategies described in [78] and [79], IWAS can also be applied to the main GWAS data containing only summary statistics. In short, IWAS uses an imaging QT to construct weights for a weighted gene-based GWAS test. The gene-based method reduces the number of tests and boosts statistical power. Also, computing gene scores via extracting genetically controlled components of an imaging QT provides potential opportunities to help interpret GWAS findings.

The above-mentioned studies developed or employed methods to examine the association between one SNP set and one QT. Lu *et al.* [74] proposed a method for examining joint association mapping between a large number (e.g., 10^5) of SNP sets and a QT. Here, the SNP sets can be defined by LD blocks or genes so that multiple SNPs can be combined to increase detection power. A linear mixed-effects model was proposed to simultaneously regress a QT on a large number of SNP sets. This model has the potential to further increase detection power via: 1) incorporating the correlation among SNP sets and 2) greatly reducing the burden of multiple comparison correction. A Bayesian latent variable selection procedure was proposed to select significant latent variables. An efficient Markov chain Monte Carlo (MCMC) algorithm was proposed to reduce the complexity of major computationally intensive steps in MCMC iterations. The empirical studies were performed on the ADNI sample to identify associations between a few imaging QTs and a number of SNP sets defined by LD blocks and genes, and yielded promising results.

Table 3 Example Studies Using Multi-SNP Methods, Where Multi-SNP–Single-QT Associations Are Examined

Ref	Notes
[70]	Joint effect of target SNPs on imaging QTs, stepwise multivariable linear regression
[71]	Multivariate gene-based voxelwise GWAS, PCA and multiple partial F test (a variant of PCReg)
[72]	Voxelwise GWAS with increased power, random field theory, semi-parametric regression model with least square kernel machines, fast permutation procedure
[73]	Imaging wide association study, weighted gene-based GWAS test, weights capturing genetic component of an imaging QT
[74]	Joint association between multiple SNP sets and an imaging QT, linear mixed-effects model, Bayesian latent variable selection

Table 3 summarizes the studies described earlier, which are designed to identify multi-SNP–single-QT associations. Compared with single-SNP methods, examining the joint effect of an SNP set on an imaging QT can potentially increase statistical power and identify multilocus or polygenic mechanisms for complex brain phenotype. In addition, the SNP sets are often defined by LD blocks, genes, pathways, known trait-associated variants, or other prior knowledge, which may offer meaningful biological insights for interpreting multi-SNP discoveries.

D. Multitrait Methods

Similar to multi-SNP methods, multitrait methods provide an alternative means to increase detection power compared with single-SNP–single-trait analyses. There are several classical strategies to perform multivariate trait analysis, as nicely summarized in [80]. One approach is to first conduct univariate analysis on each trait and then combine their results [81]. For example, a typical strategy is to select the SNP with the minimum p-value with multiple comparison correction. The second approach is to perform dimensionality reduction on the traits and then apply univariate analysis on a small number of extracted trait features. These features could simply be the average trait or first a few components from PCA [82] or

canonical correlation analysis (CCA). The third approach is to employ classical multivariate analysis methods, such as multivariate analysis of variance (MANOVA) [83] and generalized least squares (GLS) [84], [85]. In the following, we discuss a few recently proposed methods for performing multitrait analyses in brain imaging genomics.

Zhang *et al.* [80] proposed a set of new testing methods for identifying single-SNP–multi-QT associations under the framework of generalized estimation equations (GEEs) [86]. They tried to address the challenge that in multi-QT analyses, there is a lack of a uniformly powerful test. For example, given a QT set, if very few QTs are associated with the target SNP, selecting the QT with minimum p from a set of univariate SNP–QT analyses could be more powerful. On the other hand, if most of the QTs are associated with the SNP, doing a univariate analysis between the average QT and the SNP could be more powerful. With this observation, under the GEE framework, they proposed the SPU(γ) tests (i.e., the sum of powered score (U) tests) for a series of values of $\gamma = 1, 2, \ldots, \infty$, where a larger γ value tends to put higher weights on QTs with stronger associations with the SNP. As a result, SPU(∞) corresponds to the minimum p strategy and SPU(1) corresponds to the average QT strategy. Based on this, they also proposed adaptive SPU (aSPU) test. The aSPU test statistic is defined as the minimum p among all the SPU tests $T_{\text{aSPU}} = \min_{\gamma \in \Gamma} P_{\text{SPU}(\gamma)}$. In other words, aSPU was designed to be an adaptive method that automatically performs data-driven weights' adjustment and selects the most powerful weighted test from all these candidates. The empirical study was performed on an ADNI sample to pairwisely associate 20 candidate SNPs to a few imaging QT sets, and the proposed aSPU method outperformed various competing methods.

Kim *et al.* [87] further extended the aSPU test to a new test that can identify associations involving multiple SNPs. While aSPU searches for single-SNP–multi-QT associations, the new test is designed to identify multi-SNP–multi-QT associations. Similarly, under the GEE framework, they proposed the SPU(γ_1, γ_2) tests (i.e., an extension of SPU(γ) to accommodate both multiple QTs and multiple SNPs) for a series of values of $\gamma_1 = 1, 2, \ldots, \infty$ and $\gamma_2 = 1, 2, \ldots, \infty$. Here, a larger γ_1 value tends to put higher weights on QTs with stronger associations with the SNPs, and γ_2 tends to put higher weights on SNPs with stronger associations with the QTs. Based on this, they also proposed the aSPU test for an SNP set (aSPUset). The aSPUset test statistic is defined as the minimum p among all the SPU tests $T_{\text{aSPUset}} = \min_{\gamma_1 \in \Gamma_2, \gamma_2 \in \Gamma_2} P_{\text{SPU}(\gamma_1, \gamma_2)}$. Clearly, aSPUset is an extension of aSPU to identify multi-SNP–multi-QT associations using the same adaptive method that automatically performs data-driven weights' adjustment and selects the most powerful weighted test from all these candidates at both the SNP and trait levels. It has the benefit of measuring the collective effects of multiple SNPs for an increased detection power. The empirical study was performed on an ADNI sample to perform gene-based SNP-set GWAS of 12 imaging QTs within human brain default mode network (DMN). The aSPUset method outperformed competing methods, including aSPU, and identified a new gene AMOTL1 not detected by other SNP-based methods.

Kim *et al.* [88] proposed a similar aSPU test for single-SNP–multi-QT associations using a proportional odds model (POM). Most methods for mining single-SNP–multi-QT associations treated QTs as a response and the SNP as a predictor. In this approach, they treated the SNP as an ordinal response and multiple QTs as predictors and developed a similar aSPUtest under a POM framework instead of the GEE framework used in [80]. Compared with the GEE-based aSPU, the POM-based aSPU has two advantages: 1) it is easier to handle mixed types of traits (e.g., binary and quantitative) and 2) it can handle high-dimensional setting (e.g., QT number is greater than sample size). The empirical studies on ADNI data were performed to identify SNP-based genetic associations with two imaging QT sets: one containing 12 MRI-based QTs related to DMN, and the other containing functional brain connectivity network data among 18 ROIs. Compared with competing methods such as the GEE-based aSPU, the POM-based aSPU performed similarly in both studies that have a low-dimensional setting.

Hua *et al.* [44] proposed a brain imaging GWAS method on identifying single-SNP–multi-QT associations. The method includes a few components to improve detection power. First, they pooled voxel-level measures into 119 ROI-level QTs for reducing both dimensionality and voxel-level noises. Second, they treated the imaging QTs of the entire brain as a single multivariate response and used distance covariance to capture the association between all the QTs and each SNP. This approach could reduce the number of statistical tests and simultaneously embrace ROI interaction effects. Third, they proposed a new FDR-based algorithm for multiple testing adjustment, named local FDR modeling. Empirical study was performed on an ADNI sample to identify SNPs associated with 119 QTs from the entire brain.

Huang *et al.* [89] proposed a new functional GWAS (FGWAS) method for efficiently performing whole genome analysis of high-dimensional imaging QTs. First, instead of doing a univariate analysis to each SNP and each QT, they treated all the imaging QTs as a single functional response measured in the brain space. They proposed a multivariate varying coefficient model (MVCM; a function-on-scale model) to fit all the imaging QTs (as a functional phenotype) with each SNP via embracing key features of a functional phenotype, including spatial smoothness, spatial correlation, and low-dimensional representation. Second, they introduced a GSIS procedure based on global test statistics [51]. This approach selects N_{G0} important SNPs and greatly reduces the genomic search space size from N_G to $\sim N_{G0}$ for $N_{G0} \ll N_G$. Third, they developed an efficient divide-and-conquer algorithm for performing multiple comparisons and achieved a substantial perfor-

Table 4 Example Studies Using Multitrait Methods, Where Single-SNP–Multi-QT or Multi-SNP–Multi-QT Associations Are Examined

Ref	Notes
[80]	Sum of powered score tests (SPU(γ)), adaptive SPU test for multi-trait-single-SNP associations (aSPU), selection of most powerful weighted test via adjusting weights to the studied data
[87]	Sum of powered score tests (SPU(γ_1, γ_2)), adaptive SPU test for multi-trait-multi-SNP associations (aSPUset), selection of most powerful weighted test via adjusting weights to the studied data
[88]	Adaptive SPU test for multi-trait-single-SNP associations (aSPU) under a proportional odds model (POM) instead of the generalized estimation equations (GEE) framework used in [80].
[44]	Brain-wide ROI QTs as a multivariate response, distance covariance between QT set and each SNP, local FDR modeling
[89]	Functional GWAS (FGWAS), multivariate varying coefficient model (MVCM), global sure independence screening (GSIS), GWAS of functional QTs including curves, surfaces and volumes

mance gain on computational time and memory. It can handle functional phenotypes, such as 1-D curves, 2-D surfaces, and 3-D images. The empirical study on an ADNI sample was performed to identify genetic associations with functional QTs on hippocampal surfaces and yielded promising results.

Table 4 summarizes the studies discussed earlier, which are designed to identify multi-QT associations with one or more SNPs. Example strategies for performing multi-QT analyses in recent brain imaging genomics studies include adaptive sum of powered score test to identify the most powerful weighted QT score, distance covariance between QT set and each SNP to reduce the number of tests and incorporate interaction effects among QTs, and modeling all the QTs as a single functional response to embrace spatial smoothness and correlation as well as low-rank representation. Compared with single-trait methods, multi-QT genetic association analysis has the potential to not only improve detection power but also reveal complex imaging genomics associations involving multiple contributing QTs.

E. Pathway and Network Enrichment Methods

Pathway and network analyses are routinely used in genomic studies [91]. Analyzing genomic data through sets defined by biological pathways and functional interaction networks offers enormous potential to increase statistical power and translate genomic findings into meaningful biological hypotheses. For example, if we define an SNP set using a pathway of interest, we can employ the multi-SNP methods reviewed in Section III-C to examine the joint effect of this pathway-based SNP-set on any trait. Most of these multi-SNP methods use a single multivariate learning model to relate multiple SNPs to a trait. Here, we review another category of popular methods called enrichment analysis, which are widely used in the pathway and network analysis of GWAS findings. Different from the multi-SNP methods discussed earlier, an enrichment analysis typically involves two steps: 1) conduct SNP- or gene-based GWAS on a trait and 2) perform pathway or network enrichment analysis of the GWAS findings.

One type of enrichment analysis method is threshold-based (e.g., hypergeometric test or Fisher's exact test) and is used to identify pathways or subnetworks that are overrepresented by the "significant" GWAS hits. Another type of enrichment analysis method is rank-based (e.g., GSEA-SNP [92]) and uses a Kolmogorov–Smirnov-like running sum to quantify the degree to which a pathway- or network-derived gene set (GS) is overrepresented at the top of the gene list ranked by the GWAS results. These analyses are of high significance. They can identify pathways and networks related to imaging QTs or disease outcomes, which can potentially serve as the foundation for the development of diagnostic, therapeutic, and preventative approaches for complex brain disorders. In the following, we review a few example studies using pathway and network enrichment methods.

Ramanan *et al.* [93] performed a genome-wide pathway analysis of memory impairment on an ADNI sample. A composite memory measure was computed from the ADNI neuropsychological test battery and used as the QT in this study. GWAS was performed on this QT but did not yield any significant findings after multiple testing adjustments. A subsequent genome-wide pathway analysis was then conducted through applying GSA-SNP software [94] to the GWAS result and identified 27 significantly enriched canonical pathways after FDR correction. The resulting pathways include memory-related signaling pathways and pathways related to cell adhesion, neuronal differentiation and outgrowth, or inflammation. These results indicate that the pathway enrichment analysis could not only offer increased detection power but also yield valuable biological information to help mechanistic understanding.

Yao *et al.* [95] expanded the scope of enrichment analysis from GWAS to voxelwise brain imaging studies and proposed a framework for mining regional imaging genetic associations via voxelwise enrichment analysis. The main idea was to treat an ROI as a set of voxels similar to a pathway as a set of SNPs or genes in the genomic studies. A post hoc enrichment analysis was performed on the voxelwise statistics to identify ROIs overrepresented by the top voxel findings. Fisher's exact test for independence was used to calculate the enrichment p-value for each ROI. The existing ROI-based methods often collapse the voxel measures into a single value (e.g., the average) and may have limited power when only weak signals exist in part of an ROI. The enrichment-based strategy can properly address this challenge. The empirical study was performed on an ADNI sample to evaluate pairwise associations between 19 AD candidate SNPs and FDG-PET imaging QTs from 116 ROIs across the entire brain. The proposed enrichment method outperformed traditional ROI and voxelwise approaches and identified a number of new significant associations. Some of these new findings were supported by evidences from tissue-specific brain transcriptome data.

Yao *et al.* [90] expanded the scope of enrichment analysis from GWAS to brain imaging genomics studies. They proposed a new 2-D enrichment analysis paradigm, called imaging genetic enrichment analysis (IGEA). IGEA jointly considers meaningful GSs and BCs and aims to identify

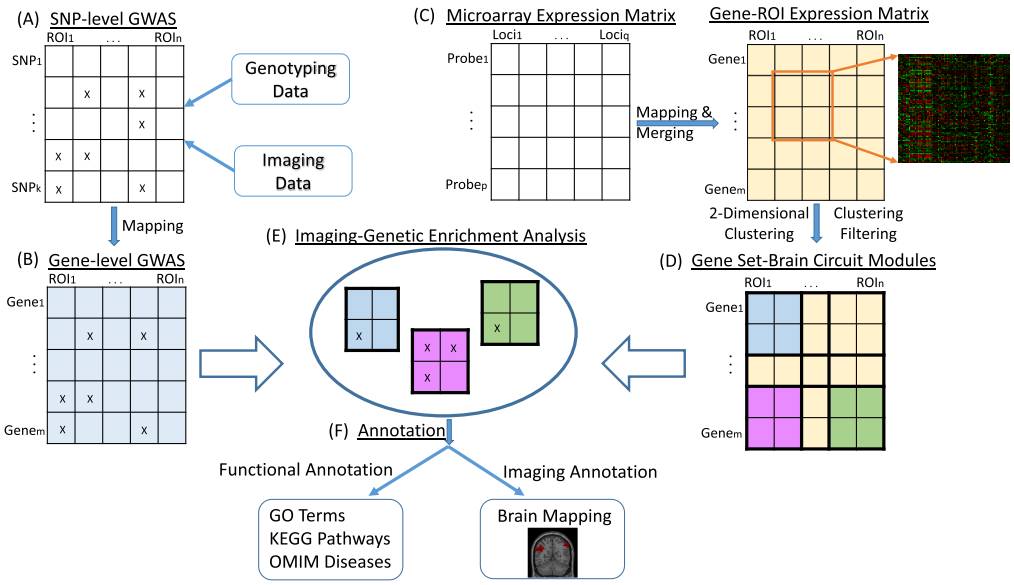


Fig. 4. ***IGEA framework proposed in [90]. Images are reproduced here from a Springer open-access article [90].***

GS-BC pairs overrepresented by SNP–QT findings from BWGW imaging genetic association study. To demonstrate the IGEA framework, they used the whole-brain transcriptome data from the Allen Human Brain Atlas (AHBA) [96] to construct GS and BC modules so that, within each module, genes share similar expression patterns across ROIs and ROIs share similar expression patterns across genes. Fig. 4 shows the IGEA workflow: (A) perform SNP-level GWAS of brain-wide imaging QTs, (B) map SNP-level GWAS findings to gene-level summary statistics, (C) construct gene–ROI expression matrix from AHBA data, (D) construct GS-BC modules by performing 2-D hierarchical clustering on gene–ROI expression matrix and then filter out 2-D clusters with an average correlation below a user-given threshold, (E) perform IGEA by mapping gene-based GWAS findings to the identified GS-BC modules, and (F) for each enriched GS-BC module, examine the GS using gene ontology (GO) terms [97], Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [98], and Online Mendelian Inheritance in Man (OMIM) disease databases (https://omim.org/), and map the BC to the brain. The empirical study using the brain transcriptome data from AHBA and brain imaging genetics data from ADNI identified 25 significant high-level GS-BC modules and showed the promise of IGEA on revealing high-level imaging genomics associations.

Similar to pathways, biological networks are also valuable prior knowledge that can assist GWAS to identify meaningful high-level genomics associations with a target phenotype. For example, network-based GWAS aims to identify phenotype-associated modules from biological networks [100]. This high-level association evaluates the collective effect of all the SNPs/genes within the network module on the phenotype and thus provides not only increased detection power but also meaningful biological interpretation. Yao *et al.* [99] proposed a module detection method for brain imaging genomics studies using tissue-specific biological networks. Fig. 5 shows its workflow. First, GWAS is performed on a target imaging QT. Second, the GWAS results are reprioritized using the NetWAS approach [101]. NetWAS couples machine learning methods [e.g., support vector machines (SVMs) and ridge regression] with a tissue-specific functional interaction network [102] (specific to the imaging QT in our case) to rerank the GWAS results. Using network topology information, SNPs connected to more top findings tend to be pushed more toward the top of the reranked list. As a result, the top reprioritized findings tend to be more densely connected than the top findings in the original GWAS. Thus, the third step is to identify densely connected modules using only interactions among these top reprioritized findings. Finally, enrichment analysis is applied to these modules to identify the ones overrepresented by the original GWAS findings. The empirical study was performed on an ADNI sample to identify modules related to the mean FDG-PET measure in amygdala and yielded promising results.

Table 5 summarizes the studies discussed earlier, which are designed to detect high-level imaging genomic associ-

Table 5 Example Studies Using Pathway and Network Enrichment Methods, Which Aim to Detect High-Level Imaging Genomic Associations Related to Pathways, Networks, or BCs

Ref	Notes
[91]	A review of pathway and network analysis of genomic data
[93]	Pathway analysis of memory impairment, GSA-SNP software
[95]	ROI enrichment analysis based on voxelwise findings
[90]	Two dimensional Imaging Genetic Enrichment Analysis (IGEA)
[99]	Tissue-specific network (specific to the imaging QT), network module detection, NetWAS re-prioritization

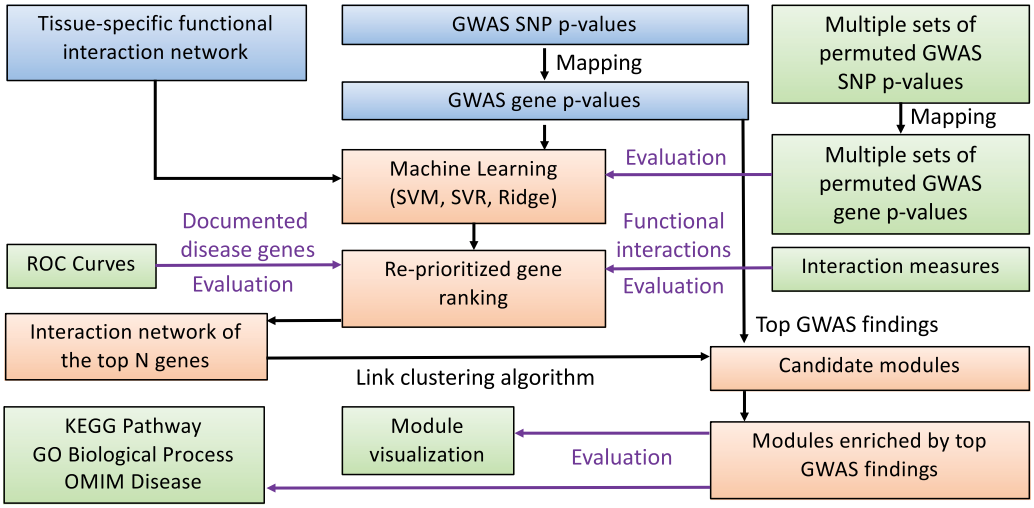


Fig. 5. ***Workflow for identifying functional interaction modules from the tissue-specific network using imaging GWAS findings. Images are reproduced here with permission from Oxford University Press [99].***

ations related to pathways, networks, or BCs. The brain imaging genomics studies usually apply the standard enrichment methods widely used in the genomic domain, including both threshold- and rank-based approaches. In addition to these enrichment calculation methods, various related strategies have been proposed to address specific issues in brain imaging genomics. For example, the enrichment analysis can be transferred from the genomic domain to the imaging domain to perform an ROI enrichment analysis based on voxelwise findings [95]. It can also be extended to 2-D IGEA to mine high-level imaging genetic associations based on massive BWGW SNP–QT results [90]. In addition, given the recent availability of tissue-specific networks, the imaging GWAS-based module identification can be extended to use the functional interaction network specific to the studied imaging QT (i.e., tissue from the corresponding brain region) [99]. In sum, the enrichment methods examine the collective effect of an SNP/GS, a QT set, or both, and have the potential to increase detection power. Also, the examined SNP or QT sets correspond to functionally annotated biological entities and may provide valuable insights into underlying mechanisms.

A topic relevant to enrichment analysis is prioritization. Enrichment analysis is typically performed at the end of the analysis pipeline (e.g., as a post-hoc analysis of the GWAS findings). Prioritization takes a reverse approach where valuable prior knowledge, such as pathway and network information, is used to select a small set of genes for the subsequent analyses. For example, Patel *et al.* [103] used GO [97] to build a biological process network associated with 21 AD seed genes from [63] and then performed imaging genetic analyses targeting at all the genes in the network. Lorenzi *et al.* [104] used the GTEx database (https://gtexportal.org/) to screen candidate SNPs generated from the imaging genetic analysis of a discovery sample for obtaining potential expression QT loci (eQTL) and then performed another imaging genetic analysis targeting only these prioritized loci in an independent sample. Grothe *et al.* [105] used Amyloid-PET and MRI scans to compute brain-wide spatial patterns of AD-typical amyloid deposition and neurodegeneration and then used the whole-brain gene expression database AHBA [96] to rank and prioritize genes based on their spatial correlation with the above-mentioned amyloid burden and neurodegeneration patterns. In short, the strength of gene prioritization is twofold: 1) it reduces the burden of multiple testing and has the potential to increase detection power and 2) the valuable functional annotation knowledge used for prioritization can help with biological interpretation and alleviate the risk for false discoveries. On the other hand, we should also be cautious about its possible limitations, such as bias associated with the reference atlas or prior knowledge used for prioritization and difficulty in updating findings according to the evolution of these valuable resources. Finally, in addition to enrichment analysis and prioritization, the pathway and network information can also be incorporated into advanced statistical and machine learning models to guide our search for more complicated imaging genomics associations (see [106]–[108]), which will be discussed in Sections V and VI.

F. Interaction Methods

Most brain imaging genomics association studies examine the main effects of genetic variants on imaging QTs. It is well known that these main effects can only explain a portion of heritability of the studied QTs. Missing heritability can often be attributed in part to the interaction effects (or epistatic effects) within genetic variants or between the genetic and environmental factors. These studies are facing major statistical and computational challenges since

an exponentially increasing number of possible tests (to the order of the interaction) significantly reduce the statistical power due to multiple comparison correction. Thus, a major topic in epistatic studies is to find an effective search strategy to reduce computational time and increase statistical power. In the following, we review a few example studies exploring the effects of SNP–SNP interaction or SNP–environment interaction on imaging QTs.

Zieselman *et al.* [109] presented a bioinformatics pipeline for the epistatic analysis of an MRI-based QT (i.e., mean gray matter density) using an ADNI sample. The pipeline employed two phases to dramatically reduce the search space. Phase I was focused on identifying a set of genes with significant SNP–SNP interactions, where the quantiative multifactor dimensionality reduction (QMDR) method [110] was used to examine the SNP–SNP interaction effect on the QT within each gene; 20 genes with 34 SNPs were identified. In Phase II, these genes were uploaded to the Integrative Multispecies Prediction (IMP) webserver (http://imp.princeton.edu; see [111]) to create a gene interaction network that incorporates the prior functional genomics knowledge. Up to 20 additional genes connected to the input genes with high confidence were allowed to be added to the IMP network. Ten genes (six original + four additional) with ten SNPs were identified. Finally, QMDR was used to examine all pairwise, three-way, and four-way SNP–SNP interactions among these ten SNPs. The most significant finding is a three-way interaction, including two SNPs within the olfactory gene cluster and one TRPC4 SNP. The goal of this study was to use the existing knowledge to reduce the possibility of false positives instead of identifying all possible interactions which is a much harder task to accomplish.

Meda *et al.* [112] performed a genome-wide interaction analysis (GWIA) of MRI-based atrophy measures in the hippocampus and entorhinal cortex using an ADNI sample. Their strategy to reduce the number of tests was to examine 151 million SNP pairs based on the gene–gene interaction patterns in the KEGG pathway database. Linear regression implemented in the INTERSNP software [113] was used to identify epistatic effects while controling for sex, age, education, APOE e4, and clinical status. They identified 109 SNP–SNP interactions for right hippocampal atrophy and 123 for right entorhinal cortex atrophy. These findings were overrepresented in several interesting pathways, including the calcium signaling, axon guidance, and ErbB signaling pathways.

Hibar *et al.* [114] performed a GWIA of MRI-based TLV using an ADNI sample. The EPISIS software [115] was employed to screen all possible SNP pairs based on a machine-learning algorithm called sure independence screening (SIS) [51]. SIS is a screening method that evaluates the correlation strength between each SNP pair and the outcome and selects the most associated SNP pairs. In this study, 111 SNP–SNP interaction pairs were obtained after SIS screening. All these interaction terms were then included in a single ridge regression model, where the extended Bayesian information criterion (BIC) [116] was used to identify the most relevant SNP pairs. This study identified a significant interaction between rs1345203 and rs1213205.

Ge *et al.* [117] presented a kernel machine method (KMM) to evaluate the main and interaction effects among multiple genetic and nongenetic variable sets on an imaging QT. Their model includes three separate kernels. The first one is a genetic kernel to measure the epistatic and joint effect of an SNP set on an imaging QT. The SNP sets can be defined by haplotype structure, gene, or pathway. The second one is a nongenetic kernel to measure the collective effect of multiple nongenetic factors. The third one is the Hadamard product of the above-mentioned two kernels to examine their interaction effect. Using an ADNI sample, they applied KMM to explore the interaction effects between each of 21 AD candidate genes and six cardiovascular disease (CVD) risk factors on MRI-based hippocampal volume. Two genes, CR1 and EPHA1, were identified to have such interaction effects with the CVD risk factors.

Wang *et al.* [118] proposed a set-based mixed effect model for gene–environment interaction (MixGE) on imaging QT. They reviewed major set-based association tests and grouped them into five categories: 1) burden tests (collapsing variants into a burden score); 2) adaptive burden tests (burden tests using data-adaptive weights); 3) variance component tests (examining variance of genetic effects); 4) combined tests; and 5) exponential combination tests (both combining burden and variance component tests). Their work is an extension of a combined test named mixed-effects score test (MiST) [119] to examine the gene–environment (G $\times$ E) effect on imaging QTs. The proposed MixGE method models both fixed and random effects of G $\times$ E and examines homogeneous and heterogeneous contributions from an SNP set and their interaction with environmental factors on an imaging QT. They employed score statistics instead of direct parameter estimation to accelerate the computation, which enabled the voxelwise analyses. Similar to [117], the empirical study was performed on the same ADNI sample to explore the interaction effects between each of 21 AD candidate genes and the first principal component of six CVD risk factors on hippocampal volume and voxelwise TBM data. The analysis on the hippocampal volume replicated the results of KMM [117]. The analysis on the TBM data suggested an interaction effect of ABCA7 gene and CVD risk on right superior parietal cortex.

Table 6 summarizes the studies discussed earlier, which are designed to examine the epistatic effects of genetic variants or their interaction effects with nongenetic factors on brain imaging QTs. Given the major statistical and computational challenges induced by an enormous number of possible tests, studies in the field typically employ various strategies to reduce the search space. For example, one strategy is to examine only a small set of candidate interactions with a potential biological mechanism suggested

Table 6 Example Studies Using Interaction Methods, Which Aim to Examine Epistatic Effects of Genetic Variants or Their Interaction Effects With Nongenetic Factors on Imaging QTs

Ref	Notes
[109]	QMDR, IMP, targeted epistatic analysis guided with statistical filtering and functional genomics knowledge
[112]	GWIA, targeted analysis using the KEGG gene-gene interaction patterns, linear regression using the INTERSNP software
[114]	GWIA, sure independence screening algorithm (called EPISIS), ridge regression, extended Bayesian Information Criterion (BIC)
[117]	Kernel machine method (KMM), joint modeling of epistatic and collective effect from a SNP set, collective effect of non-genetic factors, and interaction between genetic and non-genetic factors
[118]	Set-based mixed effect model for gene-environment interaction (MixGE) on imaging QT, score statistics for fast computation

by functional interaction networks or biological pathways. In this case, we should be aware of the strengths and limitations of the prioritization approach, as discussed at the end of Section III-E. Another strategy is to perform data-driven screening to focus on the analysis of a small number of most promising candidate interactions.

IV. IMAGING GENOMICS ASSOCIATIONS: META-ANALYSIS

A key challenge in imaging genomics is the relatively small effect size of genetic variants on the brain—most genetic variants account for under 1% of the variance in a brain measure, when considered individually, meaning that hundreds or even thousands of scans may be needed to detect and independently replicate an effect. An important exception to this rule appears to be the APOE gene; a common form of this gene, APOE4, is carried by around a quarter of the world's population and is associated with a roughly threefold higher lifetime risk for AD. In elderly people, this genotype is associated with a 1–2 standard deviation lower hippocampal volume [121], relative to carriers of the most common form of the gene, APOE3. Nonetheless, other common genetic variants with large effects on the brain have been extremely difficult to find; as a result, studies have expanded to ten thousand subjects or more, in an effort to find replicable associations [120].

In addition, the ability to test over a million markers in the genome for associations with brain measures means that heavy corrections are often required for multiple statistical testing. A typical genome-wide association study might test over a million independent genetic markers; to avoid reporting false positives, the genetics field established a genome-wide association significance threshold (typically $p < 0.05/10^6$ or thereabouts) before an association could be declared significant. The number of traits derived from images in an individual study might also be very large (up to 140 traits in a typical study of cortical thickness and surface area—and well over 10^6 voxels in an image or 10^4 edges in a connectivity network). If every trait is tested for genome-wide associations, this leads to even more stringent significance thresholds. Smith and Nichols [122] gave a detailed power analysis of association testing in large biobanks, noting the very large samples required. In parallel, several researchers examined the power and data requirements for well-powered studies of image-wide genome-wide associations [45], [123] and connectome-wide genome-wide association, which performs association tests at each edge in a graph or network model of brain connectivity [20], [38], [124].

Early attempts to reduce the search space in imaging genomics (by focusing on genes more likely to have effects on the brain) largely failed. Ten years ago, several hundred articles had reported associations between variants in specific candidate genes (e.g., COMT and BDNF) and an imaging trait—yet almost none of these was replicated when tested in independent data. Jahanshad *et al.* [125] pooled regional fractional anisotropy (FA) measures for 6165 subjects from 11 cohorts worldwide to evaluate the effects of 15 candidate SNPs that had been reported in the literature to show associations with white matter microstructure; not a single one of these associations was replicated in independent samples. This "crisis of reproducibility" or "power failure" has also been noted in several branches of science [126], including neuroscience [127], [128].

Modeled on the Psychiatric Genomics Consortium in psychiatric genetics, the ENIGMA Consortium (http://enigma.ini.usc.edu) was founded in 2009 to address these problems and perform large-scale genome-wide association studies for brain measures derived from MRI [129], DTI [125], and EEG [130]. ENIGMA uses a meta-analytic design to pool evidence from large numbers of cohorts worldwide. ENIGMA has since expanded to include over 50 working groups, focusing on global studies of specific brain diseases and has published the largest neuroimaging studies to date of nine brain disorders. Here, we focus on its work in imaging genetics, which can be categorized into studies of common [129] and rare [131] genetic variants and epigenetic variation [132]. These studies may be further subdivided by the data types studied (e.g., MRI and EEG) and methods used (e.g., mass-univariate meta-analysis, tests of genetic overlap between brain traits and other clinical or behavioral traits, and image- or connectome-wide testing of genetic associations). We begin with mass-univariate analyses, as they are the simplest.

Stein *et al.* [120] and Hibar *et al.* [121] identified over 20 genetic loci associated with the volumes of subcortical brain regions, including the hippocampus, amygdala, thalamus, putamen and other regions of the basal ganglia, and intracranial volume. Manhattan plots of these effects are shown in Fig. 6(a) and (b) for each structure; the evidence of association is shown for each genetic marker (on the x-axis) and each regional volume measure (on the y-axis) using a logarithmic scale, $-\log_{10}(p)$. Several aspects are notable from a methodological point of view. First, only hits that are genome-wide significant are considered reliable, by convention, due to the large number of statistical tests performed. To attempt to replicate these hits in independent data, ENIGMA partnered with

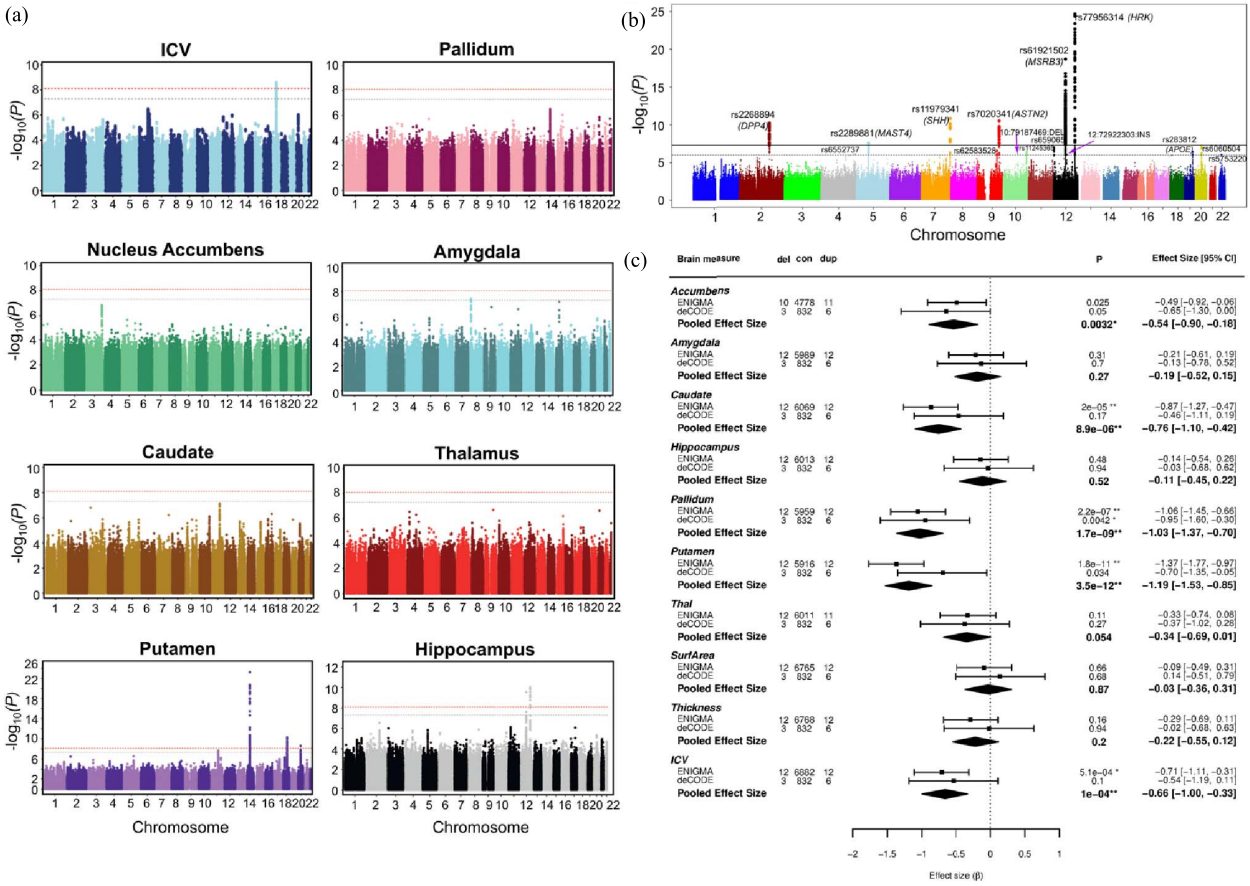


Fig. 6. ***Example ENIGMA results. (a) and (b) Manhattan plots of GWAS on ICV and subcortical volumes [120], [121]. (c) Catalog of rare variants and their effects on the brain created by partnerships among ENIGMA, deCODE Genetics, and the UK Biobank [131]. Images are reproduced here with permission from Springer Nature [120], [121], [131].***

the CHARGE Consortium on a series of articles reporting GWAS in ever-increasing sample sizes of intracranial volume [133], hippocampal volume [121], and all subcortical volumes [134]. Earlier articles performed a simple p-value lookup in the replication data; a later article performed a meta-analysis of all cohorts.

These analyses were performed using standardized protocols for quality control of the imaging and genomic data, as well as imputation of genetic data to common reference panels, such as the 1000 genomes reference panel (this step allows the same set of variants to be analyzed across cohorts even if some sites have used different genotyping chips). A later cortical GWAS [129] led to an annotated atlas of over 200 genetic loci associated with surface area and thickness measures from 70 cortical regions. Parallel work by the UK Biobank reported GWAS for MRI, DTI, and even functional MRI metrics in their first 9000 subjects scanned [12], [135]. The UK Biobank was subsequently added to the ENIGMA studies as a replication sample, showing generally strong replication [129]. A parallel set of studies also assessed the overlap between these brain-related genetic loci and genetic markers implicated in a range of brain diseases and neuropsychiatric disorders, including AD and Parkinson's disease [129], schizophrenia and bipolar disorder [35], [136], obsessive-compulsive disorder [137], Tourette syndrome [138], and even IQ [129], [139].

Holland *et al.* [140] studied the discoverability of SNPs using GWAS for a range of different traits, including image-derived measures. By modeling the effect sizes found empirically for SNPs associated with brain and behavioral traits, they noted that the rate of discovery of SNPs—and the cumulative percentage of variance explained—tends to follow an S-shaped curve. Remarkably, to discover markers that account for over half of the SNP heritability (the proportion of variance due to genotyped SNPs), they estimate that 10 000–10 000 000 participants would be needed, depending on the trait or disease studied (e.g., increasing numbers of subjects were needed to perform a well-powered GWAS of plasma cholesterol levels, regional brain volumes, schizophrenia, and major depression). Differences in SNP discoverability, for each trait, depend on the genetic architecture of each trait— the fraction of the genome that accounts for various proportions of the observed variance, the effect sizes for each SNP in this set, and the minor allele

frequency (MAF) of the variants implicated. By estimating these from empirical data, the detailed power analyses were reported.

Some Bayesian methods have been proposed to overcome the heavy statistical corrections associated with mass-univariate testing of over a million genetic loci. Smeland *et al.* [136] categorized markers as belonging to different genetic categories (e.g., lying within and outside known genes or by functional type, such as enhancers or promoters). As brain-relevant genetic loci have different prevalence in these various genetic categories, Smeland *et al.* [136] were able to use the conditional FDR method to discover some known SNPs more efficiently (i.e., in smaller samples) as well as other genetic markers not yet discovered using the existing methods. Similarly, genetic clustering—the quest to identify overlap in genetic influences between traits—has led to genetic connectomes—matrices or graphs of genetic correlations, in which traits with common genetic determination are stored in a matrix and clustered. Some researchers argue that genetic clustering of voxels in an image, edges in a network, or vertices on surface models of the cortex may yield more efficient targets for GWAS [141], [142]; such methods are just beginning to be explored.

ENIGMA is also using meta-analysis to assess effects on the brain of other types of genetic variation. ENIGMA's Epigenetics group identified two sites in the genome where methylation relates to hippocampal volume ($N = 3337$; see [132]). This type of study is computationally analogous to a GWAS, although methylation occurs at a somewhat lesser number of genetic loci, making the analysis slightly more efficient; nonetheless, thousands of subjects are still needed to detect and replicate individual associations.

As biobanks grow in size, it has become possible to discover and independently replicate effects on the brain of rare genetic loci (with a prevalence of <1:1000 individuals), such as the genetic deletions responsible for 22q deletion syndrome [143], [144]. The ENIGMA CNV Consortium [131] is performing a systematic study of these rare variants; in general, they may have a far greater effect on the brain than common variants, making their effects more efficient to replicate. Partnerships among ENIGMA, deCODE Genetics, and the UK Biobank are creating a catalog of rare variants and their effects on the brain [see Fig. 6(C)] [131]. Once the effects on the brain are known for deletions of different sizes, a second round of analyses may be required to determine how specific genetic loci within the deleted region influence the effects.

V. IMAGING GENOMICS ASSOCIATIONS: MULTIVARIATE REGRESSION

Here, we provide a review of machine learning studies that use regression models to identify complex multi-SNP and/or multi-QT associations. Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be the genetic data with p variables on n subjects. Let $\mathbf{Y} \in \mathbb{R}^{n\times q}$ be the imaging data with q variables on n subjects. We assume that each column of $\mathbf{X}$ and $\mathbf{Y}$ is normalized with zero mean and unit variance. Most of the regression models discussed earlier can be described using the following generic regularized loss function framework:

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{X}, \mathbf{Y}, \mathbf{W}) + \Sigma_{i=1}^{m} \lambda_i \mathcal{R}_i(\mathbf{W}) \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{p\times q}$ is the weight matrix for regression of $\mathbf{Y}$ on $\mathbf{X}$ and λ_i is the parameter balancing the loss function $\mathcal{L}(\mathbf{X}, \mathbf{Y}, \mathbf{W})$ and the regularization $\mathcal{R}_i(\mathbf{W})$.

A sparsity-inducing regularization term is often included in these models. The motivation is twofold. First, it is reasonable to hypothesize that only a small number of markers are relevant in the resulting imaging genomics association. The sparsity term can help identify these relevant markers. Second, the sparsity constraint can reduce the model complexity and subsequently reduce the risk of overfitting.

In the following, we discuss example studies using the four categories of methods: 1) sparse multiple regression (SMR; univariate response, $\mathbf{W}$ degrades to a vector $\mathbf{w}$); 2) sparse multivariate multiple regression (SMMR; multivariate response, $\mathbf{W}$ is a matrix); 3) sparse reduced-rank regression (SRRR; reducing the rank of $\mathbf{W}$, e.g., $\mathbf{W} = \mathbf{B}\mathbf{A}^T$); and 4) Bayesian regression and neural network (NN) models.

A. Sparse Multiple Regression

We start with a few relatively simple SMR models, where the response is a scalar. Some of these (see [106]) will be later extended into its multivariate version.

Silver *et al.* [106] proposed the pathways group lasso with adaptive weights (P-GLAW) algorithm, which is based on a group lasso model

$$\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \Sigma_{g\in\mathcal{G}} d_g \|\mathbf{w}_g\|_2 \tag{2}$$

where $\mathcal{G}$ defines the grouping structure of $\mathbf{w}$. The goal is to identify a set of SNPs from $\mathbf{X}$ to predict a single imaging QT $\mathbf{y}$. The SNPs are grouped using the pathway knowledge so that the feature selection is done at the pathway level to enhance biological interpretation and generate insightful results. The empirical study was performed on synthetic data simulated based on an ADNI sample and canonical pathways from the Molecular Signals Database (MsigDB; see [145]).

Hao *et al.* [146] proposed a tree-guided sparse learning (TGSL) method, which is also based on a group lasso model but with a tree structure

$$\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \Sigma_{i=0}^{l} \Sigma_{j=1}^{n_i} d_j^i \|\mathbf{w}_{G_j^i}\|_2 \tag{3}$$

where G_j^i indicates a predefined tree (say T) structure of $\mathbf{w}$, the tree T has l depth level, and the ith level contains

n_i nodes organized as $T_i = \{G_1^i, \ldots, G_j^i, \ldots, G_{n_i}^i\}$. The goal is to identify a set of SNPs from $\mathbf{X}$ to predict a single imaging QT $\mathbf{y}$. The SNPs are grouped using a tree structure, which groups SNPs by LD blocks and groups LD blocks by genes. The empirical study was performed on an ADNI sample to predict six target imaging QTs using SNPs from 20 AD genes.

Wang *et al.* [147] proposed a diagnosis-aligned multimodal (DAMM) method for regressing a target SNP $\mathbf{x}$ on multimodal imaging QTs ($\mathbf{Y}_m$ for $m \in [1, M]$) as follows:

$$\min_{\mathbf{W}} \Sigma_{m=1}^{M} \|\mathbf{x} - \mathbf{Y}_m \mathbf{w}_m\|_2^2 + \lambda_1 \mathcal{R}_1(\mathbf{W}) + \lambda_2 \mathcal{R}_2(\mathbf{W}) \quad (4)$$

where $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_M]$. The first regularization $\mathcal{R}_1(\mathbf{W})$ is an $l_{2,1}$-norm to select features with effects on most of the modalities. The second regularization $\mathcal{R}_2(\mathbf{W})$ is a graph Laplacian term that encourages the subjects within (between) the same diagnositic group to have similar (different) values in the projected space (i.e., these projected imaging components are aligned with diagnosis). The empirical study was performed on an ADNI sample, where the response is the APOE e4 SNP and the predictors include two modalities of ROI measures: 1) VBM measure from structural MRI and 2) hypergraph-based clustering coefficient measure from fMRI.

B. Sparse Multivariate Multiple Regression

Now, we focus on SMMR models. Wang *et al.* [148] proposed a Group-Sparse Multi-task Regression and Feature Selection (G-SMuRFS) method, which is a structured sparse model [see also Fig. 7(a)]

$$\min_{\mathbf{W}} \|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda_1 \|\mathbf{W}\|_{G_{2,1}} + \lambda_2 \|\mathbf{W}\|_{2,1} \quad (5)$$

where the group $l_{2,1}$-norm regularization ($\|\mathbf{W}\|_{G_{2,1}}$) does feature selection at the group level (e.g., LD-block) and the $l_{2,1}$-norm regularization ($\|\mathbf{W}\|_{2,1}$) does feature selection at the individual SNP level. The empirical study was performed on an ADNI sample, where 1224 SNPs from 37 AD genes were used to predict ten VBM measures and 12 FreeSurfer [150] measures, and SNPs were grouped by LD blocks.

Wang *et al.* [151] aimed to use longitudinal imaging QT data ($\mathbf{Y}_k$ for $k \in [1, t]$) to predict SNP data ($\mathbf{X}$) and proposed the following task-correlated longitudinal sparse regression (TCLSR) model (each time point treated as a task):

$$\min_{\mathbf{W}} \Sigma_{k=1}^{t} \|\mathbf{X} - \mathbf{Y}_k \mathbf{W}_k\|_F^2 + \lambda_1 \mathcal{R}_1(\mathbf{W}) + \lambda_2 \mathcal{R}_2(\mathbf{W}) \quad (6)$$

where $\mathbf{W} = [\mathbf{W}_1, \ldots, \mathbf{W}_t]$ [the same as that shown in Fig. 7(b)], $\mathcal{R}_1(\mathbf{W})$ is a trace norm to approximate a low-rank representation of $\mathbf{W}$, and $\mathcal{R}_2(\mathbf{W})$ is an $l_{2,1}$-norm to select features with effects at most of the time points. The empirical study was performed on an ADNI sample to predict 1224 SNPs from 37 AD genes using longitudinal imaging QTs.

Wang *et al.* [149] studied the same problem as in [151] and proposed a new model temporal structure autolearning (TSAL) as follows [see also Fig. 7(b)]:

$$\min_{\mathbf{W}} \Sigma_{k=1}^{t} \|\mathbf{X} - \mathbf{Y}_k \mathbf{W}_k\|_F^2 + \lambda_1 \mathcal{R}_1(\mathbf{W}) + \lambda_2 \mathcal{R}_2(\mathbf{W}) \quad (7)$$

where $\mathcal{R}_1(\mathbf{W})$ is a Schatten p-norm regularization term to identify low-rank structures [e.g., four green boxes sharing similar patterns in Fig. 7(b)] and $\mathcal{R}_2(\mathbf{W})$ is a $l_{2,1}$-norm to select SNPs correlated with most QTs over the time [e.g., the red box in Fig. 7(b)]. Of note, compared with TCLSR [see (6)], Schatten p-norm approximates rank minimization better than the trace norm [152], and $l_{2,0+}$-norm can achieve a more sparse solution than $l_{2,1}$-norm. The empirical study was performed on an ADNI sample, where longitudinal imaging QTs were used to predict 3576 SNPs from 153 AD candidate genes.

Zhou *et al.* [153] proposed a joint projection learning and sparse regression (JPLSR) model for identifying multi-SNP–multi-QT association. JPLSR model takes the following form [different from the generic form shown in (1)]:

$$\begin{aligned} \min_{\mathbf{W}_1, \mathbf{W}_2, \mathbf{P}} \; & \|(\mathbf{Y} - \mathbf{X}\mathbf{W}_2)^T\|_{2,1} + \lambda_1 \|\mathbf{X}\mathbf{W}_2\mathbf{P} - \mathbf{Y}\mathbf{W}_1\|_F^2 \\ & + \lambda_2 \mathcal{R}(\mathbf{X}, \mathbf{Y}, \mathbf{W}_1, \mathbf{W}_2, \mathbf{P}) + \lambda_3 \|\mathbf{W_1}\|_{2,1} + \lambda_3 \|\mathbf{W_2}\|_{2,1} \\ \text{s.t.} \quad & \mathbf{P}\mathbf{P}^T = \mathbf{I}. \end{aligned} \quad (8)$$

The first term is the loss function to find the multi-SNP–multi-QT association. The second term is to project the SNP data and imaging QT data into a joint latent space to aid association discovery. The third term combines two graph Laplacian terms (one for SNP data and one for imaging data) to encourage the genetic and imaging components projected to the latent space which are aligned with diagnosis. The fourth and fifth terms are two $l_{2,1}$-norms for selecting relevant imaging and SNP features, respectively. The empirical study was performed on an ADNI sample to relate 93 ROI-based imaging QTs to 3123 SNPs from top AD candidate genes.

C. Sparse Reduced-Rank Regression

Here, we focus on reviewing studies using SRRR, which is a special type of multivariate multiple regression models for identifying multi-SNP–multi-QT associations from high-dimensional imaging genomic data. The major goal is to minimize the rank of the $(p \times q)$ regression matrix $\mathbf{W}$. Assuming that $\mathbf{W}$ has a reduced rank of $r < \min(p, q)$, Vounou *et al.* [154] proposed to rewrite $\mathbf{W}$ as the product of a $(p \times r)$ matrix $\mathbf{B}$ and $(q \times r)$ matrix $\mathbf{A}$: $\mathbf{W} = \mathbf{B}\mathbf{A}^T$. In [154], they studied the following rank-one model

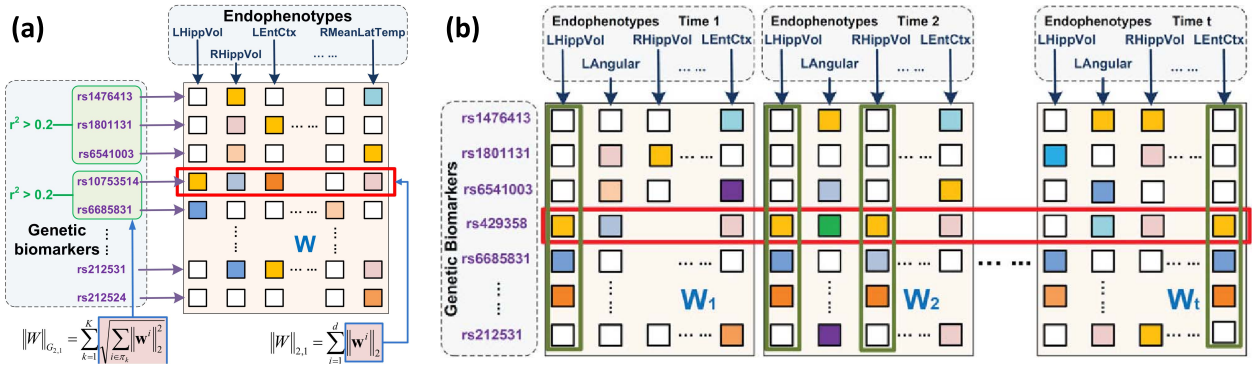


Fig. 7. *Example structured SMMR models, where only regression weight matrices W are shown here. Let X be genotype data and Y be imaging QT data. (a) Illustration of the G-SMuRFS model [148] ($\min_{\mathbf{W}} \|\mathbf{Y}-\mathbf{XW}\|_F^2 + \lambda_1 \|\mathbf{W}\|_{G_{2,1}} + \lambda_2 \|\mathbf{W}\|_{2,1}$), where the group $l_{2,1}$-norm regularization ($\|\mathbf{W}\|_{G_{2,1}}$) does feature selection at the group level (e.g., LD-block), and the $l_{2,1}$-norm regularization ($\|\mathbf{W}\|_{2,1}$) does feature selection at the individual SNP level. Image is reproduced here with permission from Oxford University Press [148]. (b) Illustration of the TSAL model [149] ($\min_{\mathbf{W}} \Sigma_{k=1}^{t} \|\mathbf{X}-\mathbf{Y_k W_k}\|_F^2 + \lambda_1 \mathcal{R}_1(\mathbf{W}) + \lambda_2 \mathcal{R}_2(\mathbf{W})$), where $\mathcal{R}_1(\mathbf{W})$ is a Schatten p-norm regularization term to identify low-rank structures (e.g., four green boxes sharing similar patterns) and $\mathcal{R}_2(\mathbf{W})$ is an $l_{2,1}$-norm to select SNPs correlated with most QTs over time (e.g., the red box). Image is reproduced here with permission from Mary Ann Liebert, Inc. [149].*

(i.e., **A** and **B** become two vectors **a** and **b**):

$$\min_{\mathbf{a},\mathbf{b}} \|\mathbf{Y} - \mathbf{Xba}^{\mathbf{T}}\|_F^2 + \lambda_1 \|\mathbf{a}\|_1 + \lambda_2 \|\mathbf{b}\|_1 \tag{9}$$

where the l_1 term is applied to both **a** and **b** for sparse feature selection. This model was evaluated using the synthetic imaging genetic data simulated using an ADNI sample.

Vounou *et al.* [155] applied a slightly modified version of the above-mentioned model [see (9)] to an ADNI sample, where they use genome-wide SNP data to predict voxelwise longitudinal imaging QTs. They first applied a penalized linear discriminant analysis (LDA) for voxel filtering to identify disease-relevant imaging QTs and then employed the following SRRR model to predict QT data **Y** from SNP data **X**:

$$\min_{\mathbf{a},\mathbf{b}} \|\mathbf{Y} - \mathbf{Xba}^{\mathbf{T}}\|_F^2 + \lambda \|\mathbf{b}\|_1 \tag{10}$$

where the l_1 term is applied for SNP selection. A data resampling scheme was used to identify SNPs with high selection probability.

In [107], Silver *et al.* integrated the P-GLAW idea [see (2)] into the SRRR framework [see (9) and proposed the following pathways SRRR (P-SRRR) model:

$$\min_{\mathbf{a},\mathbf{b}} \|\mathbf{Y} - \mathbf{Xba}^{\mathbf{T}}\|_F^2 + \lambda \Sigma_{g \in \mathcal{G}} d_g \|\mathbf{b}_g\| \tag{11}$$

where $\mathcal{G}$ defines the grouping structure of **b**. The goal is to identify a set of SNPs from **X** to predict a set of AD-related imaging QT **Y**. The SNPs are grouped using the pathway knowledge so that the feature selection is done at the pathway level. The empirical study was performed on an ADNI sample with KEGG canonical pathways from MsigDB [145].

Zhu *et al.* [156] proposed a structured SRRR (S-SRRR) model for regressing brain-wide imaging QT data **Y** on genome-wide SNP data **X** as follows:

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{B}} \|\mathbf{Y} - \mathbf{XBA}^{\mathbf{T}}\|_F^2 + \lambda_1 \|\mathbf{A}\|_{2,1} + \lambda_2 \|\mathbf{B}\|_{2,1} \\ &\text{s.t. } \mathbf{A}^T\mathbf{A} = \mathbf{I} \end{aligned} \tag{12}$$

where the $l_{2,1}$-norm regularizes **A** and **B** in a rowwise fashion for effective selection of SNP and QT features. The empirical study was performed on an ADNI sample to relate 2098 SNPs from 153 AD candidate genes to 93 imaging QTs.

Zhu *et al.* [157] employed the graph self-representation method [158] to model a sparse matrix $\mathbf{S} \in \mathbb{R}^{p \times p}$ capturing the internal partial correlations among the SNP data **X** as follows:

$$\begin{aligned} &\min_{\mathbf{S}} \|\mathbf{X} - \mathbf{XS}\|_F^2 + \lambda_1 \|\mathbf{S}\|_1 + \lambda_2 \|\mathbf{S}\|_{2,1} \\ &\text{s.t. } \text{diag}(\mathbf{S}) = \mathbf{0} \end{aligned} \tag{13}$$

where the constraint $\text{diag}(\mathbf{S}) = \mathbf{0}$ was imposed to avoid generating the trivial solution. Integrating the above-mentioned model [see (13)] into the S-SRRR model [see (12)], Zhu *et al.* [157] proposed the following graph-regularized S-SRRR (GRS-SRRR) model for regressing **Y** on **X** given **S** as a graph constraint:

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{B},\mathbf{S}} \|\mathbf{Y} - \mathbf{XBA}^{\mathbf{T}}\|_F^2 + \lambda_1 \|\mathbf{X} - \mathbf{XS}\|_F^2 \\ &\qquad + \lambda_2 \|\mathbf{S}\|_1 + \lambda_3 \|[\mathbf{B},\mathbf{S}]\|_{2,1}, \\ &\text{s.t. } \mathbf{A}^T\mathbf{A} = \mathbf{I} \text{ and } \text{diag}(\mathbf{S}) = \mathbf{0}. \end{aligned} \tag{14}$$

The empirical study was performed on the same ADNI sample as in [156].

Zhu *et al.* [159] modified the GRS-SRRR model [see (14)] into the following robust GRS-SRRR (RGRS-

SRRR) model:

$$\min_{\mathbf{A},\mathbf{B},\mathbf{S}} \sqrt{\|\mathbf{Y}-\mathbf{XBA^T}\|_{2,1}} + \lambda_1\sqrt{\|\mathbf{X}-\mathbf{XS}\|_{2,1}} + \lambda_2\|\mathbf{S}\|_1 + \lambda_3\|\mathbf{[B,S]}\|_{2,1}$$
$$\text{s.t. } \mathbf{A}^T\mathbf{A}=\mathbf{I} \text{ and } \operatorname{diag}(\mathbf{S})=\mathbf{0}. \quad (15)$$

Here, $\|\mathbf{Y}-\mathbf{XBA^T}\|_{2,1}$ and $\|\mathbf{X}-\mathbf{XS}\|_{2,1}$ are the robust versions of $\|\mathbf{Y}-\mathbf{XBA^T}\|_F^2$ and $\|\mathbf{X}-\mathbf{XS}\|_F^2$, respecitively, according to [158] and [160]. The empirical study was performed on an ADNI sample with 90 imaging QTs and 3996 SNPs from 153 AD candidate genes.

D. Bayesian Regression and Neural Network Models

While many regularized multivariate regression models have been proposed in brain imaging genomics, several Bayesian methods have been studied to achieve similar goals. For example, inspired by G-SMuRFS [148], Greenlaw *et al.* [161] proposed a Bayesian group sparse multitask regression (BGSMTR) model for identifying multi-SNP–multi-QT associations while embracing the group structure (e.g., LD blocks and genes) within the SNP data. While G-SMuRFS only provided a point estimate of the regression coefficients, BGSMTR was proposed to allow for full posterior inference, such as obtaining interval estimates for the regression parameters. The model was designed as an adapted version of the Bayesian group lasso [162], [163] to accommodate multivariate responses as well as variable selection at both SNP and gene levels. The empirical study was performed on an ADNI sample to predict 56 imaging QTs using 486 SNPs from 33 AD candidate genes.

There are also Bayesian models designed for reduced-rank regression. Zhu *et al.* [164] proposed a Bayesian generalized low-rank regression (GLRR) model for analyzing both high-dimensional imaging responses and covariates. Similar to SRRR, GLRR used a low-rank representation to approximate the high-dimensional weight matrix. It also modeled the high-dimensional covariance matrix of imaging responses with a dynamic factor model. Bayesian local hypothesis testing was proposed to identify significant SNP effects on imaging QTs while controlling for multiple comparisons. An efficient MCMC algorithm was developed for posterior computation. The empirical study was performed on an ADNI sample to evaluate the effects of 1071 SNPs from 40 AD candidate genes on 93 ROI-based volume measures.

Lu *et al.* [165] extended the above-mentioned GLRR model [164] into a Bayesian longitudinal low-rank regression (L2R2) model for examining genetic effects on longitudinal imaging responses. L2R2 includes three innovative components. The first one is a low-rank matrix to approximate regression weight matrices and gene–age interaction. The second one is to use penalized splines for characterizing the overall time effect. The third one is a sparse factor analysis model coupled with random effects to embrace spatiotemporal correlations of longitudinal imaging QTs. An efficient MCMC algorithm was used for posterior computation. The empirical study was performed on an ADNI sample to evaluate the effects of 1071 SNPs from 40 AD candidate genes on longitudinal imaging measures of 93 ROIs.

NN models, despite underexplored in brain imaging genomics, have started to attract recent attention. Wang *et al.* [166] proposed an Additive Model via Feedforward Neural networks with random weight (FNAM). This model was inspired by and adapted from the feedforward neural networks with random weight (FNNRWs) [167] to enjoy the advantages of: 1) modeling the nonlinear associations between SNPs and QTs and 2) computational efficiency over neural nets with backpropagation. The improvement of FNAM over FNNRW is that FNAM considers the role of each feature independently in the prediction and, thus, one can estimate its contribution to help model interpretation. The empirical study was performed on an ADNI sample to examine the genetic effects of 3123 SNPs from 153 AD candidate genes on 90 VBM measures and 90 FreeSurfer measures.

E. Summary

Table 7 summarizes multivariate analysis methods used in the studies discussed earlier, which aim to reveal complex imaging genomics associations between multivariate SNP data and imaging QT data. At a high level, the methods discussed in Sections V-A–V-C share a common rationale: they all use regularized regression models to relate SNPs to imaging QTs. While the SMR models in Section V-A aim to identify multi-SNP–single-QT associations, the SMMR models in Section V-B and the SRRR models in Section V-C are designed to identify multi-SNP–multi-QT associations. The SRRR models may be thought of as a special case of the SMMR models, where the regression coefficient matrix $\mathbf{W}$ in SMMR is explicitly described as a low-rank version $\mathbf{W} = \mathbf{BA}^T$ in SRRR. In general, these models share some common benefits: 1) the regression coefficients directly capture the SNP–QT relations and thus are easy to interpret and 2) using a single model to analyze the studied SNP and QT data eliminates the need for multiple testing correction and improves the detection power. One pitfall with these models is the high dimensionality of the data, which increases the risk of overfitting. To address this challenge, various regularizations are used in these models to simplify model complexity, incorporate biologically meaningful structure, and thus reduce the overfitting risk. For example, sparsity can be imposed by using the l_1- or $l_{2,1}$-norm to simply model complexity (e.g., in G-SMuRFS and SRRR). Meaningful biological structures (e.g., LD block, gene, and pathway) can be embraced by using group lasso or group $l_{2,1}$-norm (e.g., in P-GLAW, TGSL, and P-SRRR). Rank minimization can also be modeled as a regularization term (e.g., in TCLSR and TSAL) to address spatial

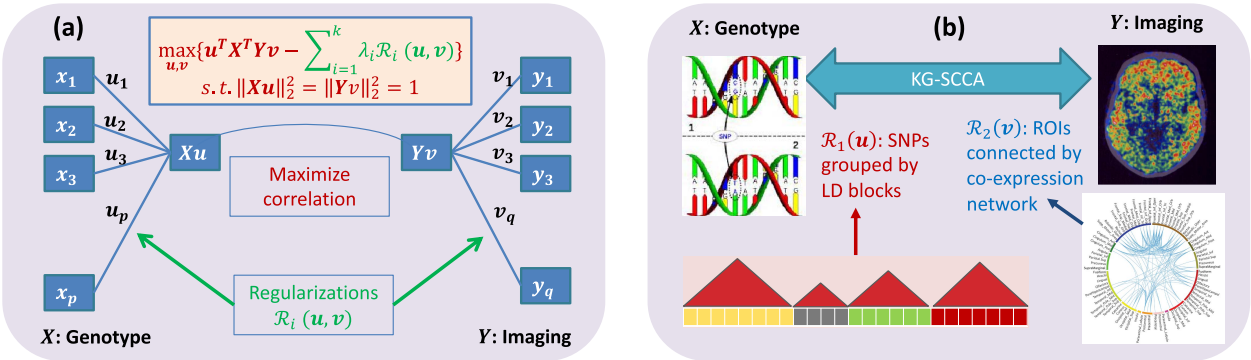


Fig. 8. *(a) Schematic of a generic reguarlized CCA framework for brain imaging genomics, which aims to find a genetic component Xu and an imaging component Yv so that their correlation (i.e., $u^T X^T Yv$ s.t. $\|Xu\|_2^2 = \|Yv\|_2^2 = 1$) is maximized under one or more regularizations $\mathcal{R}_i(u,v)$. For example, the conventional SCCA model [168] is formed by introducing two l_1-norm terms: $\mathcal{R}_1(u) = \|u\|_1$ and $\mathcal{R}_2(v) = \|v\|_1$. (b) Schematic of KG-SCCA [108]. Two regularizations are introduced into the reguarlized CCA framework shown in (a). On the genomic side, $\mathcal{R}_1(u)$ is a group l_1 term, where SNPs are grouped by LD blocks. On the imaging side, $\mathcal{R}_2(v)$ is a network-guided regularization term (similar to graph Laplacian), where ROIs are connected if they share similar coexpression patterns across the genes from the amyloid pathway. Network inset image is reproduced here with permission from Oxford University Press [108].*

or temporal correlation and reduce model complexity. Besides the above-mentioned regression models, Bayesian methods have also been studied to achieve similar goals. NN methods, although underexplored in this field have started to appear to address brain imaging genomics problems.

VI. IMAGING GENOMICS ASSOCIATIONS: BIMULTIVARIATE CORRELATION

Besides regression models, another category of prominent methods developed for brain imaging genomics studies are bimultivariate correlation models, such as sparse CCA (SCCA) and parallel-independent component analysis (pICA). Similar to the regression model discussed earlier, the sparsity is also encouraged in these correlation models to reduce model complexity and the risk of overfitting, as well as identify relevant biomarkers. Here, we discuss a few example studies using these strategies to identify complex multi-SNP–multi-QT associations. We will cover: 1) fundamental SCCA models; 2) enhanced SCCA models; 3) multimodal and longitudinal SCCA models; and 4) other bimultivariate correlation models.

A. Fundamental SCCA Models

Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be the genetic data with p variables on n subjects. Let $\mathbf{Y} \in \mathbb{R}^{n\times q}$ be the imaging data with q variables on n subjects. We assume that each column of $\mathbf{X}$ and $\mathbf{Y}$ is normalized with zero mean and unit variance. The most popular bimultivariate correlation models used in brain imaging genomics are SCCA and its variants with various regularization terms. These models can typically be described using the following generic regularized CCA form:

$$\max_{\mathbf{u},\mathbf{v}} \ \mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{v} - \Sigma_{i=1}^{k}\lambda_i\mathcal{R}_i(\mathbf{u},\mathbf{v})$$
$$\text{s.t. } \|\mathbf{X}\mathbf{u}\|_2^2 = \|\mathbf{Y}\mathbf{v}\|_2^2 = 1. \tag{16}$$

A schematic of this reguarlized CCA framework is shown in Fig. 8(a) in the context of brain imaging genomics. The goal is to find a genetic component $\mathbf{Xu}$ (i.e., a linear combination of the SNPs) and an imaging component $\mathbf{Yv}$ (i.e., a linear combination of the imaging QTs) so that their correlation (i.e., $\mathbf{u}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{v}$ s.t. $\|\mathbf{X}\mathbf{u}\|_2^2 = \|\mathbf{Y}\mathbf{v}\|_2^2 = 1$) is maximized under one or more regularization terms $\mathcal{R}_i(\mathbf{u},\mathbf{v})$. For example, the conventional SCCA model [168] is formed by introducing two l_1-norm terms: $\mathcal{R}_1(\mathbf{u}) = \|\mathbf{u}\|_1$ and $\mathcal{R}_2(\mathbf{v}) = \|\mathbf{v}\|_1$. Various other regularization terms can be defined to achieve different goals, such as incorporating group/network structure or other prior knowledge in brain imaging genomics data. In the following, we discuss a few example studies using regularized SCCA strategies.

Du *et al.* [169] proposed a structure-aware SCCA (SCCA) model by introducing into (16) two group l_1-norms: $\mathcal{R}_1(\mathbf{u}) = \Sigma_{g\in\mathcal{G}_1}\|\mathbf{u}_g\|_2$ and $\mathcal{R}_2(\mathbf{v}) = \Sigma_{g\in\mathcal{G}_2}\|\mathbf{v}_g\|_2$. The LD blocks were used to form the SNP grouping structure $\mathcal{G}_1$. The ROIs were used to form the voxelwise imaging QT grouping structure $\mathcal{G}_1$. An empirical study was performed on an ADNI sample to identify multi-SNP–multi-QT associations between the voxelwise QTs and APOE SNPs.

Yan *et al.* [108] proposed a knowledge-guided SCCA (KG-SCCA) by introducing into (16) the following two regularization terms [see Fig. 8(b)]. On the genomic side, $\mathcal{R}_1(\mathbf{u})$ is a group l_1 term, where SNPs are grouped by LD blocks. On the imaging side, $\mathcal{R}_2(\mathbf{v})$ is a network-guided

Table 7 Example Studies Using Multivariate Regression, Which Aim to Reveal Complex Imaging Genomics Associations Between Multivariate SNP Data and Imaging QT Data

Ref	Notes
[106]	P-GLAW (pathways group lasso with adaptive weights), multi-SNP-single-QT, group lasso, SNPs grouped by pathway
[146]	TGSL (tree-guided sparse learning), multi-SNP-single-QT, group lasso, tree-like group structure (SNPs grouped by LD block, LD blocks grouped by gene)
[147]	DAMM (diagnosis-aligned multimodal regression), single-SNP-multi-QT, select ROIs with genetic effects at most modalities, learning diagnosis-related components in the projected space
[148]	G-SMuRFS (Group-Sparse Multi-task Regression and Feature Selection), multi-SNP-multi-QT, group $l_{2,1}$ for feature selection at LD block level, $l_{2,1}$ for feature selection at SNP level.
[151]	TCLSR (task-correlated longitudinal sparse regression), longitudinal imaging QTs to predict SNPs, each time point treated as a task, trace norm for weight matrix rank minimization, $l_{2,1}$ norm for selecting imaging QTs with effects at most of the time points
[149]	TSAL (temporal structure auto-learning), longitudinal imaging QTs to predict SNPs, Schatten p-norm for weight matrix rank minimization, $l_{2,0+}$ norm for selecting imaging QTs with effects at most of the time points
[153]	JPLSR (joint projection learning and sparse regression), multi-SNP-multi-QT, projecting SNP and QT data into a joint latent space, SNP and QT components aligned with diagnosis, $l_{2,1}$ norm for selection of SNP and QT features
[154]	SRRR (sparse reduced rank regression), multi-SNP-multi-QT, reduced rank loss function, l_1 norm for selecting SNP and QT features, evaluation on ROI-based simulation data
[155]	SRRR (sparse reduced rank regression), multi-SNP-multi-QT, reduced rank loss function, penalized LDA to select diagnosis-related QT, l_1 norm and re-sampling for SNP identification, evaluation on voxelwise ADNI data
[107]	P-SRRR (pathways SRRR), integration of P-GLAW and SRRR, group lasso on SNP side, SNPs grouped by pathway, identifying QT-related pathways
[156]	S-SRRR (structured SRRR), reduced rank loss function, $l_{2,1}$ norm for selecting SNP and QT features
[157]	GRS-SRRR (graph-regularized S-SRRR), incorporation of graph self-representation on the SNP side into S-SRRR
[159]	RGRS-SRRR (robust GRS-SRRR), robust version of reduced rank loss function and graph self-representation loss function
[161]	BGSMTR (Bayesian group sparse multi-task regression), variable selection at both SNP and gene level, full posterior inference
[164]	GLRR (Bayesian generalized low rank regression), low rank approximation of weight matrix, dynamic factor model for imaging covariance, efficient MCMC algorithm for posterior computation
[165]	L2R2 (Bayesian longitudinal low rank regression), SNP effects on longitudinal imaging QTs, low rank approximation of weight matrix and gene-age interaction, penalized splines for overall time effect, efficient MCMC algorithm for posterior computation
[166]	FNAM (Additive Model via Feedforward Neural networks with random weight), modeling non-linear associations, computational efficiency, flexibility and interpretability of additive models

regularization term (similar to graph Laplacian), where ROIs are connected if they share similar coexpression patterns across the genes from the amyloid pathway. AHBA [170] was used to get the gene expression data across the brain. An empirical study was performed on an ADNI sample to identify multi-SNP–multi-QT associations between amyloid imaging QTs and APOE SNPs.

B. Enhanced SCCA Models

As shown in Section VI-A, there are three types of regularizations used in SCCA models: 1) l_1-norm for flat sparsity; 2) group l_1-norm for group sparisity; and 3) graph Laplacian-type norm to encourage the joint selection of features connected in a graph. In the following, we discuss a few enhanced SCCA models that are designed to improve some of the above-mentioned norms.

Du *et al.* [171] proposed an SCCA framework using a generic nonconvex penalty (GNC-SCCA) to address the challenge that the l_1-norm overpenalizes large coefficients and may introduce estimation bias. They tested seven nonconvex penalties for replacing the l_1 term in an l_1-based SCCA. These nonconvex penalties were designed to reduce the estimation bias. An empirical study was performed on an ADNI sample to identify multi-SNP–multi-QT associations between voxelwise QTs and 163 SNPs from AD genes.

Although the ideal sparsity-inducing term is l_0-norm, it is computationally intractable. Thus, l_1-norm is typically used to approximate l_0-norm. Given that the truncated l_1-norm better approximates l_0, Du *et al.* [172] proposed a truncated l_1-norm penalized SCCA (TLP-SCCA) via replacing l_1-norm with truncated l_1-norm and a truncated group lasso SCCA (TGL-SCCA) via replacing group lasso with truncated group lasso. An empirical study was performed on an ADNI sample to identify multi-SNP–multi-QT associations between voxelwise QTs and 58 SNPs from AD-related genes, where QTs were grouped by ROI and SNPs were grouped by LD block.

GraphNet was proposed in [173] as a regression model with combined graph Laplacian and l_1-norm regularization terms $\|\mathbf{u}\|_{GN} = \mathbf{u}^T\mathbf{L}\mathbf{u} + \beta\|\mathbf{u}\|_1$ where $\mathbf{L}$ is the Laplacian matrix of a given graph. Du *et al.* [174] proposed an absolute value-based GraphNet SCCA (AGN-SCCA) model, which incorporated an extended version of GraphNet regularization into the SCCA framework. The AGN regularizations are modeled as follows:

$$\mathcal{R}_1(\mathbf{u}) = \|\mathbf{u}\|_{\text{AGN}} = |\mathbf{u}|^T\mathbf{L}_1|\mathbf{u}| + \beta_1\|\mathbf{u}\|_1 \tag{17}$$

$$\mathcal{R}_2(\mathbf{v}) = \|\mathbf{v}\|_{\text{AGN}} = |\mathbf{v}|^T\mathbf{L}_2|\mathbf{v}| + \beta_2\|\mathbf{v}\|_1 \tag{18}$$

where $\mathbf{L}_1$ and $\mathbf{L}_2$ are Laplacian matrices of the correlation matrices of $\mathbf{X}$ and $\mathbf{Y}$. Here, they used data-driven correlation as graph constraint to encourage the selection of correlated features together. The newly added absolute value operation allows for the joint selection of both positively and negatively correlated features. An empirical study was performed on an ADNI sample to identify multi-SNP–multi-QT associations between ROI-based imaging QTs and 58 SNPs from AD-related genes.

Gossmann *et al.* [175] proposed a FDR-corrected SCCA (FDR-SCCA) procedure to introduce an FDR concept to SCCA and develop a method to control FDR. The existing SCCA methods determine the sparsity parameter using model fit criteria, such as cross validation and permutation. There is a lack of theoretical results to identify an appropriate level of sparsity for true signal discovery. This article proposed a method to define the FDR for canonical weight vectors in SCCA and used it as a statistical criterion to determine the model sparsity level. An empirical study was performed on an imaging genomics sample from the Philadelphia Neurodevelopmental Cohort (PNC) [176] to relate nearly 100 000 SNPs to nearly 5000 functional connectivity measures extracted from the fMRI data.

C. Multimodal and Longitudinal SCCA Models

The SCCA models discussed earlier aim to relate the SNP data to the imaging data of one single modality at one single time point. Attempts have also been made to extend these models to handle multimodal or longitudinal imaging data. We review a few example studies here.

Du *et al.* [177] proposed a multitask SCCA (MTSCCA) model to identify bimultivariate associations between SNP data and multimodal imaging data. Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be the SNP data and $\mathbf{Y}_j \in \mathbb{R}^{n\times q}$ ($j \in [1, M]$) be the imaging data of M modalities. MTSCCA is designed as

$$\begin{aligned}\max_{\mathbf{U},\mathbf{V}}\ & \Sigma_{j=1}^{M}\mathbf{u_j^T X^T Y_j v_j} - \lambda_1\|\mathbf{U}\|_{2,1} \\ & - \lambda_2\|\mathbf{U}\|_{G_{2,1}} - \lambda_3\|\mathbf{V}\|_{2,1} \\ \text{s.t. } & \|\mathbf{X}\mathbf{u}_j\|_2^2 = \|\mathbf{Y}_j\mathbf{v}_j\|_2^2 = 1 \end{aligned} \tag{19}$$

where $\mathbf{U} = [\mathbf{u}_1\ \mathbf{u}_2\ \cdots\ \mathbf{u}_M]$ and $\mathbf{V} = [\mathbf{v}_1\ \mathbf{v}_2\ \cdots\ \mathbf{v}_M]$. Here, the canonical correlation is maximized for each modality separately. The first regularization $\|\mathbf{U}\|_{2,1}$ is an $l_{2,1}$ term for SNP feature selection. The second regularization $\|\mathbf{U}\|_{G_{2,1}}$ is a group $l_{2,1}$ term for SNP feature selection at the group level (e.g., LD blocks). The third regularization $\|\mathbf{V}\|_{2,1}$ is an $l_{2,1}$ term for imaging feature selection across all the modalities. A fast optimization algorithm was implemented and applied to an ADNI sample to identify associations between over 150 000 SNPs from chromosome 19 and ROI-based QTs from three imaging modalities (VBM, FDG-PET, and Amyloid-PET).

Hao *et al.* [178] proposed a temporally constrained group SCCA (TG-SCCA) model to identify genetic association with longitudinal imaging QTs. Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be the SNP data and $\mathbf{Y}_j \in \mathbb{R}^{n\times q}$ ($j \in [1, t]$) be the imaging data at t time points. TG-SCCA is designed as

$$\begin{aligned}\max_{\mathbf{u},\mathbf{V}}\ & \Sigma_{j=1}^{t}\mathbf{u^T X^T Y_j v_j} - \lambda_1\|\mathbf{u}\|_1 \\ & - \lambda_2\|\mathbf{V}\|_{2,1} - \lambda_3\Sigma_{j=1}^{t-1}\|\mathbf{v}_{j+1} - \mathbf{v}_j\|_1 \\ \text{s.t. } & \|\mathbf{X}\mathbf{u}\|_2^2 = \|\mathbf{Y}_j\mathbf{v}_j\|_2^2 = 1 \end{aligned} \tag{20}$$

where $\mathbf{V} = [\mathbf{v}_1\ \mathbf{v}_2\ \cdots\ \mathbf{v}_t]$. Here, the canonical correlation is maximized for each time point separately while maintaining the genetic component the same across all the time points. The first regularization $\|\mathbf{u}\|_1$ is an l_1-norm for SNP feature selection. The second regularization $\|\mathbf{V}\|_{2,1}$ is an $l_{2,1}$ term for imaging feature selection across all the time points. The third regularization $\Sigma_{j=1}^{t-1}\|\mathbf{v}_{j+1} - \mathbf{v}_j\|_1$ is a fused lasso term to constrain the weight difference between two neighboring time points. An empirical study was performed on an ADNI sample to identify associations between 85 APOE SNPs and longitudinal VBM QTs from 116 ROIs at four time points.

Du *et al.* [179] proposed another longitudinal imaging genetics model based on MTSCCA [177]. It is named temporal MTSCCA (T-MTSCCA) and designed as

$$\begin{aligned}\max_{\mathbf{U},\mathbf{V}}\ & \Sigma_{j=1}^{t}\mathbf{u_j^T X^T Y_j v_j} - \lambda_1\mathcal{R}_1(\mathbf{U}) - \lambda_2\mathcal{R}_2(\mathbf{V}) \\ \text{s.t. } & \|\mathbf{X}\mathbf{u}_j\|_2^2 = \|\mathbf{Y}_j\mathbf{v}_j\|_2^2 = 1 \end{aligned} \tag{21}$$

where $\mathbf{U} = [\mathbf{u}_1\ \mathbf{u}_2\ \ldots\ \mathbf{u}_t]$ and $\mathbf{V} = [\mathbf{v}_1\ \mathbf{v}_2\ \ldots\ \mathbf{v}_t]$. Here, the canonical correlation is maximized for each time point separately. The regularization $\mathcal{R}_1(\mathbf{U}$ on the genomic side contains three components: one l_1-norm and one $l_{2,1}$-norm for feature selection at SNP level, and one group $l_{2,1}$-norm for feature selection at group level (e.g., LD block). The regularization $\mathcal{R}_2(\mathbf{V}$ on the imaging side contains three components: 1) an l_1-norm for imaging feature selection using flat sparsity; 2) an $l_{2,1}$-norm for selection imaging features associated at most time points; and 3) a fused pairwise $l_{2,1}$-norm ($\text{FP}_{2,1}$-norm) for joint selection of the same QT at neighboring time points. Compared with the nonconvex fused lasso used in TG-SCCA [178], $\text{FP}_{2,1}$-norm is convex and thus easy to optimize. An empirical study was performed on an ADNI sample to identify associations between 1085 APOE SNPs and longitudinal VBM QTs from 90 ROIs at four time points.

D. Other Bimultivariate Correlation Models

We now discuss a few other bimultivariate correlation models. Le Floch *et al.* [180] proposed a two-step procedure, named FSPLS (filtering + sparse partial least squares), to identify associations between high-dimensional SNP and imaging QT data (e.g., empirical study of real data including 94 subjects with 600 000 SNPs and 34 fMRI QTs). The first step of FSPLS selected top SNPs with minimal p-values via massive univariate association analysis between each SNP–QT pair using linear regression based on an additive genetic model. The second step of FSPLS applied a single sparse partial least squares (SPLS) model to the selected SNP data and full QT data to identify a multi-SNP–multi-QT association. Empirical studies on both simulated and real high-dimensional SNP and imaging QT data demonstrated that FSPLS outperformed several competing methods using other regularization and dimensionality reduction strategies coupled with PLS or kernel CCA models. This article also illustrated that the SRRR, SCCA, and SPLS models are mathematically equivalent methods, up to specific assumptions on the covariance matrix.

Fang *et al.* [181] proposed a greedy projected distance correlation (G-PDC) method to examine pairwise gene–ROI associations, where each gene contains a number of SNPs and each ROI contains a number of voxels. Distance correlation measures statistical dependence between two random vectors (e.g., gene versus ROI) and can model nonlinear relationship between them. Projected distance correlation measures conditional dependence based on distance correlation [182]. In this article, given a gene–ROI

Table 8 Example Studies Using Bimultivariate Correlation Methods, Which Aim to Identify Multi-SNP–Multi-QT Associations From High-Dimensional Imaging Genomic Data

Ref	Notes
[169]	S2CCA (structure aware SCCA), group l_1 norm on both SNP and QT sides, SNPs grouped by LD block, QTs grouped by ROI
[108]	KG-SCCA (knowledge-guided SCCA), group l_1 norm on genetic side (SNPs grouped by LD block), graph Laplacian type norm on imaging side (ROIs connected by co-expression network)
[171]	GNC-SCCA (generic non-convex penalty SCCA), seven non-convex penalties replacing l_1 norm to reduce estimation bias
[172]	TLP-SCCA (truncated l_1-norm penalized SCCA), TGL-SCCA (truncated group lasso SCCA), better approximation of l_0 norm, voxels grouped by ROI, SNPs grouped by LD block
[174]	AGN-SCCA (absolute value based GraphNet SCCA), incorporation of a GraphNet variant into SCCA, joint selection of both positively and negatively correlated features
[175]	FDR-corrected SCCA, incorporation of FDR concept into SCCA
[177]	MTSCCA (multi-task SCCA), relating SNP to multimodal imaging QTs, $l_{2,1}$ norm for SNP selection and QT selection
[178]	TG-SCCA (temporally constrained group SCCA), l_1 for SNP selection, $l_{2,1}$ for ROI selection (over time), fussed lasso for smoothing weights between neighbouring time points
[179]	T-MTSCCA (temporal multi-task SCCA), l_1 and $l_{2,1}$ for SNP and QT selection, fused pairwise $l_{2,1}$ norm for smoothing weights between neighbouring time points
[180]	FSPLS (filtering + sparse Partial Least Square), two step procedure, univariate filtering, sparse PLS with l_1 regularization
[181]	G-PDC (Greedy projected distance correlation), examination of pairwise gene-ROI associations, an efficient algorithm
[183]	DCCA (Distance CCA), identification of SNP set and QT set with the highest distance correlation
[186]	pICA (parallel independent component analysis), joint maximization of within-modality component independence and between-modality component correlation

pair, the goal is to test their independence while controlling for all the other SNPs and voxels. Fang *et al.* [181] proposed an efficient G-PDC algorithm to enable large-scale imaging genomics analysis. An empirical study was performed on the PNC data [176] to examine the pairwise association between 221 ROIs (containing 27 168 voxels) and 2035 genes (containing 63 010 SNPs).

Hu *et al.* [183] integrated distance correlation model into the CCA framework and proposed a distance CCA (DCCA) method. The G-PDC method described earlier performs pairwise analysis for each possible gene–ROI combination and is still facing large burden for multiple testing correction. The DCCA model was proposed to overcome this limitation by identifying a set of original SNPs and a set of original imaging QTs with the highest distance correlation. The approach was to first construct a distance kernel function and then solve an optimization problem. An empirical study was performed on the PNC data [176] to examine the pairwise association between 264 ROIs (containing 27 384 voxels) and 736 genes (containing 21 487 SNPs).

pICA [184], [185] is another well-established strategy for mining multi-SNP–multi-QT associations. It is a joint estimation procedure to extract imaging components and genetic components for achieving two goals: 1) maximizing independence among components within each modality using an entropy term and 2) maximizing components' correlation between two modalities. Meda *et al.* [186] applied pICA to an ADNI sample for identifying multi-SNP–multi-QT associations between the genome-wide SNPs and brain-wide ROI-based imaging QTs.

E. Summary

Table 8 summarizes bimultivariate correlation methods used in the studies discussed earlier, which aim to identify multi-SNP–multi-QT associations from high-dimensional imaging genomic data. Most of these strategies are regularized SCCA models. Similar to the regression models in Section V, these SCCA models also employ l_1 or $l_{2,1}$-norm for feature selection, group l_1 or $l_{2,1}$-norm for feature selection at group level, and graph Laplacian for graph-guided learning. Multimodal and longitudinal SCCA models often include $l_{2,1}$-norm for feature selection across modalities or time points as well as fussed lasso or fused pairwise $l_{2,1}$-norm for smoothing neighboring weights along the temporal dimension. Other bimultivariate correlation models include: 1) SPLS that is mathematically equivalent to SRRR and SCCA under certain assumptions on the covariance matrix; 2) distance correlation that can model nonlinear associations; and 3) parallel ICA models for joint maximization of within-modality component independence and between-modality component correlation.

VII. INTEGRATING IMAGING AND GENOMICS FOR OUTCOME PREDICTION

In addition to identifying imaging genomics associations, another active research topic in brain imaging genomics is how to integrate brain imaging and genomics data for prediction of outcomes of interest, such as disease stage, impairment score, and progression status. A relevant interesting topic is to learn the associations among genomics, imaging, and the outcome to help understand the biological pathway from genetics to brain structure and function, and to cognitive, behavior, and diagnostic outcomes. In this section, we first focus on methods for outcome prediction and then review methods for joint association learning and outcome prediction.

A. Outcome Prediction

We discuss a few example studies using the existing conventional prediction models, newly developed machine learning approaches, and state-of-the-art deep learning methods. Of note, all these studies were performed using brain imaging genomics data from the ADNI cohort.

We start with some studies using conventional predictive models. For example, Dukart *et al.* [187] examined the role of multimodal imaging (MRI, FDG-PET, and Amyloid-PET), neuropsychological, and genetic data as potential biomarkers for identifying MCI patients who will convert to AD in the future. They first built naive Bayes classifiers to distinguish AD and CN participants using different combinations of the above-mentioned data modalities. After

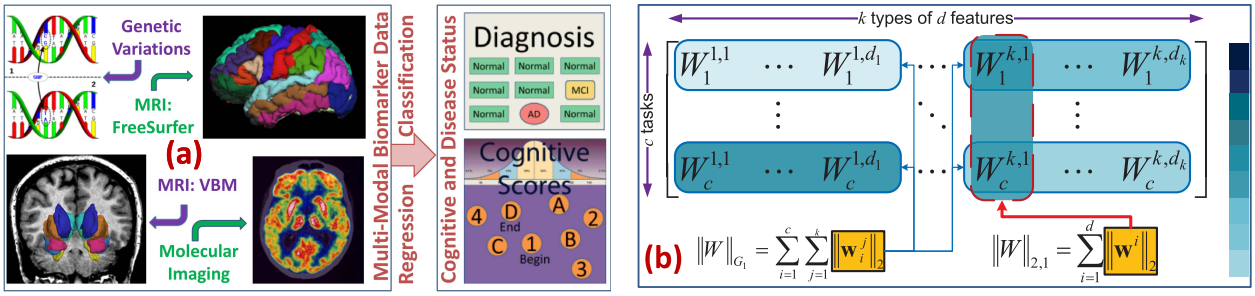


Fig. 9. *(a) JCRMML framework [193] performs joint classification and regression via multimodal multitask learning to identify disease-sensitive and cognition-relevant biomarkers from brain imaging genomic data. The identified biomarkers could predict not only disease status but also cognitive functions to help us better understand the underlying mechanism from gene to brain structure and function and to cognition and disease. (b) Illustration of the JCRMML feature weight matrix W^T. The group l_1-norm (G_1-norm) learns the groupwise weights for features within a single modality for each task (i.e., outcome) and the $l_{2,1}$-norm selects features associated with most tasks. Images are reproduced here with permission from Oxford University Press [193].*

that, they applied the learned classifier to the MCI cohort for predicting AD conversion status. They achieved 76% accuracy using FDG-PET data and 87% accuracy using multimodal imaging and genetic data. This shows the promise of the data integration strategy in the context of AD outcome prediction.

Filipovych *et al.* [188] proposed a method to create a composite imaging genetic score for predicting MCI conversion to AD. On the imaging side, they used a nonlinear pattern recognition method "COMPARE" [189] to identify AD-relevant volumetric regions. After that, a nonlinear SVM was applied to imaging measures from these regions to get an imaging score for each individual. On the genomic side, a linear SVM was used to classify AD vs CN, which yielded a polygenic AD-related genetic score for each subject. Finally, a composite imaging genetic score was created as a weighted sum of the imaging score and the genetic score. The empirical study showed that the proposed composite score improved the prediction accuracy.

Kauppi *et al.* [190] performed survival analysis using the Cox proportional Hazard model to predict time to progression from MCI to AD via integrating a PHS, an imaging-based atrophy score, and the MMSE score. The PHS was generated using the ADGC data [66], as described in Section III-B. The atrophy score was generated from volumetric measures of a few AD-related ROIs using an LDA to distinguish AD versus CN (see [191] and [192] for more details). The empirical study showed that combining PHS with atrophy and MMSE significantly improved the prediction performance compared with models without PHS.

Besides conventional prediction methods, new machine learning models have also been proposed for outcome prediction using brain imaging genomics data. For example, Wang *et al.* [193] proposed a joint classification and regression framework for multimodal multitask learning (JCRMML). JCRMML was designed to use multimodal imaging (MRI and FDG-PET) and genetic data for joint prediction of diagnositic and cognitive outcomes and, at the same time, to identify disease-sensitive and cognition-relevant imaging and genetic biomarkers [see Fig. 9(a)]. It is formulated as a regularized multivariate linear model with feature weight matrix $\|\mathbf{W}^T\|$ shown in Fig. 9(b), where a task indicates an outcome response. The loss function includes a logistic regression component for disease classification and a linear regression component for cognitive score regression. JCRMML has two regularization terms. One group l_1 term $\|\mathbf{W}\|_{G_1}$ is used for learning groupwise weights for features within a single modality for each task (i.e., a diagnostic or cognitive outcome). One $l_{2,1}$ term $\|\mathbf{W}\|_{2,1}$ is used for selecting features associated with most tasks (i.e., outcomes). The empirical study yielded improved performance on prediction both diagnostic and cognitive outcomes compared with several competing methods.

Zhang *et al.* [194] examined several machine learning strategies for AD prediction via combining multimodal imaging (MRI and FDG-PET), CSF, and SNP data. Specifically, they compared three state-of-the-art feature selection methods. The first is a multiple kernel learning (MKL) method named SimpleMKL [195]. The second is a high-order graph matching-based feature selection (HGM-FS) [196]. The third is sparse multimodel learning (SMML) [197]. The AD prediction model was learned in three steps: 1) a feature selection method was applied to select discriminative features; 2) each selected feature was multiplied by its learned weight to form a new feature vector; and 3) a linear SVM was applied to the new feature vectors to learn a predictor. Empirical studies yielded a few findings: 1) FDG-PET was the modality with the best prediction accuracy; 2) adding SNP data to other modalities could improve prediction accuracy; and 3) HGM-FS worked the best among three feature selection methods.

Peng *et al.* [198] proposed a structured sparse kernel learning (SSKL) model for AD prediction using multimodal imaging (MRI and FDG-PET) and SNP data. They described each feature with a kernel and used the modality

information to group kernels to facilitate variable selection at both feature and group levels. An innovative structured sparsity regularization term was further introduced to enable feature sparsity within each modality but encourage nonsparse solution modality wisely. The intuition is based on the hypothesis that different modalities offer complementary information and including modalities with weaker predictive power may help capture valuable complementary information. Their empirical study yielded promising results.

Singanamalli *et al.* [199] proposed a Cascaded Multiview Canonical Correlation (CaMCCo) for classifying CN, MCI, and AD using multidimensional imaging, genetics, biomarker, and cognitive data. The cascaded framework first classified all subjects as CN versus cognitively impaired (CI) and further classified CI subjects as MCI versus AD. For each binary classification, the class label was used as a separate variable set. Integrating the class label set with all the other modalities, supervised multiview CCA (sMVCCA) [200] was employed to obtain a low-dimensional representation of each involved modality, followed by a modality selection step using the diagnostic information. Naive Bayes classification method was then applied to the fused representation of selected modalities to learn a classifier. Empirical study showed that fusion of selected modalities outperformed that using each individual modality and that integrating all the modalities.

Although NN models have been highly successful in making prediction for many recent applications in various fields such as computer vision and natural language processing, they have not been widely used in brain imaging genomics. This could be largely attributed to the limited sample size and high dimensionality of the existing imaging genomics data. Some attempts have been made to address this challenge. In the following, we review a couple of recent studies using NN methods for AD outcome prediction via integrating brain imaging genomics data.

For example, Zhou *et al.* [201] presented a three-stage deep feature learning and fusion framework to detect disease status (e.g., CN/MCI/AD) via integrating MRI, FDG-PET, and SNP data. In the first stage, they learned feature representation for each modality independently. In the second stage, they used the features learned in Stage 1 to learn joint latent features for each pair of modalities. In the third stage, they learned the diagnostic label using the features learned in Stage 2. This framework can address several challenges: 1) learning high-level features for each modality in Stage 1 could alleviate data heterogeneity issue and 2) using the maximum number of all available samples at each stage could help address both the high-dimension-low-sample-size and incomplete modality data issues. Their empirical study showed very promising results that the proposed NN method outperformed a number of non-NN-based competing methods.

Ning *et al.* [202] proposed another NN framework to detect AD or MCI-to-AD conversion using MRI and SNP data. Their strategy to address high-dimension-low-sample-size is twofold: 1) instead of examining all the SNPs and imaging QTs, their analysis only targeted at 16 AD-related QTs and 19 AD-related SNPs to reduce the dimensionality and 2) they designed a relatively simple NN with two hidden layers, and explored 2, 4, 8, up to 64 nodes in each layer to reduce the model complexity. The proposed NN was fully connected between the layers, coupled with shortcut connections linking all the input nodes directly to the output layer. Their empirical study showed promising results that the proposed NN model outperformed a linear regression model.

Table 9 Example Studies Using Machine Learning Methods for Outcome Prediction via Integrating Imaging and Genomics Data

Ref	Notes
[187]	Naive Bayes classifier, predicting MCI-to-AD conversion
[188]	Composite multivariate polygenic and neuroimaging score, predicting MCI-to-AD conversion
[190]	Cox proportional hazard model, predicting time to progression from MCI to AD, integrating PHS, atrophy score and MMSE
[193]	JCRMML (joint classification and regression framework for multimodal multitask learning), joint logistic regression and linear regression, feature selection at modality level for each outcome, feature selection across all the outcomes
[194]	Multiple kernel learning (MKL), high-order graph matching based feature selection (HGM-FS), sparse multimodal learning (SMML), AD prediction using MRI, FDG-PET, CSF and SNP data
[198]	SSKL (structured sparse kernel learning), sparsity inside modalities, dense combination between modalities
[199]	CaMCCo (Cascaded Multi-view Canonical Correlation), supervised multiview CCA with class label as a new variable set
[201]	Stage-wise deep neural network, addressing issues such as data heterogeneity, high-dimension-low-sample-size & incomplete data
[202]	Neural network model with two hidden layers, AD outcome prediction using 16 imaging QTs and 19 SNPs

Table 9 summarizes the example studies using machine learning methods for outcome prediction via integrating imaging and genomics data. Some studies directly applied conventional learning methods to the combined data sets and showed improved performances. Some studies developed new learning models to address various challenges, such as feature selection at a group level, and joint classification and regression. With a couple of successful attempts, NN models have started to attract attention in the field of brain imaging genomics.

B. Joint Association Learning and Outcome Prediction

Here, we review a few example studies exploring the associations among genomics, imaging, and outcome. These include four SCCA-based studies [203]–[206], one study using classic mediation analysis [207] and one study using a newly proposed Bayesian method [208]. While two studies [206], [207] performed the analyses using the PNC data, the other four studies were conducted on the ADNI data.

Yan *et al.* [203] proposed a discriminative SCCA model in order to identify disease-relevant imaging proteomics associations. Instead of SNP data, Yan *et al.* [203] analyzed the protein expression data collected from CSF and plasma and studied their relationship to imaging QTs and

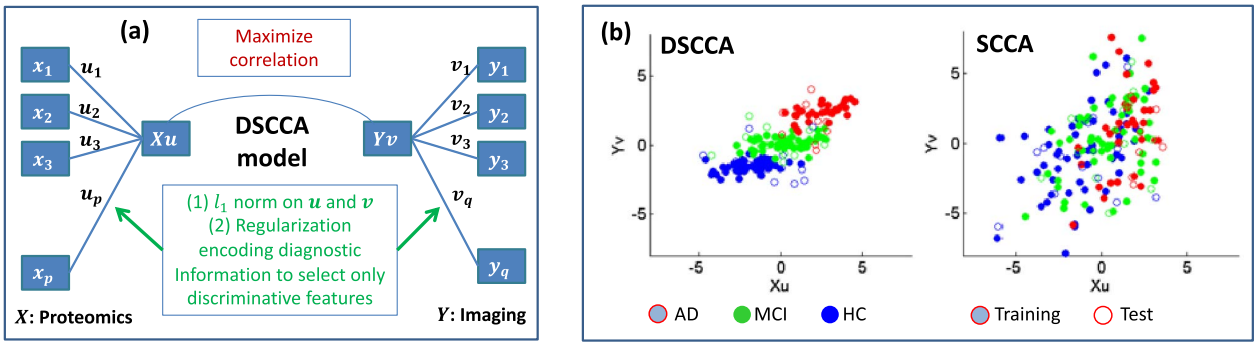


Fig. 10. *(a) Schematic of the DSCCA model [203]. DSCCA incorporates regularization into SCCA to encourage the identification of canonical components with discriminative power. (b) Imaging component Yv is plotted against the proteomic component Xu. DSCCA components are clearly more discriminative than SCCA components. Image in (b) is reproduced here from an open-access chapter by World Scientific Publishing Company [203].*

multiclass diagnostic label (CN, MCI, and AD). Fig. 10(a) shows a schematic of the DSCCA model. It introduced a new graph Laplacian regularization to the standard SCCA framework. The graph is defined on the subjects, where subjects within the same diagnositic group are connected. This regularization encourages the identification of canonical components with discriminative power. Fig. 10(b) shows a comparision between DSCCA and SCCA, where the imaging component **Yv** is plotted against the proteomic component **Xu**. It is clear that the components identified by DSCCA have more discriminative power than those by SCCA. The empirical study using cross validation showed that DSCCA yielded higher canonical coefficient (CC) on the test data than SCCA.

Hao *et al.* [204] proposed an alternative strategy to identify pairwise associations among genomics, imaging, and outcome(s). This was directly implemented by a three-way SCCA, which was a joint learning model by combining three pairwise SCCA models to learn a single component for each modality (i.e., genomics, imaging, or outcome). Two empirical studies were performed on ADNI imaging genomics data: one using a set of cognitive scores as outcome, and the other using diagnostic status as outcome. In a cross-validation setting, both studies using three-way SCCA yielded higher CCs on the test data than that using SCCA.

Du *et al.* [205] proposed a joint learning model by combining SCCA and regression (SCCAR) to identify diagnosis-relevant imaging genomics associations. Let **z** be the outcome data. The model is defined as

$$\min_{\mathbf{u},\mathbf{v}} \frac{1}{2}\|\mathbf{z}-\mathbf{Y}\mathbf{v}\|_2^2 - \mathbf{u}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{Y}\mathbf{v} + \lambda_1\mathcal{R}_1(\mathbf{u}) + \lambda_2\mathcal{R}_2(\mathbf{v})$$
$$\text{s.t. } \|\mathbf{X}\mathbf{u}\|_2^2 = \|\mathbf{Y}\mathbf{v}\|_2^2 = 1. \tag{22}$$

Here, they would like to jointly learn the imaging component **Yv** so that it could predict the outcome **z** (see the first regression term) and is correlated with the genomic component (see the second CCA term). In the empirical study, they used l_1-norm for both $\mathcal{R}_1(\mathbf{u})$ and $\mathcal{R}_2(\mathbf{v})$. The cross-validation results showed that SCCAR could identify stronger canonical correlations than SCCA in the test data.

Zille *et al.* [206] proposed a Multi-Task Collaborative Regression (MT-CoReg) method to extract outcome-relevant variables that are coexpressed in both imaging and genomics modalities. Similar to SCCAR, MT-CoReg was also formulated as a joint learning model by integrating SCCA and linear regression. The major difference is that MT-CoReg allows the imaging component used in the linear regression to predict outcome to be different from that used in the SCCA to correlated with the genetic component. An empirical study was performed on the PNC cohort to analyze the SNP and fMRI data for the study of learning ability as outcome and yielded promising results.

Bi *et al.* [207] performed a genome-wide mediation analysis in order to detect complicated mechanisms of genetic inferences on the outcome implicitly through intermediate imaging QTs. The study was performed on the PNC cohort, analyzing 445 205 SNPs, 204 imaging QTs, and 104 psychiatric and cognitive traits as outcomes. Mediation analysis was performed at the individual marker level using a three-stage procedure: 1) GWAS was performed to identify significant SNP–QT pairs; 2) each outcome was regressed against each candidate SNP; and 3) each outcome was regressed against each identified SNP and its associated QT. A mediation relationship is established if the SNP is significant in 1) and 2), QT is significant in 3), and the absolute effect size of the SNP is smaller in 3) than 1). Their analysis identified an NMNAT2 SNP associated with a psychiatric trait through the volume of the left superior frontal region.

Performing brain-wide genome-wide analysis at the single marker level faces a major challenge on multiple comparison correction. To overcome this limitation, a common approach is to learn one single multivariate multiple regression model coupled with some sparsity-inducing regularization. Batmanghelich *et al.* [208] proposed such a Bayesian method for probabilistic modeling of imaging, genetics, and diagnosis. The goal of this method is to jointly learn the following two predictive relationships in

Table 10 Example Studies for Joint Learning of Imaging Genomics Associations and Outcome Prediction Model

Ref	Notes
[203]	DSCCA (discriminative SCCA), disease-relevant imaging proteomics associations, graph laplacian
[204]	Three way SCCA among genomics, imaging and outcomes
[205]	SCCAR (joint learning by combining SCCA and regression)
[206]	MT-CoReg (Multi-Task Collabora-tive Regression), joint regression and SCCA model
[207]	Genome-wide mediation analysis, genetic influence on phenotypica outcome mediated by imaging endophenotype
[208]	Bayesian model to identify imaging QTs that have genetic basis and are associated to diagnosis

a single Bayesian model: 1) using imaging QTs to predict diagnosis and 2) using SNPs to predict imaging QTs. The joint model can help identify a set of imaging QTs that not only have a genetic basis but also are associated with diagnostic status. Their empirical study on the ADNI data yielded promising results.

Table 10 summarizes the example studies for joint learning of imaging genomics associations and outcome prediction model. Four of these studies introduced into the standard SCCA framework one or more components that incorporate outcome information. Empirical studies demonstrated that including outcome information as additional constraints could identify stronger imaging genomics associations, indicating that this strategy has the potential to capture true signals and reduce model overfitting.

VIII. CONCLUSION AND DISCUSSION

A. Summary of Learning Problems and Reviewed Methods

We have reviewed three categories of learning problems in brain imaging genomics, as shown in Fig. 1(a). In the first category, we focused on the learning problem of heritability estimation of brain imaging QTs. The heritability of a trait is defined as the proportion of its observed variance explained by the genetic factors. Given high-dimensional brain imaging data, heritability estimation can be used as a screening tool to extract heritable QTs as attractive targets for in-depth genetic analyses. We discussed two types of methods for heritability estimation: one based on data collected using twin or family designs, and the other based on genome-wide genotyping data.

In the second category, we focused on the problem of learning imaging genomics associations, a major theme studied in brain imaging genomics to gain new insights into the genetic and molecular mechanisms of the brain structure and function. Given the high-dimensionality-small-sample-size challenge that we are facing in brain imaging genomics, a wide range of methods have been proposed to increase statistical power and enhance biological interpretation via reducing dimensionality, measuring collective effects, and incorporating prior knowledge. We first reviewed a few fundamental strategies, including single-SNP–single-QT methods, PRSs, multi-SNP methods, multi-trait methods, pathway and network enrichment methods, and interaction methods. We then discussed the important topics of power and sample size and reviewed relevant meta-analysis strategies. After that, we reviewed two major types of multi-SNP–multi-QT methods: multivariate regression models and bimultivariate correlation models.

In the third category, we focused on the learning problem of integrating imaging and genomics for outcome prediction. This is an important topic studied in brain imaging genomics to gain valuable insights into the outcome-relevant neurobiological mechanisms at the genetic, molecular, and macroscale brain system levels. Imaging and genomics data capture the subject's characteristics at different scales and from different perspectives and are naturally considered to contain complementary information for improved outcome prediction. Various machine learning and deep learning methods have been proposed to address relevant data integration challenges, such as high dimensionality, small sample size, heterogeneity, and incompleteness. We reviewed these learning strategies for outcome prediction using both brain imaging and genomics data, as well as joint learning strategies that could not only identify associations between imaging and genomics data but also use them to accurately predict outcomes.

B. Biomedical Application Considerations

Fig. 1(b) summarizes some biomedical application considerations regarding the studied data sets across multiple disciplines, including brain imaging, genomics, and clinical outcome research. Careful consideration of the data characteristics and relevant biological structure and knowledge can often provide valuable guidance on the selection of an appropriate method for practical applications. A brain imaging genomics application involves the integrated analysis of brain imaging data, genomics data, and optionally clinical outcome data.

First, let us take a look at brain imaging data. Imaging QTs can be extracted from brain scans at multiple scales (e.g., voxels, ROIs, and connectivity matrix). In the following, we discuss a few example strategies for dealing with analytic challenges with these QTs. Although voxelwise analysis (see [57]) can capture the finest details in the brain, it is often underpowered due to its heavy burden of multiple comparison correction and high spatial correlation. There are several strategies to overcome this limitation: 1) use methods, such as RFT (see [72]), to reduce the multiple testing burden via embracing spatial correlation; 2) collapse voxel measures into ROI measures to greatly reduce the number of statistical tests (see [37]); 3) measure collective effect of all voxels within an ROI to reduce the test number (see [95]); and 4) perform only targeted SNP analysis (see [49]). Compared with voxelwise analysis, ROI-based analysis has a greatly reduced multiple testing burden but may not be able to capture the detailed spatial patterns. One strategy to leverage this issue is to first identify a small number of interesting SNPs from ROI-based analysis and then map their effects onto the

brain in a voxelwise fashion (see [37]). Connectivity matrices are another type of high-dimensional imaging QTs. To alleviate the multiple testing burden, besides conducting targeted QT analyses, one can perform heritability analysis to select only highly heritable connectivity QTs for in-depth genetic analysis (see [38]).

Brain imaging data can be collected with multiple modalities. Given the availability of multimodal imaging data, one can employ multimodal learning strategies (see [193]) to make full use of the complementary information offered by multiple imaging modalities. One may also consider methods, such as in [201], to address potential challenges related to multimodal imaging data (e.g., high dimensionality, small sample size, heterogeneity, and incompleteness). In addition, brain imaging data can be longitudinal. A longitudinal QT offers a unique power to capture progressive pattern a cross-sectional QT cannot describe and thus is an important biomarker to study. One simple approach could examine some summary statistics of a longitudinal QT (see [48]). One can also employ more complicated longitudinal learning models (see [149] and [151]) to identify more detailed longitudinal patterns. Finally, there are different types of prior knowledge and structure that can be used to group and connect imaging QTs. For example, voxels can be grouped by ROIs (see [169]), and ROIs can be grouped by network components (e.g., DMN [88]) and connected by brain networks (see [108]). Incorporating this prior knowledge into the learning model can help alleviate overfitting and yield biologically interpretable findings.

Second, let us take a look at the genomic data. Traditional GWAS performs univariate analysis at the SNP level, with a huge burden on multiple testing correction. To address this challenge, the following are a few possible strategies: 1) examine a few target SNPs (see [49]) or a PRS (see Section III-B); 2) perform analysis at the SNP-set level (e.g., LD block and gene) (see [71]); 3) perform enrichment analysis using pathways and networks (see Section III-E); and 4) examine a single model involving multiple SNPs (see Sections V and VI). Here, the LD blocks, genes, pathways, and functional interaction networks are biologically meaningful knowledge and structures. They can also be incorporated into the multivariate learning models to reduce overfitting and improve model interpretability.

Third, let us take a look at the clinical outcome data, such as disease stage, impairment score, and progression status. These are critical data sources for the study of brain disorders. There are several strategies to perform outcome-relevant brain imaging genomics studies. One is to first identify outcome-relevant imaging QTs and then reveal its genetic basis. This can be done as a two-step procedure (see [207]) or via a joint learning model (see [208]). The second strategy is to combine imaging and genomics data for an improved outcome prediction (see Section VII-A). The third strategy is to use outcome information to guide the search for imaging genomics associations, which can often reduce overfitting and identify stronger associations (see [203]).

C. Statistical and Machine Learning Considerations

Fig. 1(c) summarizes some statistical and machine learning considerations for brain imaging genomics. The first important consideration is the statistical power since the existing brain imaging data sets typically have high dimensionality and relatively small sample size. The following are a few strategies on how to increase study power. First, compared with case-control analyses, QT studies are shown to have increased statistical power [4], [209]. The second strategy is to employ more powerful multiple testing correction methods by taking into consideration the correlation within imaging and genomics data (see [44]). The third strategy is to increase the sample size via mega- or meta-analysis on combined data set from multiple collaborative sites (see Section IV). The fourth strategy is to reduce the test number by pooling low-level measures into high-level ones (e.g., averaging voxel measures into ROI measures and aggregating SNP statistics into gene statistics) or simply by applying a single multivariate model involving all the studied SNPs and QTs.

Another important methodological consideration is how to control overfitting and reduce spurious findings for multivariate learning models. To reduce the risk of overfitting, the data fitting flexibility of a learning model should be properly controlled. One strategy is to reduce the number of variables in the model via dimensionality reduction. For example, one can condense fine-level SNP/voxel measures into high-level gene/ROI components (see [37] and [71]). Another strategy is to include regularization terms in the model to control data fitting flexibility. For example, to increase the feature selection stability, we can group SNPs by LD block (see [169]). To help biological interpretation, we can group SNPs by gene, pathway or network, and/or ROIs by brain network (see [107], [108], and [146]). In addition, incorporating outcome information into the learning model can help select outcome-relevant SNP and QT markers and reduce overfitting (see [203]).

There are a few other methodological considerations that we briefly discuss in the following.

1) To help biological interpretation, we can incorporate prior knowledge and structure into the learning methods and try to identify associations between meaningful biological entities, such as genes, pathways, ROIs, and genetic and brain networks. One strategy is to perform GWAS enrichment analysis (see [90] and [99]) to measure collective effects at the set level. This can reduce the number of tests and increase the detection power. Another strategy is to regularize the learning model using these sources of prior knowledge and structure to guide our search for meaningful associations (see [106]–[108]

and [146]). In both cases, findings are associated with meaningful functional annotation implicating potential biological mechanism and interpretation, which make them less likely to be false discoveries.

2) Scalability is often an important consideration in BWGW studies, particularly if one wants to perform analyses at the voxelwise level. Several efficient algorithms (see [50] and [72]) have been proposed to address this consideration. One effective strategy is a GSIS procedure used in [50], which can greatly reduce the search space size from $N_s N_v$ to $\sim N_0 N_v$ for $N_0 \ll N_s$. Here, N_s is the number of SNPs and N_v is the number of voxels. Another valuable strategy is a fast permutation procedure, proposed in [72], which uses a parametric tail approximation to provide accurate p estimations in an efficient manner.
3) Biased sampling is another potential cause for spurious findings. Most GWAS studies (e.g., ADNI) are based on the case-control design, and the data are typically a biased sample of the target population. Directly correlating imaging QTs (as secondary traits) with genotype may lead to biased inference generating misleading results. This issue has been considered in several studies (see [52] and [55]). Although the standard linear analysis was found to be generally valid on the ADNI data in [52], simulation studies in [55] showed that linear regression models without adjusting for biased sampling demonstrated severely inflated Type I error rates in some cases. In general, caution should be taken while analyzing imaging QT data as secondary phenotypes in case-control studies.
4) Gene–gene interaction has also been studied to identify epistatic genetic effects on imaging QT and to help address miss heritability. Given an exponentially increasing number of possible tests, a major topic in epistatic studies is to find an effective search strategy to reduce computational time and increase statistical power. One strategy is to examine only a subset of candidate interactions with a potential biological mechanism suggested by functional interaction networks or biological pathways (see [109] and [112]). Another strategy is to perform data-driven screening to focus on the analysis of a small number of most promising candidate interactions (see [109] and [114]).

D. Scientific and Clinical Impact

Our previous reviews of ADNI brain imaging genomics findings [4], [5] indicated that numerous genes contributing to increased risk for or protection against AD have been identified and replicated using multimodal brain imaging data. These findings implicated immune, mitochondrial, cell cycle/fate, and other biological processes and advanced the mechanistic understanding of AD. In the following, we briefly discuss a few new example findings with potential scientific and clinical impacts.

According to the most recent ENIGMA review article [2], the consortium's GWAS analyses have revealed over 200 genetic loci associated with cortical thickness or surface area and over 40 common genetic variants associated with subcortical volumes. In addition, the recent UK Biobank GWAS of 3144 brain imaging QTs identified 148 clusters of SNP–QT associations [12]. These results have provided substantial new insights into the genetic landscape of the brain and offered a great scientific value that could impact and advance research on normal brain development and aging, and neurological and psychiatric disorders.

Given the timelines set in place by the National Alzheimer Project Act (NAPA) (e.g., the goal of effectively treating or preventing AD and related dementias by 2025) and that many clinical trials of therapies for AD have failed in recent years, it becomes an extremely important and timely topic to study brain imaging genomics in AD. In particular, these efforts could accelerate progress in better understanding of the genetic, molecular, and neurobiological mechanisms of AD and have a subsequent translational impact on disease modeling and drug development. For example, recent ADNI studies have yielded prominent imaging genomics findings, such as BCHE and IL1RAP with amyloid QTs [210], [211], PARP1, CARD10, REST, FASTKD2, and ADORA2A with hippocampal morphometry [212]–[215], INPP5D with cerebral blood flow [216], and APOE with multimodal imaging QTs [48], [108], [174]. Some of these findings have contributed to genetically based drug targets leading to novel disease model systems [e.g., creation of the IL1RAP knockout mouse [217] and nomination of INPP5D as a modeling target (http://agora.ampadportal.org)].

Finally, for many novel statistical and machine learning methods reviewed here, the authors often used the ADNI data to demonstrate the power of the methods to detect interesting and novel imaging genomics signals. Some yielded confirmatory findings matching previous studies, showing the effectiveness of these methods. Some identified novel signals missed by the existing methods, showing improved detection power. Of note, the generalizability of findings from many of these new methods needs to be evaluated in additional independent data sets to demonstrate their broader impact on the future.

E. Related Work and Future Directions

In this article, we mostly reviewed the methods developed and employed for analyzing ADNI and ENIGMA data. Similar methods have been investigated in the study of other neurological and psychiatric disorders. For example, the pICA method was first proposed and then widely used in studies of psychiatric disorders [218]. Various SCCA and other multivariate models (see [219]–[223]) have been developed and employed in brain imaging genomics applications to study psychiatric disorders. Additional details

are available in [224], where the authors provided a recent review on neuroimaging genomics analyses and their translational potential to diagnosis and treatment in mental disorders.

Due to the open-science nature of the ADNI project and the large-scale global alliance formed by ENIGMA, a large number of researchers around the world have had the chance to analyze the ADNI and ENIGMA data, resulting in a major growth of literature in new statistical and machine learning methods for brain imaging genomics. Of note, the generalization of many of these new methods remains to be evaluated in other independent data sets, which will be an interesting and promising future direction. In particular, given the rapid growth and sheer number of these new developments, we observe no lack of innovation and expect to see the impact of these methods or their enhanced versions to permeate biomedical studies in brain imaging genomics.

Integrating imaging and omics data is also an active research topic in cancer studies, which is often referred to as radiogenomics [225]. In these studies, in addition to SNP data, multiomics data (e.g., transcriptomics, proteomics, metabolomics, and epigenomics) are often collected from the actual tumor tissues. Therefore, relating multiomics data to imaging data becomes a study focus. Note that the omics data are tissue-specific. Thus, the methods reviewed in this article are mostly focused on relating SNP data to imaging QTs, mainly due to the lack of the available brain tissue in these *in vivo* studies. However, with the increasing accumulation of brain samples in some landmark studies (e.g., AMP-AD [226]), more and more omics data will be available for the study of brain disorders. A promising future direction is to adapt many radiogenomics approaches developed for cancer research to the study of brain imaging genomics.

As we aim at understanding mechanisms and pathways, another challenge in brain imaging genomics is how to handle spurious correlations leading to erroneous conclusions. Thus, replication in independent cohorts will be an important step to complete in order to identify true signals. Some sources of spurious correlations, such as overfitting and biased sampling, have been studied as described earlier. However, systematic investigation of various confounding factors is an underexplored topic and warrants further investigation.

Deep learning models have been highly successful in addressing data-driven problems in biology and medicine [227]. However, they have not been widely used in brain imaging genomics, partly due to the limited sample size and high dimensionality of the existing imaging and genomics data sets. Some recent attempts have been made to develop effective deep learning models for outcome prediction via integrating brain imaging genomics data (see [201]). Given that deep learning has been producing impressive results in both medical image analysis [228], [229] and multiomics research [230], it is a promising future direction to develop deep learning methods for solving pressing problems in brain imaging genomics.

Given the unprecedented scale, complexity, and heterogeneity of the fast-growing big data in brain imaging genomics, we are facing a variety of other methodological challenges that suggest promising and exciting future research directions as follows.

1) Although multicohort integrative data analysis can offer increased statistical power, one major obstacle is that the available data modalities often vary across different studies. Thus, one promising direction is to develop novel machine learning or transfer learning methods that can effectively handle incomplete data modalities and facilitate multicohort data integration.
2) Most methods reviewed here analyzed genotyping data and were not designed for examining the whole genome/exome sequencing (WGS/WES) data. The rapid growth of WGS/WES data in brain imaging genomics calls for new statistical and machine learning methods that can properly handle their ultrahigh dimensionality and resolution as well as effectively identify both common and rare genetic variants related to imaging QTs.
3) There is also an urgent need for novel scalable computational strategies to support large-scale consortium-based collaborative efforts. For consortia with one single centralized data repository, cloud-based computational and informatics tools are needed to enable the users to directly analyze large-scale data in the cloud. For consortia with multiple local data repositories, distributed computation methods and frameworks could be established to handle the decentralized data sets.

The rapid growth of brain imaging genomics as an emerging data science field is greatly attributed to the public availability of valuable imaging and genomics data sets. For example, due to the open-science nature of the ADNI project, hundreds of publications using ADNI imaging genomics data have been produced in the past decade, yielding not only innovative machine learning methods but also novel biomedical discoveries. Similar to ADNI and ENIGMA, more and more landmark studies are producing big data, including multidimensional imaging and omics modalities, and make them available to the research community. Some example landmark studies are shown in the following:

1) ADNI [1];
2) ENIGMA [7], [8];
3) UK Biobank [3];
4) Human Connectome Project (HCP) [231];
5) Accelerating Medicines Partnership AD (AMP-AD) [226];
6) Mind Clinical Imaging Consortium (MCIC) [232];
7) Pediatric Imaging, Neurocognition, and Genetics study (PING) [233];

8) Parkinson's Progression Markers Initiative (PPMI) [234];
9) The Cancer Genome Atlas (TCGA) [235];
10) The Cancer Imaging Archive (TCIA) [236].

With this growing availability of brain imaging genomics data, we anticipate to observe many more advances in machine learning and their applications to brain imaging genomics, which will significantly contribute to biomedical discoveries in brain science and the study of brain disorders. ■

Acknowledgment

The authors would like to thank B. Thirion and three anonymous reviewers for providing highly valuable comments and feedback on this article.

REFERENCES

[1] M. W. Weiner *et al.*, "Recent publications from the Alzheimer's disease neuroimaging initiative: Reviewing progress toward improved ad clinical trials," *Alzheimer's Dementia*, vol. 13, no. 4, pp. e1–e85, 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/28342697

[2] P. Thompson *et al.*, "ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries," *PsyArXiv*, pp. 1–41, Jul. 2019. [Online]. Available: https://psyarxiv.com/qnsh7/. doi: 10.31234/osf.io/qnsh7.

[3] C. Sudlow *et al.*, "UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS Med.*, vol. 12, no. 3, 2015, Art. no. e1001779. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25826379

[4] L. Shen *et al.*, "Genetic analysis of quantitative phenotypes in AD and MCI: Imaging, cognition and biomarkers," *Brain Imag. Behav.*, vol. 8, no. 2, pp. 183–207, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24092460

[5] A. J. Saykin *et al.*, "Genetic studies of quantitative MCI and AD phenotypes in ADNI: Progress, opportunities, and plans," *Alzheimer's Dement*, vol. 11, no. 7, pp. 792–814, 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26194313

[6] D. P. Veitch *et al.*, "Understanding disease progression and improving Alzheimer's disease clinical trials: Recent highlights from the Alzheimer's disease neuroimaging initiative," *Alzheimer's Dementia*, vol. 15, no. 1, pp. 106–152, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30321505

[7] P. M. Thompson *et al.*, "The ENIGMA consortium: Large-scale collaborative analyses of neuroimaging and genetic data," *Brain Imag. Behav.*, vol. 8, no. 2, pp. 153–182, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24399358

[8] P. M. Thompson *et al.*, "ENIGMA and the individual: Predicting factors that affect the brain in 35 countries worldwide," *NeuroImage*, vol. 145, pp. 389–408, Jan. 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26658930

[9] C. E. Bearden and P. M. Thompson, "Emerging global initiatives in neurogenetics: The enhancing neuroimaging genetics through meta-analysis (ENIGMA) consortium," *Neuron*, vol. 94, no. 2, pp. 232–236, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28426957

[10] C. Bycroft *et al.*, "The UK Biobank resource with deep phenotyping and genomic data," *Nature*, vol. 562, no. 7726, pp. 203–209, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30305743

[11] F. Alfaro-Almagro *et al.*, "Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank," *NeuroImage*, vol. 166, pp. 400–424, Feb. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29079522

[12] L. T. Elliott *et al.*, "Genome-wide association studies of brain imaging phenotypes in UK Biobank," *Nature*, vol. 562, no. 7726, pp. 210–216, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30305740

[13] J. Liu and V. D. Calhoun, "A review of multivariate analyses in imaging genetics," *Frontiers Neuroinform.*, vol. 8, p. 29, Mar. 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24723883

[14] J. Yan, L. Du, X. Yao, and L. Shen, "Machine learning in brain imaging genomics," in *Machine Learning and Medical Imaging*. New York, NY, USA: Academic, 2016, ch. 14, pp. 411–434. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780128040768000141

[15] M. S. Mufford, D. J. Stein, S. Dalvie, N. A. Groenewold, P. M. Thompson, and N. Jahanshad, "Neuroimaging genomics in psychiatry—A translational approach," *Genome Med.*, vol. 9, no. 1, p. 102, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29179742

[16] J. Liu, J. Chen, N. Perrone-Bizzozero, and V. D. Calhoun, "A perspective of the cross-tissue interplay of genetics, epigenetics, and transcriptomics, and their relation to brain based phenotypes in schizophrenia," *Frontiers Genet.*, vol. 9, p. 343, Aug. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30190726

[17] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature Med.*, vol. 25, no. 1, pp. 44–56, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30617339

[18] P. M. Visscher, W. G. Hill, and N. R. Wray, "Heritability in the genomics era–concepts and misconceptions," *Nature Rev. Genet.*, vol. 9, no. 4, pp. 255–266, 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18319743

[19] P. M. Thompson *et al.*, "Genetic influences on brain structure," *Nature Neurosci.*, vol. 4, no. 12, pp. 1253–1258, 2001. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/11694885

[20] P. M. Thompson, T. Ge, D. C. Glahn, N. Jahanshad, and T. E. Nichols, "Genetics of the connectome," *NeuroImage*, vol. 80, pp. 475–488, Oct. 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23707675

[21] C. Brun *et al.*, "A tensor-based morphometry study of genetic influences on brain structure using a new fluid registration method," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*, vol. 11. Berlin, Germany: Springer, 2008, pp. 914–921. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18982692

[22] Y.-Y. Chou *et al.*, "Mapping genetic influences on ventricular structure in twins," *NeuroImage*, vol. 44, no. 4, pp. 1312–1323, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19041405

[23] K. K. Shen *et al.*, "Investigating brain connectivity heritability in a twin study using diffusion imaging data," *NeuroImage*, vol. 100, pp. 628–641, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24973604

[24] K.-K. Shen *et al.*, "Heritability and genetic correlation between the cerebral cortex and associated white matter connections," *Hum. Brain Mapping*, vol. 37, no. 6, pp. 2331–2347, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27006297

[25] Y. Fu *et al.*, "Genetic influences on resting-state functional networks: A *twin study*," *Hum. Brain Mapping*, vol. 36, no. 10, pp. 3959–3972, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26147340

[26] P. Kochunov *et al.*, "Homogenizing estimates of heritability among SOLAR-Eclipse, OpenMx, APACE, and FPHI software packages in neuroimaging data," *Frontiers Neuroinform.*, vol. 13, p. 16, Mar. 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30914942

[27] X. Chen *et al.*, "Accelerated estimation and permutation inference for ACE modeling," *Hum. Brain Mapping*, vol. 40, no. 12, pp. 3488–3507, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31037793

[28] H. Ganjgahi, A. M. Winkler, D. C. Glahn, J. Blangero, P. Kochunov, and T. E. Nichols, "Fast and powerful heritability inference for family-based neuroimaging studies," *NeuroImage*, vol. 115, pp. 256–268, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25812717

[29] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher, "GCTA: A tool for genome-wide complex trait analysis," *Amer. J. Hum. Genet.*, vol. 88, no. 1, pp. 76–82, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21167468

[30] T. Ge *et al.*, "Massively expedited genome-wide heritability analysis (MEGHA)," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 8, pp. 2479–2484, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25675487

[31] T. Ge *et al.*, "Multidimensional heritability analysis of neuroanatomical shape," *Nature Commun*, vol. 7, Nov. 2016, Art. no. 13291. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27845344

[32] T. Ge, C.-Y. Chen, B. M. Neale, M. R. Sabuncu, and J. W. Smoller, "Phenome-wide heritability analysis of the UK Biobank," *PLoS Genet.*, vol. 13, no. 4, 2017, Art. no. e1006711. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28388634

[33] B. K. Bulik-Sullivan *et al.*, "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies," *Nature Genet.*, vol. 47, no. 3, pp. 291–295, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25642630

[34] M. E. Rentería *et al.*, "Genetic architecture of subcortical brain regions: Common and region-specific genetic contributions," *Genes, Brain Behav.*, vol. 13, no. 8, pp. 821–830, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25199620

[35] P. H. Lee *et al.*, "Partitioning heritability analysis reveals a shared genetic basis of brain anatomy and schizophrenia," *Mol. Psychiatry*, vol. 21, no. 12, pp. 1680–1689, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27725656

[36] R. M. Brouwer *et al.*, "Genetic influences on individual differences in longitudinal changes in global and subcortical brain volumes: Results of the ENIGMA plasticity working group," *Hum. Brain Mapping*, vol. 38, no. 9, pp. 4444–4458, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28580697

[37] L. Shen *et al.*, "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*, vol. 53, no. 3, pp. 1051–1063, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20100581

[38] N. Jahanshad *et al.*, "Genome-wide scan of healthy human connectome discovers SPON1 gene variant influencing dementia severity," *Proc.*

Nat. Acad. Sci. USA, vol. 110, no. 12, pp. 4768–4773, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23471985

[39] M. A. Lindquist and A. Mejia, "Zen and the art of multiple comparisons," *Psychosomatic Med.*, vol. 77, no. 2, pp. 114–125, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25647751

[40] R. A. Armstrong, "When to use the Bonferroni correction," *Ophthalmic Physiol. Opt.*, vol. 34, no. 5, pp. 502–508, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24697967

[41] K. J. Worsley, J. E. Taylor, F. Tomaiuolo, and J. Lerch, "Unified univariate and multivariate random field theory," *NeuroImage*, vol. 23, no. Supplement 1, pp. S189–S195, 2004. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1053811904003982?via%3Dihub

[42] M. K. Chung, Y. Wang, S.-G. Huang, and I. Lyu, "Rapid acceleration of the permutation test via slow random walks in the permutation group," 2018, *arXiv:1812.06696*. [Online]. Available: https://arxiv.org/abs/1812.06696

[43] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Statist. Soc. B, Methodol.*, vol. 57, no. 1, pp. 289–300, Jan. 1995. [Online]. Available: http://www.jstor.org/stable/2346101

[44] W.-Y. Hua, T. E. Nichols, D. Ghosh, and AlzheimerŠs Disease Neuroimaging Initiative, "Multiple comparison procedures for neuroimaging genomewide association studies," *Biostatistics*, vol. 16, no. 1, pp. 17–30, 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24963012

[45] J. L. Stein *et al.*, "Genome-wide analysis reveals novel genes influencing temporal lobe structure with relevance to neurodegeneration in Alzheimer's disease," *NeuroImage*, vol. 51, no. 2, pp. 542–554, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/B6WNP-4YH56B3-2/2/0a240fe93fd981a1c214d859629daff8

[46] M. A. Scelsi *et al.*, "Genetic study of multimodal imaging Alzheimer's disease progression score implicates novel loci," *Brain*, vol. 141, no. 7, pp. 2167–2180, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29860282

[47] M. C. Donohue *et al.*, "Estimating long-term multivariate progression from short-term data," *Alzheimer's Dementia*, vol. 10, no. 5, pp. S400–S410, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24656849

[48] S. L. Risacher *et al.*, "*APOE* effect on Alzheimer's disease biomarkers in older adults with significant memory concern," *Alzheimer's Dementia*, vol. 11, no. 12, pp. 1417–1429, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25960448

[49] A. J. Ho *et al.*, "A commonly carried allele of the obesity-related FTO gene is associated with reduced brain volume in the healthy elderly," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 18, pp. 8404–8409, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/20404173

[50] M. Huang *et al.*, "FVGWAS: Fast voxelwise genome wide association analysis of large-scale imaging genetic data," *NeuroImage*, vol. 118, pp. 613–627, Sep. 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26025292

[51] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *J. Roy. Statist. Soc. B, Stat. Methodol.*, vol. 70, no. 5, pp. 849–911, 2008, doi: 10.1111/j.1467-9868.2008.00674.x.

[52] J. Kim, W. Pan, and Alzheimer's Disease Neuroimaging Initiative, "A cautionary note on using secondary phenotypes in neuroimaging genetic studies," *NeuroImage*, vol. 121, pp. 136–145, Nov. 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26220747

[53] E. D. Schifano, L. Li, D. C. Christiani, and X. Lin, "Genome-wide association analysis for multiple continuous secondary phenotypes," *Amer. J. Hum. Genet.*, vol. 92, no. 5, pp. 744–759, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23643383

[54] D. Y. Lin and D. Zeng, "Proper analysis of secondary phenotype data in case-control association studies," *Genet. Epidemiol.*, vol. 33, no. 3, pp. 256–265, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19051285

[55] W. Zhu *et al.*, "Genome-wide association analysis of secondary imaging phenotypes from the Alzheimer's disease neuroimaging initiative study," *NeuroImage*, vol. 146, pp. 983–1002, Feb. 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/27717770

[56] E. J. T. Tchetgen, "A general regression framework for a secondary outcome in case–control studies," *Biostatistics*, vol. 15, no. 1, pp. 117–128, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24152770

[57] J. L. Stein *et al.*, "Voxelwise genome-wide association study (vGWAS)," *NeuroImage*, vol. 53, no. 3, pp. 1160–1174, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20171287

[58] F. Dudbridge, "Power and predictive accuracy of polygenic risk scores," *PLoS Genet.*, vol. 9, no. 3, 2013, Art. no. e1003348. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23555274

[59] D. Dima and G. Breen, "Polygenic risk scores in imaging genetics: Usefulness and applications," *J. Psychopharmacol.*, vol. 29, no. 8, pp. 867–871, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25944849

[60] D. Chasioti, J. Yan, K. Nho, and A. J. Saykin, "Progress in polygenic composite scores in Alzheimer's and other complex diseases," *Trends Genet.*, vol. 35, no. 5, pp. 371–382, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30922659

[61] E. C. Mormino *et al.*, "Polygenic risk of Alzheimer disease is associated with early-and late-life processes," *Neurology*, vol. 87, no. 5, pp. 481–488, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27385740

[62] M. R. Sabuncu *et al.*, "The association between a polygenic Alzheimer score and cortical thickness in clinically normal subjects," *Cerebral Cortex*, vol. 22, no. 11, pp. 2653–2661, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22169231

[63] J.-C. Lambert *et al.*, "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease," *Nature Genet.*, vol. 45, no. 12, pp. 1452–1458, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24162737

[64] D. Harold *et al.*, "Genome-wide association study identifies variants at *CLU* and *PICALM* associated with Alzheimer's disease," *Nature Genet.*, vol. 41, no. 10, pp. 1088–1093, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19734902

[65] C. H. Tan *et al.*, "Polygenic hazard score, amyloid deposition and Alzheimer's neurodegeneration," *Brain*, vol. 142, no. 2, pp. 460–470, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30689776

[66] R. S. Desikan *et al.*, "Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score," *PLoS Med.*, vol. 14, no. 3, 2017, Art. no. e1002258. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28323831

[67] J. Euesden, C. M. Lewis, and P. F. O'Reilly, "PRSice: Polygenic risk score software," *Bioinformatics*, vol. 31, no. 9, pp. 1466–1468, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25550326

[68] A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, and M. J. Daly, "Clinical use of current polygenic risk scores may exacerbate health disparities," *Nature Genet.*, vol. 51, no. 4, pp. 584–591, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30926966

[69] L. Ibanez, F. H. G. Farias, U. Dube, K. A. Mihindukulasuriya, and O. Harari, "Polygenic risk scores in neurodegenerative diseases: A review," *Current Genet. Med. Rep.*, vol. 7, no. 1, pp. 22–29, 2019. [Online]. Available: https://doi.org/10.1007/s40142-019-0158-0

[70] L. G. Apostolova *et al.*, "Associations of the top 20 Alzheimer disease risk variants with brain amyloidosis," *JAMA Neurol.*, vol. 75, no. 3, pp. 328–341, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29340569

[71] D. P. Hibar *et al.*, "Voxelwise gene-wide association study (vGeneWAS): Multivariate gene-based association testing in 731 elderly subjects," *NeuroImage*, vol. 56, no. 4, pp. 1875–1891, 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21497199

[72] T. Ge, J. Feng, D. P. Hibar, P. M. Thompson, and T. E. Nichols, "Increasing power for voxel-wise genome-wide association studies: the random field theory, least square kernel machines and fast permutation procedures," *NeuroImage*, vol. 63, no. 2, pp. 858–873, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22800732

[73] Z. Xu, C. Wu, W. Pan, and Alzheimer's Disease Neuroimaging Initiative, "Imaging-wide association study: Integrating imaging endophenotypes in GWAS," *NeuroImage*, vol. 159, pp. 159–169, Oct. 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28736311

[74] Z.-H. Lu *et al.*, "Multiple SNP set analysis for genome-wide association studies through Bayesian latent variable selection," *Genet. Epidemiol.*, vol. 39, no. 8, pp. 664–677, 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26515609

[75] K. Wang and D. Abbott, "A principal components regression approach to multilocus genetic association studies," *Genet. Epidemiol.*, vol. 32, no. 2, pp. 108–118, 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/17849491

[76] D. Liu, X. Lin, and D. Ghosh, "Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models," *Biometrics*, vol. 63, no. 4, pp. 1079–1088, 2007. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18078480

[77] A. Gusev *et al.*, "Integrative approaches for large-scale transcriptome-wide association studies," *Nature Genet.*, vol. 48, no. 3, pp. 245–252, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26854917

[78] I. Y. Kwak and W. Pan, "Adaptive gene- and pathway-trait association testing with GWAS summary statistics," *Bioinformatics*, vol. 32, no. 8, pp. 1178–1184, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26656570

[79] G. R. Svishcheva, N. M. Belonogova, I. V. Zorkoltseva, A. V. Kirichenko, and T. I. Axenovich, "Gene-based association tests using GWAS summary statistics," *Bioinformatics*, vol. 35, no. 19, pp. 3701–3708, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30860568

[80] Y. Zhang, Z. Xu, X. Shen, W. Pan, and Alzheimer's Disease Neuroimaging Initiative, "Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data," *NeuroImage*, vol. 96, pp. 309–325, Aug. 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24704269

[81] Q. Yang, H. Wu, C. Y. Guo, and C. S. Fox, "Analyze multivariate phenotypes in genetic association studies by combining univariate association tests," *Genet. Epidemiol.*, vol. 34, no. 5, pp. 444–454, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/20583287

[82] H. Lan, J. P. Stoehr, S. T. Nadler, K. L. Schueler, B. S. Yandell, and A. D. Attie, "Dimension reduction for mapping mRNA abundance as quantitative traits," *Genetics*, vol. 164, no. 4, pp. 1607–1614, 2003. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/12930764

[83] M. A. R. Ferreira and S. M. Purcell, "A multivariate test of association," *Bioinformatics*, vol. 25, no. 1, pp. 132–133, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19019849

[84] X. Li, S. Basu, M. B. Miller, W. G. Iacono, and M. McGue, "A rapid generalized least squares model for a genome-wide quantitative trait association analysis in families," *Hum. Heredity*, vol. 71, no. 1, pp. 67–82, 2011. [Online].

Available: https://www.ncbi.nlm.nih.gov/pubmed/21474944
[85] A. Korte, B. J. Vilhjálmsson, V. Segura, A. Platt, Q. Long, and M. Nordborg, "A mixed-model approach for genome-wide association studies of correlated traits in structured populations," *Nature Genet.*, vol. 44, no. 9, pp. 1066–1071, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22902788
[86] K.-Y. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, vol. 73, no. 1, pp. 13–22, 1986. [Online]. Available: https://doi.org/10.1093/biomet/73.1.13
[87] J. Kim, Y. Zhang, W. Pan, and Alzheimer's Disease Neuroimaging Initiative, "Powerful and adaptive testing for multi-trait and multi-SNP associations with GWAS and sequencing data," *Genetics*, vol. 203, no. 2, pp. 715–731, 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/27075728
[88] J. Kim, W. Pan, and Alzheimer's Disease Neuroimaging Initiative, "Adaptive testing for multiple traits in a proportional odds model with applications to detect SNP-brain network associations," *Genet. Epidemiol.*, vol. 41, no. 3, pp. 259–277, 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/28191669
[89] C. Huang *et al.*, "FGWAS: Functional genome wide association analysis," *NeuroImage*, vol. 159, pp. 107–121, Oct. 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28735012
[90] X. Yao *et al.*, "Two-dimensional enrichment analysis for mining high-level imaging genetic associations," *Brain Inf.*, vol. 4, no. 1, pp. 27–37, 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/27747820
[91] V. K. Ramanan, L. Shen, J. H. Moore, and A. J. Saykin, "Pathway analysis of genomic data: Concepts, methods, and prospects for future development," *Trends Genet.*, vol. 28, no. 7, pp. 323–332, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22480918
[92] M. Holden, S. Deng, L. Wojnowski, and B. Kulle, "GSEA-SNP: Applying gene set enrichment analysis to SNP data from genome-wide association studies," *Bioinformatics*, vol. 24, no. 23, pp. 2784–2785, 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18854360
[93] V. K. Ramanan *et al.*, "Genome-wide pathway analysis of memory impairment in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort implicates gene candidates, canonical pathways, and networks," *Brain Imag. Behav.*, vol. 6, no. 4, pp. 634–648, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22865056
[94] D. Nam, J. Kim, S.-Y. Kim, and S. Kim, "GSA-SNP: A general approach for gene set analysis of polymorphisms," *Nucleic Acids Res.*, vol. 38, pp. W749–W754, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/20501604
[95] X. Yao *et al.*, "Mining regional imaging genetic associations via Voxel-wise enrichment analysis," in *Proc. IEEE Int. Conf. Biomed. Health Informat. (BHI)*, May 2019, pp. 1–4.
[96] Allen Institute for Brain Science. (2013). *Allen Human Brain Atlas: Technical White Paper: Microarray Data Normalization*. [Online]. Available: http://www.brain-map.org/ and http://help.brain-map.org/download/attachments/2818165/Normalization_WhitePaper.pdf
[97] The Gene Ontology Consortium, "The gene ontology project in 2008," *Nucleic Acids Res.*, vol. 36, pp. D440–D444, Jan. 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/17984083
[98] M. Kanehisa and S. Goto, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/10592173
[99] X. Yao *et al.*, "Tissue-specific network-based genome wide study of amygdala imaging phenotypes to identify functional interaction modules," *Bioinformatics*, vol. 33, no. 20, pp. 3250–3257, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28575147
[100] J. Yan, S. L. Risacher, L. Shen, and A. J. Saykin, "Network approaches to systems biology analysis of complex disease: Integrative methods for multi-omics data," *Briefings Bioinf.*, vol. 19, no. 6, pp. 1370–1381, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28679163
[101] A. Song *et al.*, "Network-based analysis of genetic variants associated with hippocampal volume in Alzheimer's disease: a study of ADNI cohorts," *BioData Mining*, vol. 9, no. 1, p. 3, 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26788126
[102] C. S. Greene *et al.*, "Understanding multicellular function and disease with human tissue-specific networks," *Nature Genet.*, vol. 47, no. 6, pp. 569–576, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25915600
[103] S. Patel, M. T. M. Park, Alzheimer's Disease Neuroimaging Initiative, M. M. Chakravarty, and J. Knight, "Gene prioritization for imaging genetics studies using gene ontology and a stratified false discovery rate approach," *Frontiers Neuroinform.*, vol. 10, p. 14, Apr. 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27092072
[104] M. Lorenzi *et al.*, "Susceptibility of brain atrophy to TRIB3 in Alzheimer's disease, evidence from functional prioritization in imaging genetics," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 12, pp. 3162–3167, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29511103
[105] M. J. Grothe *et al.*, "Molecular properties underlying regional vulnerability to Alzheimer's disease pathology," *Brain*, vol. 141, no. 9, pp. 2755–2771, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30016411
[106] M. Silver, G. Montana, and Alzheimer's Disease Neuroimaging Initiative, "Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps," *Stat. Appl. Genet. Mol. Biol.*, vol. 11, no. 1, pp. 1–43, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22499682
[107] M. Silver, E. Janousova, X. Hua, P. M. Thompson, G. Montana, and Alzheimer's Disease Neuroimaging Initiative, "Identification of gene pathways implicated in alzheimer's disease using longitudinal imaging phenotypes with sparse regression," *NeuroImage*, vol. 63, no. 3, pp. 1681–1694, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22982105
[108] J. Yan *et al.*, "Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm," *Bioinformatics*, vol. 30, no. 17, pp. i564–i571, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25161248
[109] A. L. Zieselman *et al.*, "Computational genetics analysis of grey matter density in Alzheimer's disease," *BioData Mining*, vol. 7, p. 17, Dec. 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25165488
[110] J. Gui *et al.*, "A simple and computationally efficient approach to multifactor dimensionality reduction analysis of gene-gene interactions for quantitative traits," *PLoS ONE*, vol. 8, no. 6, 2013, Art. no. e66545. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23805232
[111] A. K. Wong, C. Y. Park, C. S. Greene, L. A. Bongo, Y. Guan, and O. G. Troyanskaya, "IMP: A multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks," *Nucleic Acids Res.*, vol. 40, pp. W484–W490, Jun. 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22684505
[112] S. A. Meda *et al.*, "Genetic interactions associated with 12-month atrophy in hippocampus and entorhinal cortex in Alzheimer's Disease Neuroimaging Initiative," *Neurobiol. Aging*, vol. 34, no. 5, pp. 1518.e9–1518.e18, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23107432
[113] C. Herold, M. Steffens, F. F. Brockschmidt, M. P. Baur, and T. Becker, "INTERSNP: Genome-wide interaction analysis guided by a priori information," *Bioinformatics*, vol. 25, no. 24, pp. 3275–3281, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19837719
[114] D. P. Hibar *et al.*, "Genome-wide interaction analysis reveals replicated epistatic effects on brain structure," *Neurobiol. Aging*, vol. 36, pp. S151–S158, Jan. 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25264344
[115] M. Ueki and G. Tamiya, "Ultrahigh-dimensional variable selection method for whole-genome gene-gene interaction analysis," *BMC Bioinf.*, vol. 13, p. 72, May 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22554139
[116] J. Chen and Z. Chen, "Extended Bayesian information criteria for model selection with large model spaces," *Biometrika*, vol. 95, no. 3, pp. 759–771, 2008, doi: 10.1093/biomet/asn034.
[117] T. Ge *et al.*, "A kernel machine method for detecting effects of interaction between multidimensional variable sets: An imaging genetics application," *NeuroImage*, vol. 109, pp. 505–514, Apr. 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25600633
[118] C. Wang *et al.*, "A set-based mixed effect model for gene-environment interaction and its application to neuroimaging phenotypes," *Frontiers Neurosci.*, vol. 11, p. 191, Apr. 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/28428742
[119] J. Sun, Y. Zheng, and L. Hsu, "A unified mixed-effects model for rare-variant association in sequencing studies," *Genet. Epidemiol.*, vol. 37, no. 4, pp. 334–344, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23483651
[120] J. L. Stein *et al.*, "Identification of common variants associated with human hippocampal and intracranial volumes," *Nature Genet.*, vol. 44, no. 5, pp. 552–561, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22504417
[121] D. P. Hibar *et al.*, "Common genetic variants influence human subcortical brain structures," *Nature*, vol. 520, no. 7546, pp. 224–229, 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25607358
[122] S. M. Smith and T. E. Nichols, "Statistical challenges in 'big data' human neuroimaging," *Neuron*, vol. 97, no. 2, pp. 263–268, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29346749
[123] S. E. Medland, N. Jahanshad, B. M. Neale, and P. M. Thompson, "Whole-genome analyses of whole-brain data: Working within an expanded search space," *Nature Neurosci*, vol. 17, no. 6, pp. 791–800, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24866045
[124] P. M. Thompson, D. P. Hibar, J. L. Stein, G. Prasad, and N. Jahanshad, "Genetics of the Connectome and the ENIGMA Project," in *Micro-, Meso-and Macro-Connectomics of the Brain*. Cham, Switzerland: Springer, 2016, pp. 147–164. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28590671
[125] N. Jahanshad *et al.*, "Do candidate genes affect the brain's white matter microstructure? Large-scale evaluation of 6,165 diffusion MRI scans," *BioRxiv*, Jan. 2017, Art. no. 107987. [Online]. Available: http://biorxiv.org/content/early/2017/02/20/107987.abstract. doi: 10.1101/107987.
[126] J. P. Ioannidis, "Why most published research findings are false," *PLoS Med.*, vol. 2, no. 8, p. e124, 2005. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/16060722
[127] K. S. Button *et al.*, "Power failure: Why small sample size undermines the reliability of neuroscience," *Nature Rev. Neurosci.*, vol. 14, no. 5, pp. 365–376, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23571845
[128] K. S. Button, "Double-dipping revisited," *Nature Neurosci.*, vol. 22, no. 5, pp. 688–690, 2019.

[Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31011228

[129] K. L. Grasby *et al.*, "The genetic architecture of the human cerebral cortex," *BioRxiv*, Sep. 2018, Art. no. 399402. [Online]. Available: http://biorxiv.org/content/early/2018/09/03/399402.abstract. doi: 10.1101/399402.

[130] D. J. A. Smit *et al.*, "Genome-wide association analysis links multiple psychiatric liability genes to oscillatory brain activity," *Hum. Brain Mapping*, vol. 39, no. 11, pp. 4183–4195, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29947131

[131] I. E. Sønderby *et al.*, "Dose response of the 16p11.2 distal copy number variant on intracranial volume and basal ganglia," *Mol. Psychiatry*, Oct. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30283035. doi: 10.1038/s41380-018-0118-1.

[132] T. Jia *et al.*, "Epigenome-wide meta-analysis of blood DNA methylation and its association with subcortical volumes: Findings from the ENIGMA Epigenetics Working Group," *BioRxiv*, Nov. 2018, Art. no. 460444. [Online]. Available: http://biorxiv.org/content/early/2018/11/05/460444.abstract. doi: 10.1101/460444.

[133] H. H. Adams *et al.*, "Novel genetic loci underlying human intracranial volume identified through genome-wide association," *Nature Neurosci.*, vol. 19, no. 12, pp. 1569–1582, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27694991

[134] C. L. Satizabal *et al.*, "Genetic architecture of subcortical brain structures in over 40,000 individuals worldwide," *BioRxiv*, Aug. 2017, Art. no. 173831. [Online]. Available: http://biorxiv.org/content/early/2017/08/28/173831.abstract. doi: 10.1101/173831.

[135] F. Alfaro-Almagro *et al.*, "Image processing and quality control for the first 10,000 brain imaging datasets from UK biobank," *NeuroImage*, vol. 166, pp. 400–424, Feb. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29079522

[136] O. B. Smeland *et al.*, "Genetic overlap between schizophrenia and volumes of hippocampus, putamen, and intracranial volume indicates shared molecular genetic mechanisms," *Schizophrenia Bull.*, vol. 44, no. 4, pp. 854–864, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29136250

[137] D. P. Hibar *et al.*, "Significant concordance of genetic variation that increases both the risk for obsessive-compulsive disorder and the volumes of the nucleus accumbens and putamen," *Brit. J. Psychiatry*, vol. 213, no. 1, pp. 430–436, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29947313

[138] M. Mufford *et al.*, "Concordance of genetic variation that increases risk for Tourette Syndrome and that influences its underlying neurocircuitry," *Transl. Psychiatry*, vol. 9, no. 1, 2019, Art. no. 120. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30902966

[139] P. R. Jansen *et al.*, "GWAS of brain volume on 54,407 individuals and cross-trait analysis with intelligence identifies shared genomic loci and genes," *BioRxiv*, Jan. 2019, Art. no. 613489. [Online]. Available: http://biorxiv.org/content/early/2019/04/19/613489.abstract. doi: 10.1101/613489.

[140] D. Holland *et al.*, "Beyond SNP heritability: Polygenicity and discoverability of phenotypes estimated with a univariate Gaussian mixture model," *BioRxiv*, Jan. 2018, Art. no. 133132. [Online]. Available: http://biorxiv.org/content/early/2018/12/17/498550.abstract. doi: 10.1101/133132.

[141] M.-C. Chiang *et al.*, "Gene network effects on brain microstructure and intellectual performance identified in 472 twins," *J. Neurosci.*, vol. 32, no. 25, pp. 8732–8745, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22723713

[142] C.-H. Chen *et al.*, "Genetic topography of brain morphology," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 42, pp. 17089–17094, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24082094

[143] D. Sun *et al.*, "Large-scale mapping of cortical alterations in 22q11.2 deletion syndrome: Convergence with idiopathic psychosis and effects of deletion size," *Mol. Psychiatry*, Jun. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29895892. doi: 10.1038/s41380-018-0078-5.

[144] J. E. Villalón-Reina *et al.*, "Altered white matter microstructure in 22q11.2 deletion syndrome: A multisite diffusion tensor imaging study," *Mol. Psychiatry*, Jul. 2019. doi: 10.1038/s41380-019-0450-0.

[145] A. Subramanian *et al.*, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 43, pp. 15545–15550, 2005. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/16199517

[146] X. Hao *et al.*, "Identifying candidate genetic associations with MRI-derived AD-related ROI via tree-guided sparse learning," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, to be published. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29993890

[147] M. Wang, X. Hao, J. Huang, W. Shao, and D. Zhang, "Discovering network phenotype between genetic risk factors and disease status via diagnosis-aligned multi-modality regression method in Alzheimer's disease," *Bioinformatics*, vol. 35, no. 11, pp. 1948–1957, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30395195

[148] H. Wang *et al.*, "Identifying quantitative trait loci via group-sparse multitask regression and feature selection: An imaging genetics study of the ADNI cohort," *Bioinformatics*, vol. 28, no. 2, pp. 229–237, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22155867

[149] X. Wang *et al.*, "Longitudinal genotype–phenotype association study through temporal structure auto-learning predictive model," *J. Comput. Biol.*, vol. 25, no. 7, pp. 809–824, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30011249

[150] B. Fischl *et al.*, "Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain," *Neuron*, vol. 33, no. 3, pp. 341–355, 2002. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/11832223

[151] H. Wang *et al.*, "From phenotype to genotype: An association study of longitudinal phenotypic markers to Alzheimer's disease relevant SNPs," *Bioinformatics*, vol. 28, no. 18, pp. i619–i625, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22962490

[152] F. Nie, H. Huang, and C. Ding, "Low-rank matrix recovery via efficient schatten p-norm minimization," in *Proc. 26th AAAI Conf. Artif. Intell.*, 2012, pp. 655–661.

[153] T. Zhou, K.-H. Thung, M. Liu, and D. Shen, "Brain-wide genome-wide association study for Alzheimer's disease via joint projection learning and sparse regression model," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 1, pp. 165–175, Jan. 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29993426

[154] M. Vounou, T. E. Nichols, G. Montana, and Alzheimer's Disease Neuroimaging Initiative, "Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach," *NeuroImage*, vol. 53, no. 3, pp. 1147–1159, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20624472

[155] M. Vounou *et al.*, "Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in alzheimer's disease," *NeuroImage*, vol. 60, no. 1, pp. 700–716, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22209813

[156] X. Zhu, H.-I. Suk, H. Huang, and D. Shen, "Structured sparse low-rank regression model for brain-wide and genome-wide associations," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*, vol. 9900. Cham, Switzerland: Springer, 2016, pp. 344–352. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28530001

[157] X. Zhu, H.-I. Suk, H. Huang, and D. Shen, "Low-rank graph-regularized structured sparse regression for identifying genetic biomarkers," *IEEE Trans. Big Data*, vol. 3, no. 4, pp. 405–414, Dec. 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29725610

[158] R. Hu, X. Zhu, D. Cheng, W. He, Y. Yan, J. Song, and S. Zhang, "Graph self-representation method for unsupervised feature selection," *Neurocomputing*, vol. 220, pp. 130–137, Jan. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231216305458

[159] X. Zhu, W. Zhang, Y. Fan, and Alzheimer's Disease Neuroimaging Initiative, "A robust reduced rank graph regression method for neuroimaging genetic analysis," *Neuroinformatics*, vol. 16, nos. 3–4, pp. 351–361, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29907892

[160] X. Zhu, X. Li, S. Zhang, C. Ju, and X. Wu, "Robust joint graph sparse coding for unsupervised spectral feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 6, pp. 1263–1275, Jun. 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26955053

[161] K. Greenlaw, E. Szefer, J. Graham, M. Lesperance, F. S. Nathoo, and Alzheimer's Disease Neuroimaging Initiative, "A Bayesian group sparse multi-task regression model for imaging genetics," *Bioinformatics*, vol. 33, no. 16, pp. 2513–2522, 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/28419235

[162] T. Park and G. Casella, "The Bayesian lasso," *J. Amer. Stat. Assoc.*, vol. 103, no. 482, pp. 681–686, 2008, doi: 10.1198/016214508000000337.

[163] M. Kyung, J. Gill, M. Ghosh, and G. Casella, "Penalized regression, standard errors, and Bayesian lassos," *Bayesian Anal.*, vol. 5, no. 2, pp. 369–411, 2010. [Online]. Available: https://projecteuclid.org:443/euclid.ba/1340218343

[164] H. Zhu, Z. Khondker, Z. Lu, J. G. Ibrahim, and Alzheimer's Disease Neuroimaging Initiative, "Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers," *J. Amer. Stat. Assoc.*, vol. 109, no. 507, pp. 990–997, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25349462

[165] Z. H. Lu, Z. Khondker, J. G. Ibrahim, Y. Wang, H. Zhu, and Alzheimer's Disease Neuroimaging Initiative, "Disease Neuroimaging, "Bayesian longitudinal low-rank regression models for imaging genetic data from longitudinal studies," *NeuroImage*, vol. 149, pp. 305–322, Apr. 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/28143775

[166] X. Wang *et al.*, "Quantitative trait loci identification for brain endophenotypes via new additive model with random networks," *Bioinformatics*, vol. 34, no. 17, pp. i866–i874, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30423101

[167] W. F. Schmidt, M. A. Kraaijveld, and R. P. W. Duin, "Feedforward neural networks with random weights," in *Proc. 11th IAPR Int. Conf. Pattern Recognit. Conf. B, Pattern Recognit. Methodol. Syst.*, vol. 2, 1992, pp. 1–4.

[168] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19377034

[169] L. Du *et al.*, "A novel structure-aware sparse learning algorithm for brain imaging genetics," in *Medical Image Computing and Computer-Assisted*

Intervention—MICCAI, vol. 17. Cham, Switzerland: Springer, 2014, pp. 329–336. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25320816

[170] H. Zeng *et al.*, "Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures," *Cell*, vol. 149, no. 2, pp. 483–496, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22500809

[171] L. Du *et al.*, "Pattern discovery in brain imaging genetics via SCCA modeling with a generic non-convex penalty," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 14052. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29070790

[172] L. Du *et al.*, "A novel SCCA approach via truncated ℓ_1-norm and truncated group lasso for brain imaging genetics," *Bioinformatics*, vol. 34, no. 2, pp. 278–285, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28968815

[173] L. Grosenick, B. Klingenberg, K. Katovich, B. Knutson, and J. E. Taylor, "Interpretable whole-brain prediction analysis with GraphNet," *NeuroImage*, vol. 72, pp. 304–321, May 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23298747

[174] L. Du *et al.*, "Structured sparse canonical correlation analysis for brain imaging genetics: An improved GraphNet method," *Bioinformatics*, vol. 32, no. 10, pp. 1544–1551, 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26801960

[175] A. Gossmann, P. Zille, V. Calhoun, and Y.-P. Wang, "FDR-corrected sparse canonical correlation analysis with applications to imaging genomics," *IEEE Trans. Med. Imag.*, vol. 37, no. 8, pp. 1761–1774, Aug. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29993802

[176] T. D. Satterthwaite *et al.*, "Neuroimaging of the Philadelphia neurodevelopmental cohort," *NeuroImage*, vol. 86, pp. 544–553, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23921101

[177] L. Du *et al.*, "Fast multi-task SCCA learning with feature selection for multi-modal brain imaging genetics," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, Dec. 2018, pp. 356–361. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30881731

[178] X. Hao *et al.*, "Identification of associations between genotypes and longitudinal phenotypes via temporally-constrained group sparse canonical correlation analysis," *Bioinformatics*, vol. 33, no. 14, pp. i341–i349, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28881979

[179] L. Du *et al.*, "Identifying progressive imaging genetic patterns via multi-task sparse canonical correlation analysis: A longitudinal study of the ADNI cohort," *Bioinformatics*, vol. 35, no. 14, pp. i474–i483, 2019.

[180] É. L. Floch *et al.*, "Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares," *NeuroImage*, vol. 63, no. 1, pp. 11–24, Oct. 2012.

[181] J. Fang *et al.*, "Fast and accurate detection of complex imaging genetics associations based on greedy projected distance correlation," *IEEE Trans. Med. Imag.*, vol. 37, no. 4, pp. 860–870, Apr. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29990017

[182] J. Fan, Y. Feng, and L. Xia, "A projection based conditional dependence measure with applications to high-dimensional undirected graphical models," 2016, *arXiv:1501.01617*. [Online]. Available: https://arxiv.org/abs/1501.01617

[183] W. Hu, A. Zhang, B. Cai, V. Calhoun, and Y. P. Wang, "Distance canonical correlation analysis with application to an imaging-genetic study," *J. Med. Imag., Bellingham*, vol. 6, no. 2, 2019, Art. no. 026501. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31001569

[184] J. Liu, O. Demirci, and V. D. Calhoun, "A parallel independent component analysis approach to investigate genomic influence on brain function," *IEEE Signal Process. Lett.*, vol. 15, pp. 413–416, 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19834575

[185] V. D. Calhoun, J. Liu, and T. Adali, "A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data," *NeuroImage*, vol. 45, no. 1, p. S163–S172, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19059344

[186] S. A. Meda *et al.*, "A large scale multivariate parallel ICA method reveals novel imaging-genetic relationships for Alzheimer's disease in the ADNI cohort," *NeuroImage*, vol. 60, no. 3, pp. 1608–1621, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22245343

[187] J. Dukart, F. Sambataro, and A. Bertolino, "Accurate prediction of conversion to alzheimer's disease using imaging, genetic, and neuropsychological biomarkers," *J. Alzheimer's Disease*, vol. 49, no. 4, pp. 1143–1159, 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26599054

[188] R. Filipovych, B. Gaonkar, and C. Davatzikos, "A composite multivariate polygenic and neuroimaging score for prediction of conversion to Alzheimer's disease," in *Proc. Int. Workshop Pattern Recognit. Neuroimag.*, 2012, pp. 105–108. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24899230

[189] Y. Fan, D. Shen, R. C. Gur, R. E. Gur, and C. Davatzikos, "COMPARE: Classification of morphological patterns using adaptive regional elements," *IEEE Trans. Med. Imag.*, vol. 26, no. 1, pp. 93–105, Jan. 2007. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/17243588

[190] K. Kauppi *et al.*, "Combining polygenic hazard score with volumetric MRI and cognitive measures improves prediction of progression from mild cognitive impairment to Alzheimer's disease," *Frontiers Neurosci.*, vol. 12, p. 260, Apr. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29760643

[191] L. K. McEvoy *et al.*, "Alzheimer disease: Quantitative structural neuroimaging for detection and prediction of clinical and structural changes in mild cognitive impairment," *Radiology*, vol. 251, no. 1, pp. 195–205, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19201945

[192] L. K. McEvoy *et al.*, "Mild cognitive impairment: Baseline and longitudinal structural MR imaging measures improve predictive prognosis," *Radiology*, vol. 259, no. 3, pp. 834–843, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21471273

[193] H. Wang *et al.*, "Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning," *Bioinformatics*, vol. 28, no. 12, pp. i127–i136, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22689752

[194] Z. Zhang, H. Huang, D. Shen, and Alzheimer's Disease Neuroimaging Initiative, "Integrative analysis of multi-dimensional imaging genomics data for Alzheimer's disease prediction," *Frontiers Aging Neurosci.*, vol. 6, p. 260, Oct. 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25368574

[195] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *J. Mach. Learn. Res.*, vol. 9, pp. 2491–2521, Nov. 2008.

[196] F. Liu, H.-I. Suk, C.-Y. Wee, H. Chen, and D. Shen, "High-order graph matching based feature selection for Alzheimer's disease identification," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*, vol. 16. Berlin, Germany: Springer, 2013, pp. 311–318. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24579155

[197] H. Wang, F. Nie, H. Huang, and C. Ding, "Heterogeneous visual features fusion via sparse multimodal machine," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3097–3102.

[198] J. Peng, L. An, X. Zhu, Y. Jin, and D. Shen, "Structured sparse kernel learning for imaging genetics based Alzheimer's disease diagnosis," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*, vol. 9901. Cham, Switzerland: Springer, 2016, pp. 70–78. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28580458

[199] A. Singanamalli, H. Wang, A. Madabhushi, and Alzheimer's Disease Neuroimaging Initiative, "Cascaded multi-view canonical correlation (CaMCCo) for early diagnosis of Alzheimer's disease via fusion of clinical, imaging and omic features," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 8137. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28811553

[200] G. Lee *et al.*, "Supervised multi-view canonical correlation analysis (sMVCCA): Integrating histologic and proteomic features for predicting recurrent prostate cancer," *IEEE Trans. Med. Imag.*, vol. 34, no. 1, pp. 284–297, Jan. 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25203987

[201] T. Zhou, K.-H. Thung, X. Zhu, and D. Shen, "Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis," *Hum. Brain Mapping*, vol. 40, no. 3, pp. 1001–1016, 2019.

[202] K. Ning *et al.*, "Classifying Alzheimer's disease with brain imaging and genetic data using a neural network framework," *Neurobiol. Aging*, vol. 68, pp. 151–158, Aug. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29784544

[203] J. Yan, S. L. Risacher, K. Nho, A. J. Saykin, and L. Shen, "Identification of discriminative imaging proteomics associations in Alzheimer's disease via a novel sparse correlation model," in *Proc. Pacific Symp. Biocomput.*, vol. 22, 2017, pp. 94–104. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27896965

[204] X. Hao *et al.*, "Mining outcome-relevant brain imaging genetic associations via three-way sparse canonical correlation analysis in Alzheimer's disease," *Sci. Rep.*, vol. 7, Mar. 2017, Art. no. 44272. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/28291242

[205] L. Du *et al.*, "Diagnosis status guided brain imaging genetics via integrated regression and sparse canonical correlation analysis," in *Proc. IEEE Int. Sym. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 356–359.

[206] P. Zille, V. D. Calhoun, and Y.-P. Wang, "Enforcing co-expression within a brain-imaging genomics regression framework," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2561–2571, Dec. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28678703

[207] X. Bi, L. Yang, T. Li, B. Wang, H. Zhu, and H. Zhang, "Genome-wide mediation analysis of psychiatric and cognitive traits through imaging phenotypes," *Hum. Brain Mapping*, vol. 38, no. 8, pp. 4088–4097, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28544218

[208] N. K. Batmanghelich, A. Dalca, G. Quon, M. Sabuncu, and P. Golland, "Probabilistic modeling of imaging, genetics and diagnosis," *IEEE Trans. Med. Imag.*, vol. 35, no. 7, pp. 1765–1779, Jul. 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26886973

[209] S. G. Potkin *et al.*, "Genome-wide strategies for discovering genetic influences on cognition and cognitive disorders: Methodological considerations," *Cogn. Neuropsychiatry*, vol. 14, nos. 4–5, pp. 391–418, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19634037

[210] V. K. Ramanan *et al.*, "APOE and BCHE as modulators of cerebral amyloid deposition: A florbetapir PET genome-wide association study," *Mol. Psychiatry*, vol. 19, no. 3, pp. 351–357, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23419831

[211] V. K. Ramanan *et al.*, "GWAS of longitudinal amyloid accumulation on 18F-florbetapir PET in Alzheimer's disease implicates microglial activation gene IL1RAP," *Brain*, vol. 138,

pp. 3076–3088, Oct. 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26268530

[212] K. Nho *et al.*, "Whole-exome sequencing and imaging genetics identify functional variants for rate of change in hippocampal volume in mild cognitive impairment," *Mol. Psychiatry*, vol. 18, no. 7, pp. 781–787, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23608917

[213] K. Nho *et al.*, "Protective variant for hippocampal atrophy identified by whole exome sequencing," *Ann. Neurol.*, vol. 77, no. 3, pp. 547–552, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25559091

[214] V. K. Ramanan *et al.*, "*FASTKD2* is associated with memory and hippocampal structure in older adults," *Mol. Psychiatry*, vol. 20, no. 10, pp. 1197–1204, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25385369

[215] E. Horgusluoglu-Moloch *et al.*, "Targeted neurogenesis pathway-based gene analysis identifies ADORA2A associated with hippocampal volume in mild cognitive impairment and Alzheimer's disease," *Neurobiol. Aging*, vol. 60, pp. 92–103, Dec. 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28941407

[216] X. Yao *et al.*, "Targeted genetic analysis of cerebral blood flow imaging phenotypes implicates the INPP5D gene," *Neurobiol. Aging*, vol. 81, pp. 213–221, Sep. 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31319229

[217] The MODEL-AD Consortium. (2019). *JAX Stock #003284: IL-1R ACP KO Mouse Strain*. [Online]. Available: https://www.jax.org/strain/003284

[218] G. D. Pearlson, J. Liu, and V. D. Calhoun, "An introductory review of parallel independent component analysis (p-ICA) and a guide to applying p-ICA to genetic data and imaging phenotypes to identify disease-associated biological pathways and systems in common complex disorders," *Frontiers Genet.*, vol. 6, p. 276, Sep. 2015.

[219] D. Lin, V. D. Calhoun, and Y.-P. Wang, "Correspondence between fMRI and SNP data by group sparse canonical correlation analysis," *Med. Image Anal.*, vol. 18, no. 6, pp. 891–902, Aug. 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24247004

[220] J. Fang, D. Lin, S. C. Schulz, Z. Xu, V. D. Calhoun, and Y.-P. Wang, "Joint sparse canonical correlation analysis for detecting differential imaging genetics modules," *Bioinformatics*, vol. 32, no. 22, pp. 3480–3488, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27466625

[221] W. Hu *et al.*, "Adaptive sparse multiple canonical correlation analysis with application to imaging (epi)genomics study of schizophrenia," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 2, pp. 390–399, Feb. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29364120

[222] M. A. Alam, H.-Y. Lin, H.-W. Deng, V. D. Calhoun, and Y.-P. Wang, "A kernel machine method for detecting higher order interactions in multimodal datasets: Application to schizophrenia," *J. Neurosci. Methods*, vol. 309, pp. 161–174, Nov. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30184473

[223] M. Wang, T.-Z. Huang, J. Fang, V. D. Calhoun, and Y.-P. Wang, "Integration of imaging (epi)genomics data for the study of schizophrenia using group sparse joint nonnegative matrix factorization," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, to be published. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30762565

[224] J. Chen, J. Liu, and V. D. Calhoun, "Translational potential of neuroimaging genomic analyses to diagnosis and treatment in mental disorders," *Proc. IEEE*, vol. 107, no. 5, pp. 912–927, May 2019.

[225] L. Antonelli, M. R. Guarracino, L. Maddalena, and M. Sangiovanni, "Integrating imaging and omics data: A review," *Biomed. Signal Process. Control*, vol. 52, pp. 264–280, Jul. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1746809419301326

[226] R. J. Hodes and N. Buckholtz, "Accelerating medicines partnership: Alzheimer's disease (AMP-AD) knowledge portal aids Alzheimer's drug discovery through open data sharing," *Expert Opinion Therapeutic Targets*, vol. 20, no. 4, pp. 389–391, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26853544

[227] T. Ching *et al.*, "Opportunities and obstacles for deep learning in biology and medicine," *J. Roy. Soc Interface*, vol. 15, no. 141, 2018, Art. no. 20170387. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29618526

[228] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28778026

[229] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift Medizinische Phys.*, vol. 29, no. 2, pp. 102–127, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30553609

[230] D. Grapov, J. Fahrmann, K. Wanichthanarak, and S. Khoomrung, "Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine," *Omics, J. Integrat. Biol.*, vol. 22, no. 10, pp. 630–636, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30124358

[231] D. C. Van Essen and M. F. Glasser, "The human connectome project: Progress and prospects," *Cerebrum, Dana Forum Brain Sci.*, vol. 2016, pp. 10–16, Sep./Oct. 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5198757/

[232] R. L. Gollub *et al.*, "The MCIC collection: A shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia," *Neuroinformatics*, vol. 11, no. 3, pp. 367–388, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23760817

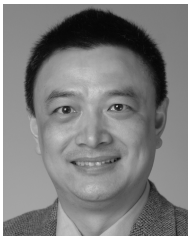
[233] T. L. Jernigan *et al.*, "The pediatric imaging, neurocognition, and genetics (PING) data repository," *NeuroImage*, vol. 124, pp. 1149–1154, Jan. 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25937488

[234] K. Marek *et al.*, "The Parkinson's progression markers initiative (PPMI)—Establishing a PD biomarker cohort," *Ann. Clin. Transl. Neurol.*, vol. 5, no. 12, pp. 1460–1477, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30564614

[235] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge," *Contemp. Oncol., Pozn*, vol. 19, no. 1A, pp. A68–A77, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25691825

[236] K. Clark *et al.*, "The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imag.*, vol. 26, no. 6, pp. 1045–1057, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23884657

ABOUT THE AUTHORS



Li Shen (Senior Member, IEEE) received the bachelor's degree from Xi'an Jiaotong University, Xi'an, China, in 1993, the master's degree from Shanghai Jiao Tong University, Shanghai, China, in 1996, and the Ph.D. degree from the Dartmouth College, Hanover, NH, USA, in 2004, all in computer science.

He is currently a Professor of informatics with the Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. His current research interests include brain imaging genetics, machine learning, medical image computing, biomedical informatics, and big data science in biomedicine.

Dr. Shen is an elected Fellow of the American Institute for Medical and Biological Engineering (AIMBE).

Paul M. Thompson received the bachelor's degree in Greek and Latin languages and mathematics and the master's degree in mathematics from Oxford University, Oxford, U.K., in 1991 and 1993, respectively, and the Ph.D. degree in neuroscience from the University of California at Los Angeles, Los Angeles, CA, USA, in 1998.

He is currently a Professor of neurology with the Imaging Genetics Center (IGC), University of Southern California, Los Angeles, CA, USA. He specializes in the field of human-brain imaging, with a research interest in mathematical and computational algorithm development for human-brain mapping. He has contributed to more than 900 publications. He currently leads the Enhancing Neuro Imaging Genetics through Meta Analysis (ENIGMA) project, a global data collection and sharing effort designed to understand how brain structure changes during the trajectory of brain atrophy, mental illness and Alzheimer's, and the underlying genetic landscape.