Computational and Structural Biotechnology Journal 19 (2021) 5800–5810

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Multivariate Analysis and Modelling of multiple Brain endOphenotypes: Let's MAMBO!

Natalia Vilor-Tejedor [a,b,c,d,1,*], Diego Garrido-Martín [b,1,*], Blanca Rodriguez-Fernandez [a], Sander Lamballais [c], Roderic Guigó [b,d], Juan Domingo Gispert [a,d,e,f]

[a] Barcelonaβeta Brain Research Center (BBRC), Pasqual Maragall Foundation, Barcelona, Spain
[b] Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Barcelona, Spain
[c] Department of Clinical Genetics, Erasmus Medical Center, Rotterdam, Netherlands
[d] Universitat Pompeu Fabra, Barcelona, Spain
[e] IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain
[f] Centro de Investigación Biomédica en Red Bioingeniería, Biomateriales y Nanomedicina, Madrid, Spain

ARTICLE INFO

Article history:
Received 22 April 2021
Received in revised form 8 October 2021
Accepted 12 October 2021
Available online 13 October 2021

Keywords:
Imaging genetics
Multiple phenotypes
Multivariate modelling
Neuroimaging
Genetics
Image-derived phenotype

ABSTRACT

Imaging genetic studies aim to test how genetic information influences brain structure and function by combining neuroimaging-based brain features and genetic data from the same individual.

Most studies focus on individual correlation and association tests between genetic variants and a single measurement of the brain. Despite the great success of univariate approaches, given the capacity of neuroimaging methods to provide a multiplicity of cerebral phenotypes, the development and application of multivariate methods become crucial.

In this article, we review novel methods and strategies focused on the analysis of multiple phenotypes and genetic data. We also discuss relevant aspects of multi-trait modelling in the context of neuroimaging data.

Contents

* Corresponding authors at: BarcelonaBeta Brain Research Center (BBRC), Pasqual Maragall Foundation. C. Wellington 30, 08005 Barcelona, Spain (N. Vilor-Tejedor). Centre for Genomic Regulation (CRG). The Barcelona Institute for Science and Technology, Barcelona, Spain. C. Doctor Aiguader, 88, 08003, Barcelona, Spain (D. Garrido-Martín).
E-mail addresses: nvilor@barcelonabeta.org (N. Vilor-Tejedor), diego.garrido@crg.eu (D. Garrido-Martín).
[1] Equal contribution.

https://doi.org/10.1016/j.csbj.2021.10.019

1. Introduction

Genetics plays an important role and provides valuable insights into the etiology of common brain diseases. Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with complex diseases [1]. However, when focusing on brain diseases, mapping genetic associations becomes particularly challenging. Among the possible reasons, their genetic and phenotypic complexity provides inconclusive results, as well as the difficulty assessing clinical diagnosis due to the existence of heterogeneity between individuals with the same diagnosis. This raises the question of using neuroimaging brain-based features as intermediate phenotypes, allowing a biologically more plausible manner to assess these complex associations [2].

Neuroimaging endophenotypes are quantitative measurements of the brain used to disentangle genetic susceptibility for complex neurological diseases and psychiatric [3]. Neuroimaging techniques can provide a wide variety of cerebral phenotypes associated with the brain's morphology (through structural magnetic resonance imaging (MRI), for instance), presence of lesions (e.g. microbleeds), measurements related to brain function (through functional MRI (fMRI) or fluorodeoxyglucose (FDG)-positron emission tomography (PET) imaging, for example) or the burden of molecular pathology (e.g amyloid and tau PET). Several studies have shown that brain endophenotypes generally seem to be shaped by genetic influences, which suggests that studying genetics and the brain jointly can refine our understanding of the etiology of neurological disease [4–6].

Neuroimaging data is commonly structured as a matrix, where each endophenotype becomes a single variable (column) for which a number of observations are available (rows). This structure is often maintained regardless of the source (e.g. disease studied), modality (e.g. sMRI, fMRI, DTI), units (e.g. voxels, volumes), number or combination of endophenotypes. Genetic data presents an analogous structure, where each variable is the genotype of a genetic variant (e.g. SNP, structural variant), observed in the same individuals from which neuroimaging data has been collected. In both cases, the variables studied are not independent, due to interconnected brain networks and linkage disequilibrium, respectively.

Imaging genetic (IG) studies started analyzing how candidate genes and genetic variants affected brain endophenotypes, using correlation or linear regression models. Later, IG studies focused on the genome-wide effects through GWAS to leverage high-throughput genetic variant data [7,8]. In both strategies, neuroimaging traits were modeled as outcome variables. However, despite the increasing availability of brain endophenotype data, most GWAS in the field test associations between genetic variants and a single brain phenotype at a time, that is, in a univariate fashion [9,10]. Assessing every brain outcome independently ignores the genetic correlation structure (i.e., pleiotropy) among multiple phenotypes, and implies a strict penalization for the significance threshold due to multiple hypothesis testing [11]. Altogether, this translates into reduced statistical power. In addition, since neuroimaging techniques already provide a quantitative measurement of the phenotype of interest across different brain subregions, multivariate approaches are particularly well suited to conduct IG studies.

Nowadays, some studies in the IG field have begun to explore the interplay between genetic variants and multiple phenotypes using multivariate approaches, aiming to identify associations at a genome-wide, whole-brain scale [12–14]. In these studies, brain features can be used either as outcomes or as independent variables in combination with genetic data, depending on the methodological approach. In contrast to univariate analysis, multivariate methods are able to leverage phenotype correlations due to shared genetic or environmental factors, while reducing the multiple testing burden, resulting in higher statistical power to identify significant associations. In addition, multi-phenotype analysis may reveal pleiotropic variants, providing new insights into the complex genetic architecture of brain endophenotypes and, eventually, helping to clarify their underlying biology.

Our aim with this review is to describe the methods currently employed in IG studies for the analysis of multiple traits, as well as those that are not widely used yet, but may be of great interest for researchers in the field in the near future. We also discuss relevant aspects of multi-trait modelling in the context of neuroimaging data.

2. Methods for the analysis of multiple phenotypes

Given the high dimensionality of neuroimaging data, methods for multi-phenotype analysis become a natural choice in IG studies. These approaches can be classified into three main groups: methods based on linear combinations of multiple phenotypes, multivariate regression models and Bayesian strategies [Fig. 1, Table 1]. Overall, they can be applied regardless of the source of the neuroimaging traits.

2.1. 2.1 Linear combinations of multiple phenotypes

These methods explore the multivariate structure of the data, aiming to select representative components and generate new responses based on combinations of multiple phenotypes.

Principal Component Analysis (PCA) is the traditional method used to reduce the number of phenotypes from large datasets [15,16]. PCA uses an orthogonal transformation to derive new phenotypes (principal components) that are linear combinations (lower-dimensional representations) of the original phenotypes. The first principal component is the direction in the orthogonal space along which the variance of the multiple phenotypes is maximized, and so on. Thus, PCA can be employed to decompose multiple phenotypes into components that can be used in subsequent univariate regression analyses as dependent variables [17,18].

The Independent Component Analysis (ICA) [19,20] and its various related algorithms are an extension of PCA. In ICA, data variables are assumed to be linear or nonlinear mixtures of unknown latent variables. The latent variables are assumed non-gaussian and mutually independent and they are called the independent components of the observed data. ICA is suitable to extract independent components from multiple phenotype measurements and it is extensively used in IG studies [21–23]. [22] performed a large scale multivariate ICA identifying significant imaging-genetic relationships for Alzheimer's Disease (AD). [21] used ICA to extract independent component values from connectivity brain measurements to assess their association with genetic variants related with the risk of schizophrenia. More recently, [23] used ICA to extract features from structural brain data, and searched for genetic variants associated with these brain-related features. In addition, multi-modal order methods based on ICA have been proposed to deal with the selection of the optimal number of inde-

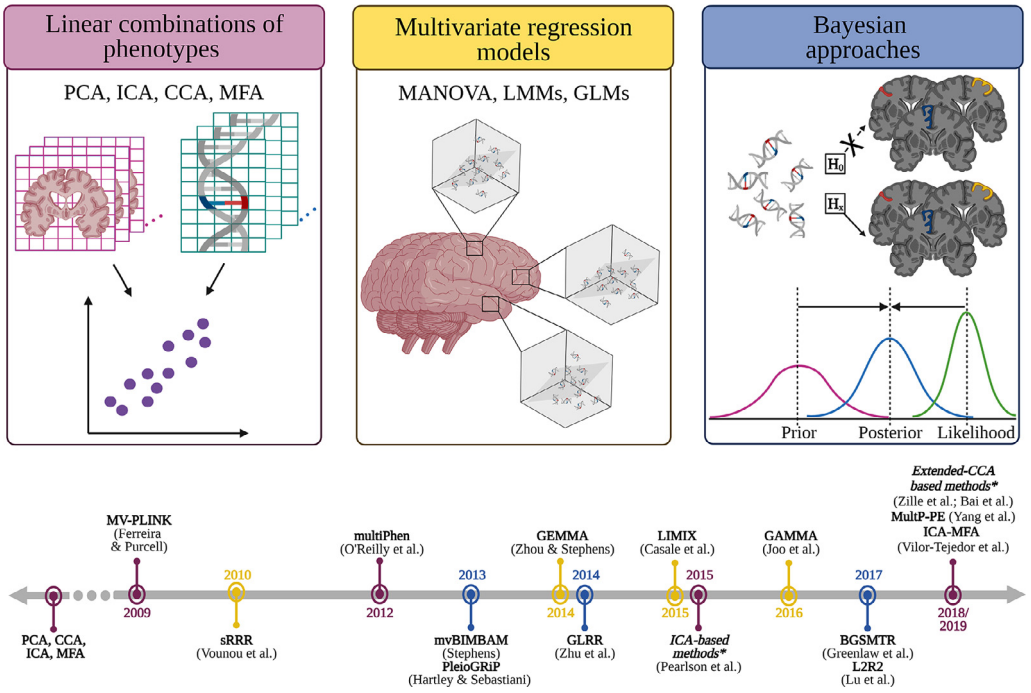

Fig. 1. Framework of analytical strategies for multiple phenotype assessment. **Year of these publications corresponds to a literature review describing this family of methods. Legend: BGSMTR: Bayesian group sparse multi-task regression model; CCA: Canonical Correlation Analysis; GLM: General Linear Model; ICA: Independent Component Analysis; ICA-MFA: Independent Multiple Factor Association Analysis for Multiblock Data; L2R2: Bayesian longitudinal low-rank regression; LMM: Linear Mixed Model; MANOVA: Multivariate Analysis of Variance; MFA; Multiple Factor Analysis; PCA: Principal Component Analysis; sRRR: sparse Reduced Rank Regression. This schematic representation was created with Biorender (©BioRender - biorender.com).*

pendent components in neuroimaging studies. For instance, [24] used healthy unrelated subjects from the WU-Minn Human Connectome Project [25] to show the improvement in the selection of components when the method follows a free order model. [26] also proposed a method to capture information at multiple model orders, showing an improvement in classifying brain patterns of schizophrenia individuals *vs* healthy controls.

The use of ICA-related methods to integrate multimodal data is becoming increasingly widespread. The main objective remains to simultaneously maximize independence and correlation linkage across variables by combining ICA with other methods such as Canonical Correlation Analysis (CCA). For instance, [27] proposed a multi-site canonical correlation analysis fusioned with ICA (MCCAR + jICA), which takes advantage of cross-information among multiple neuroimaging modalities to search for common patterns in brain endophenotypes related to brain disorders. Using this method, similar brain networks were identified in two independent cohorts, suggesting working memory deficits in schizophrenia individuals. More studies applying extended ICA-based methods in IG studies are described in [28,29].

Regarding CCA, this method allows to derive the relationship between two sets of variables measured in the same individuals (e.g. *X* and *Y*). It finds linear combinations of *X* and *Y* which have maximum correlation with each other. A comprehensive summary of CCA and its variants is presented in [30]. Specifically in the IG context, several CCA-based methods have been applied, providing higher precision, compared to previous approaches, on assessing the correlation patterns of multiple phenotypes [31–33]. Most of these methods extract variables co-occurring across imaging phenotypes and modalities together with genomic information. For instance, [32] and [33] combined sparse regression and CCA to extract significant sets of fMRI units (voxels) and genetic variants, whereas [31] combined linear regression analysis with CCA to extract features from multiple imaging phenotypes obtained in schizophrenic patients through fMRI and epigenetics. MV-PLINK is also a well known method based on CCA to simultaneously analyze multiple phenotypes. In MV-PLINK, CCA extracts the linear combination of phenotypes that has maximum correlation with a given genetic variant [34]. Although no studies applying MV-PLINK to brain phenotypes were found, interesting related works rely on this approach [35]. Alternatively, MultiPhen [36], uses ordinal regression instead of CCA to test for the association between a linear combination of phenotypes and each genetic variant. In MultiPhen, genetic data is modeled as an ordinal outcome and the multiple phenotypes are treated as the explanatory variables. An extension of MultiPhen was also developed [37], proposing a multiple phenotype procedure based on cross-validation of the prediction error (MultP-PE), which controlled well type I error rates and performed consistently better than MultiPhen.

Multiple Factor Analysis (MFA) is another well known strategy to integrate multiple phenotypes or datasets [38]. By applying MFA, we can evaluate how much the combination of the variables/datasets contributes to the inertia extracted by a unique component. Even with the potential advantages of MFA, few IG studies to date have used this methodological approach [39,40]. In [39], a sparse MFA framework was proposed. This framework first applied LASSO (least absolute shrinkage and selection operator) procedure to extract relevant genetic and brain features associated with hyperactivity and inattention domains. Then, these features were used as input for the MFA. Moreover, [40], proposed an extension of MFA based on ICA (ICA-MFA). This method implemented an independent component value decomposition of multiple datasets and multiple variables allowing non-normal and non-linear distributions. The application of this method showed the improvement in performance compared to MFA.

All these methods are characterized by being computationally faster than most of the approaches described below. However, several limitations must be considered. First, the loadings of the

Table 1
Summary of methods and software for multivariate analysis of multiple phenotypes/datasets.

Method	Type	Reference	Description	Pros	Cons	Implementation	R-function{R-package}/ Script
PCA-based methods	*Linear Combination Method*	*Pearlson et al. (2015)*	Principal Component Analysis	Computationally faster than other methods.	High dependency of input data. Lack of generalizability. Biologically meaningless.	Matlab	*https://trendscenter.org/software/fit/*
MV-PLINK	*Linear Combination Method*	*Ferreira and Purcell (2008)*	Based on Canonical Correlation Analysis			C++, Command-line	*https://genepi.qimr.edu.au/staff/manuelF/multivariate/main.html*
MELODIC	*Linear Combination Method*	*Beckman and Smith (2004)*	Probabilistic Independent Component Analysis for fMRI			GUI, Command-line	*https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/MELODIC#Research_Overview*
ICA-MFA	*Linear Combination Method*	*Vilor-Tejedor et al. (2019)*	Independent Multiple Factor Association Analysis for Multiblock Data			R/GitHub	*{ICA-MFA}*
MFA	*Linear Combination Method*	*Lê, Josse and Husson (2008)*	Multiple Factor Analysis			R/CRAN	*MFA{FactoMineR}*
MultiPhen	*Linear Combination Method*	*O'Reilly et al. (2012)*	Joint Model of Multiple Phenotypes			R/CRAN	*{MultiPhen}*
MultP-PE	*Linear Combination Method*	*Yang et al. (2019)*	Multiple Phenotypes based on cross-validation Prediction Error			R	*https://pages.mtu.edu/~shuzhang/software/MultP-PE.R*
GEMMA	*Regression-based Approach*	*Zhou and Stephens (2014)*	Genome-wide Efficient Mixed Model Association	Flexibility. Efficiency.	Strong distributional assumptions. Permutations. Requirement of multivariate normality and homoscedasticity. Inflated type I error rates.	C/C++	*http://stephenslab.uchicago.edu/software.html*
GAMMA	*Regression-based Approach*	*Jo et al. (2016)*	Generalized Analysis of Molecular variance for Mixed-model Analysis			R/C	*{vegan}; MEMMA{sommer}*
mvLMMs	*Regression-based Approach*	*Furlotte and Eskin (2015)*	Multivariate Linear Mixed Models			Python	*http://genetics.cs.ucla.edu/mvLMM*
mtSet	*Regression-based Approach*	*Casale et al. (2015)*	multi-trait Set test			Python	*https://github.com/limix/limix*
SNPtest	*Bayesian Modeling*	*Marchini et al. (2007)*	Multipoint method for genome-wide association studies via imputation of genotypes	Interpretability. Adaptability.	Need of specifying a prior probability distribution for the alternative hypothesis. High computational cost.	Java, C++, Command-line	*https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest*
PleioGRiP	*Bayesian Modeling*	*Hartley and Sebastiani (2013)*	Genetic risk prediction with pleiotropy			Java	*http://hdl.handle.net/2144/4367*
mvBIMBAM	*Bayesian Modeling*	*Stephens (2013)*	Bayesian approach for genetic association analysis of multiple related phenotypes			C/C++	*{mvBIMBAM}*

(*continued on next page*)

Table 1 (*continued*)

Method	Type	Reference	Description	Pros	Cons	Implementation	R-function{R-package}/ Script
G2RR	*Bayesian Modeling*	*Zhu et al. (2014)*	Bayesian Generalized Low-Rank Regression Models			R/GitHub	*{L2R2}*
BGSMTR	*Bayesian Modeling*	*Greenlaw et al. (2017)*	Bayesian group sparse multi-task regression model for imaging genetics			R/CRAN	*{bgsmtr}*

extracted components depend on the input data, and so slight differences between datasets may yield vastly different loadings. As such, there are constraints in generalizing findings related to the components from one dataset to another [41]. Second, these components maximize the explained (co)variance or correlation, but they may not properly represent the underlying biology. Thus, interpreting the findings can be unintuitive.

2.2. Multivariate regression models

Although multivariate regression models have been less applied to IG studies than the previous group of methods, they are probably the most versatile approaches to identify associations between multiple phenotypes and genetic variants.

These methods generally regress a matrix of m phenotypes measured in n individuals onto a set of covariates, including the genotype at the variant(s) of interest. They include variations from the general linear model, such as multivariate analysis of variance (MANOVA) or multivariate linear mixed models (mvLMMs), as well as multivariate generalized linear models (mvGLMs) and generalized estimating equations (GEEs). Regularized and nonparametric alternatives are also available.

Analogously to its univariate counterpart, MANOVA (equivalently, MANCOVA, when additional covariates need to be accounted for) decomposes the total covariance of the response variables, comparing the covariance explained by the genotype with the residual covariance. MANOVA can be performed using different test statistics (e.g. Wilks' lambda, Pillai's trace, etc.). Interestingly, MANOVA (Wilks' lambda) is equivalent to CCA when the latter is applied using a single variant at time, as in MV-PLINK [42]. This illustrates the close relationship between approaches leveraging multiple phenotypes, even if grouped here in different categories. MANOVA can be considered the direct multivariate equivalent of the univariate GWAS approach, and it has been recently compared with the latter in the context of neuroimaging studies [43].

A nonparametric alternative to MANOVA is PERMANOVA (permutational multivariate analysis of variance, also referred to as multivariate regression analysis of distance matrices) [44,45]. PERMANOVA computes the similarity (or distance) between pairs of individual samples and performs a covariance decomposition analogous to MANOVA. In contrast to MANOVA, it does not assume errors to follow multivariate normal (MVN) distributions. However, it relies on permutations for significance assessment. This results in a large p-value lower bound (the smallest p-value that can be achieved is approximately $1/(P + 1)$, where P is the number of permutations) and in running times increasing dramatically with the number of individuals and permutations, hindering its usage in large-scale GWAS studies.

In the context of GWAS studies, population stratification (large-scale, systematic differences in ancestry) and relatedness (either family structure or cryptic relatedness among individuals with no known family relationships) are well known to result in inflated type I errors [46]. Although methods such as MANOVA can account for population stratification by including the top genotype principal components as covariates in the model, they are sensitive to relatedness. In this regard, mvLMMs have become very popular in recent years, given that they can naturally incorporate relatedness as a random effect in the model. However, fitting mvLMMs (often involving restricted maximum likelihood estimation, REML) is computationally intensive and may be slow in large datasets with a large number of individuals and phenotypes, despite continuous implementation enhancements [47–49]. A widely used mvLMM implementation is available in GEMMA [48]. In GEMMA, the running time per genetic variant increases quadratically with the number of individuals, which makes computationally tractable the analysis of datasets of moderate size (n < 50,000, approximately). Analogously, only a modest number of phenotypes can be studied (e.g., $m \sim$ 2–10). As a general rule, available mvLMMs implementations scale better with respect to n than with respect to m. An exception is GAMMA [50], which scales linearly with the number of phenotypes, and cubically with the number of individuals. GAMMA combines the LMM and PERMANOVA frameworks. Hence, it requires permutations for significance assessment and presents the limitations pointed out above.

To date, most mvLMM implementations include a single variance component (i.e. relatedness). However, recent advancements have allowed to efficiently incorporate additional variance components. This is the case of mSet [51], which models the sum of the contribution from the variants in the genetic region to be tested ("set component") as a random factor, in addition to relatedness. mSet scales similarly to GEMMA with respect to n and m, although it does not assess significance of individual genetic variants, but rather of genomic regions of custom size.

MANOVA and mvLMMs assume that the model errors are MVN-distributed. Although they are relatively robust to violations of this assumption, both methods can lead to inflated type I errors in the presence of strong outliers, especially in the case of low minor allele freqüències (MAFs) [36]. As a result, normalization procedures such as rank-based inverse quantile transformation of the phenotypes (a methodology that replaces the sample quantiles by quantiles from the standard normal distribution) are commonly applied. Nevertheless, it is unclear whether these transformations always result in higher power and controlled type I errors compared to modeling the untransformed data [52].

Hence, MANOVA and mvLMMs are sometimes replaced by the more flexible framework provided by multivariate generalized linear models (mvGLMs), which allow error distributions from the exponential family, and therefore discrete outcomes. For instance, for multivariate count data (e.g. microbleed counts across brain subregions), a common choice is the multinomial-logit GLM. Pervasive overdispersion (variance of the phenotypes higher than expected) and complex correlation structures have led to the development of alternative models (boosted by the advent of

RNA-seq technologies in the field of transcriptomics), such as Dirichlet-multinomial or negative-multinomial GLMs [53,54]. Analogously to mvLMMs, mvGLMs can be extended to incorporate random factors. However, mvGL(M)Ms require one to correctly specify the model, including the correlation structure among the multiple phenotypes, which is often difficult, and fitting them is computationally demanding. As a result, they are not routinely used in multivariate GWAS studies. Some of these difficulties may be overcome by generalized estimation equations (GEE), a semi-parametric approach which just assumes marginal phenotype distributions to follow univariate generalized linear error models [55]. GEE require defining the link function and a "working" covariance matrix, being relatively robust to misspecification of the latter [56]. They have been often employed in GWAS, including in the field of IG, although they may not offer a good control of *type I error* in some scenarios [57].

The multivariate linear regression framework can be also employed to model the joint effect of many genetic variants in multiple phenotypes. However, in this scenario, the number of independent variables of interest could be larger than the number of observations. In addition, due to linkage disequilibrium (LD), multicollinearity may arise. To address these issues, regularized models, such as sparse reduced rank regression (sRRR), have been proposed [58]. For instance, sRRR was employed to identify genetic variants associated with voxel-wise longitudinal changes in the brain of AD patients [59]. Nevertheless, while these approaches allow to select a set of relevant genetic variants, they often lack an appropriate multivariate hypothesis testing setting to assess the significance of each variant's effect on the multiple phenotypes.

Multivariate regression strategies provide highly flexible and relatively efficient tools to study the genetic architecture of brain imaging endophenotypes. However, they are not exempt from limitations. Overall, they tend to make strong distributional assumptions on the error distributions, which often do not hold. Although nonparametric alternatives are available, they generally rely on permutations for significance assessment, and are thus impractical for large GWAS. In addition to multivariate normality, multivariate homoscedasticity (i.e. homogeneity of covariance matrices) is often required, and the violation of this assumption leads to markedly inflated *type I errors*, particularly in the case of low minor allele frequencies. Albeit the impact of heteroscedasticity may be reduced in mvGLMs or mixed models, as stated above, these require either defining *a priori* the variance structure (which can be difficult in large and complex biological datasets), or inferring it from the data (which is slow).

2.3. Bayesian approaches

Bayesian inference approaches have been implemented in the context of multivariate analyses in several genetic software, such as the widely used SNPtest [60], the multivariate version of BIMBAM [61] and, PleioGRiP [62]. All these software calculate a ratio between the probability distribution of the data under two models: the null hypothesis of no association and the alternative hypothesis. This ratio, called Bayes Factor (BF), represents an alternative to the classical use of p-value in frequentist approaches [63]. BFs are specially beneficial in the context of multivariate settings since they can be easily combined into a weighted average measure across different genetic variants, as implemented in SNPtest, or across several models at a given genetic variant, as implemented in mvBIMBAM. Further, mvBIMBAM introduced the idea of dividing the potentially associated genetic variants in two groups: those directly and those indirectly associated with one or more phenotypes. This division increases the interpretability of the multivariate analysis, although it is made according to statistical relationships rather than based on biology [61]. Lately, PleioGRiP introduced the use of Naive Bayesian classifiers to find pleiotropic associations between genetic variants and multiple phenotypes [64].

In the context of IG studies, several Bayesian methods have been also applied to the multivariate analysis of MRI data. [65] proposed a Bayesian generalized low rank regression model (GLRR), able to handle the high-dimensionality of the data which characterizes IG analysis. GLRR also integrates an efficient Markov chain Monte Carlo algorithm for posterior computation. The application of GLRR has led to the description of novel associations between relevant genetic variants associated with AD [66] and brain endophenotypes from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [67]. [68] developed a Bayesian longitudinal low-rank regression (L2R2) model as an adaptation of GLRR for longitudinal designs. L2R2 is characterized by approximating the low-rank matrix including penalized splines that account for overall time effect, and a sparse factor analysis model coupled with random effects to cover the within-subject spatio-temporal correlations of longitudinal imaging endophenotypes. Finally, [69] proposed a novel Bayesian group sparse multi-task regression model (BGSMTR) focused on integrating multiple structural brain endophenotypes with genetic information. This approach adapted a sparse regularization method primarily defined to group genetic variants by genes or LD blocs [70] into a method able to group genetic variants by brain imaging endophenotypes. BGSMTR was applied on the multivariate association analysis of relevant genetic variants for AD. However, the computational complexity of this method makes its application to IG studies infeasible without a previous selection of relevant genetic variants..

Bayesian methods provide interesting advantages compared to frequentist approaches. For instance, in terms of interpretability of the results. They also overcome limitations that are intrinsic to the use of p-values in assessing true associations between genetic variants and multiple phenotypes [71,72]. Moreover, they provide a setting that can be adapted to a wide range of scenarios (e.g. hierarchical models) and are able to deal with missing data problems. However, some considerations associated with Bayesian models include the need of specifying a prior probability distribution for the alternative hypothesis. The prior distributions will determine the genetic models that will be finally tested by the BF. Thus, they should be cautiously selected according to the scientific question. Additionally, Bayesian analysis often comes with a very high computational cost, especially for models involving a large number of variables [73].

3. Discussion and future perspective

The analysis of neuroimaging and genetic data is highly relevant in elucidating the etiology of neurological diseases. However, despite the intrinsically multivariate nature of neuroimage-derived phenotypes, they are still typically analyzed using univariate strategies. In addition to the increased statistical power offered by multivariate methods, the joint analysis of multiple imaging phenotypes and genetic data can help us to gain new insights on phenomena, such as pleiotropic effects, that are difficult to capture through standard univariate analyses.

In this review, we discussed methods commonly applied -or that can be potentially applied- in IG studies for the analysis of múltiple quantitative phenotypes, as well as examples of their application [Table 2]. Although we have grouped them into three broad categories (i.e. *linear combinations of multiple phenotypes and/or datasets*, *multivariate regression models*, and *Bayesian approaches*), we want to emphasize that the different methods are often related, and in some cases could even be considered complementary views of the same approach. Note also that we focused

Table 2
Summary of studies applying multiple phenotype strategies in Imaging genetic studies.

Author	Project/Consortium	Neuroimaging modality	Omics data	Statistical Modelling
Meda et al., (2012)	Alzheimer's Disease Neuroimaging Initiative	sMRI	SNPs	ICA
Vounou et al., (2012)	Alzheimer's Disease Neuroimaging Initiative	sMRI	SNPs	sRRR
Wang et al., (2012)	Alzheimer's Disease Neuroimaging Initiative	sMRI, FDG-PET	SNPs	sparse multimodal MTL
Mounce et al., (2014)	Mind Clinical Imaging Consortium	DTI	SNPs	ICA
Nazeri et al., (2014)	Alzheimer's Disease Neuroimaging Initiative	TBM, sMRI	Proteomics	PICA
Zhang et al., (2014)	Alzheimer's Disease Neuroimaging Initiative	sMRI	SNPs	GEE
Zhu et al., (2014)	Alzheimer's Disease Neuroimaging Initiative	sMRI	SNPs	L2R2
Du et al., (2015)	Alzheimer's Disease Neuroimaging Initiative	fMRI	SNPs	CCA
Peng et al., (2016)	Alzheimer's Disease Neuroimaging Initiative	MRI, PET	SNPs	Structured Sparse Kernel Learning
Greenlaw et al., (2017)	Alzheimer's Disease Neuroimaging Initiative	sMRI	SNPs	BGSMTR
Liu et al., (2017)	Alzheimer's Disease Neuroimaging Initiative	sMRI	SNPs, Gene expression	SCCA
Lu et al., (2017)	Alzheimer's Disease Neuroimaging Initiative	sMRI	SNPs	L2R2
Singanamalli et al., (2017)	Alzheimer's Disease Neuroimaging Initiative	sMRI, FDG-PET	Proteomics	CaMCCo
Yan et al., (2017)	Alzheimer's Disease Neuroimaging Initiative	sMRI	Proteomics	SCCA
Kircher et al., (2018)	FOR2107 Consortium	sMRI, fMRI	Multi-Omics	Multiple methods applied: data reduction; mixed models; machine learning
Ning et al., (2018)	Alzheimer's Disease Neuroimaging Initiative	sMRI	SNPs	Neural network
Vilor-Tejedor et al., (2018)	BRain dEvelopment and Air polluTion ultrafine particles in scHool childrEn	sMRI	SNPs	LASSO + MFA
Wachinger et al., (2018)	Alzheimer's Disease Neuroimaging Initiative	sMRI	SNPs	Mixed model
Zille et al., (2018)	Philadelphia Neurodevelopmental Cohort	fMRI	SNPs	CCA
Bai et al., (2019)	Mind Clinical Imaging Consortium	fMRI	Epigenetics	CCA
Vilor-Tejedor et al., (2019)	Rotterdam study	sMRI	SNPs	ICA-MFA
Soheili-Nezhad et al., (2020)	Alzheimer's Disease Neuroimaging Initiative	TBM, sMRI	SNPs	ICA
Vilor-Tejedor et al., (2020)	Rotterdam Study	sMRI	SNPs	Mixed model
Wu et al., (2020)	UK Biobank	sMRI	SNPs	Adaptive multi-trait association test

Legend: BGSMTR: Bayesian group sparse multi-task regression model; CaMCCo: Cascaded Multi-view Canonical Correlation; CCA: Canonical Correlation Analysis; CT: Computed tomography; DTI: Diffusion Tensor Imaging; FDG: fluorodeoxyglucose; fMRI: functional Magnetic Resonance Imaging; GEE: Generalized Estimation Equations; ICA: Independent Component Analysis; ICA-MFA: Independent Multiple Factor Association Analysis for Multiblock Data; ISGC: International Stroke Genetics Consortium; L2R2: Bayesian longitudinal low-rank regression; LASSO: Least Absolute Shrinkage and Selection Operator; MANOVA: Multivariate Analysis of Variance; MFA; Multiple Factor Analysis; MTL: Multi-Task Learning; PET: Positron Emission Tomography; PICA: Parallel Independent Component Analysis; SCCA: Sparse Canonical Correlation Analysis; sMRI: structural Magnetic Resonance Imaging; SNPs: Single Nucleotide Polymorphisms; sRRR: sparse Reduced Rank Regression;TBM: Tensor-Based Morphometry.

on quantitative endophenotypes, and that most methods described here may not be suited for the analysis of categorical data. Finally, further considerations that should be taken into account in current and future IG studies are discussed below.

3.1. Statistical power

Although the analysis of multiple phenotypes reduces the number of statistical comparisons, large samples are still required to increase statistical power. Statistical power can be boosted through the use of larger sample sizes, particularly by joining global consortia collaborations [74,75] or by using publicly available data from biobanks and other resources [76]. Nevertheless, it can also be improved by reducing the measurement error in both fields. Moreover, advances in DNA sequencing and reductions in costs associated with its acquisition, will ease the way for obtaining whole genome sequences. For neuroimaging modalities, such as PET data, the problem is to acquire sufficiently large datasets, given the use of ionizing radiation (and associated costs). In addition, improvements on the quality control of neuroimaging data, specifically on increasingly large datasets, could also help to obtain precise MRI-derived quantifications [77]. Finally, we can increase statistical power by using smarter data analysis, and developing more powerful statistical techniques. For instance, the combination of different multivariate approaches has also been proposed as an alternative for increasing significantly association findings [61].

3.2. Multivariate vs meta-analysis (summary-statistic-based) approaches

The multivariate methods described above require individual-level data on phenotypes and genotypes, and estimate the phenotype variances and covariances directly from the observed measurements. Although not discussed here, several methods have been proposed to leverage univariate summary-statistics (estimated effect sizes and standard errors, *Z*-statistics or association p-values) from published GWAS in a *meta*-analysis-like approach. Some examples include S_{Hom}/S_{Het} [78], metaCCA [79], MTAG [80] or MTAR [81]. They often rely on the assumption that, under the null hypothesis, summary Z-statistics from univariate tests performed on individual phenotypes have an asymptotic multivariate normal distribution with correlation equal to the phenotype correlation matrix. In recent years, this strategy has become very popular, being employed for the analysis of multiple phenotypes in a wide variety of scenarios, including IG studies [82–84]. The main advantage of these approaches is their ability to carry out fast GWAS analyses across many phenotypes and with very large sample sizes, without requiring complex and time-consuming access to individual-level data. Nevertheless, they are limited by the nature of the univariate analyses upon which they are based (sample size, power, assumptions, data transformations, etc.). In addition, estimating the phenotype correlation matrix from summary statistics is not trivial, and can be affected, among others, by LD or the phenotype heritability, which if not accounted properly may lead to biased estimates [81,85].

3.3. Longitudinal designs

Much of the existing work in the field, as we previously described, analyzes neuroimaging data cross-sectionally, which does not provide information about changes in the structure and functionality of the brain over time. A better understanding of the longitudinal trajectories of brain endophenotypes would improve our knowledge of the underlying biological characterisation of complex neurological diseases [86–88]. For example, appli-

cation of mixed-effect models to longitudinal brain imaging data has helped to identify a large number of significant genetic-multiphenotype associations [89,90]. Overall, this is an important topic that needs to be further investigated.

3.4. Integration of different types of Omics and Imaging modalities

A critical challenge in IG studies is to model an even greater complexity of genetic effects on the brain. Most neuroimaging studies have examined only two sources of biological data at a time (e.g., genetic variants and structural MRI). However, combining multiple omics would provide a better understanding of the underlying biological mechanisms of neurological diseases [91,92]. In addition, integrating multiple sources of neuroimaging -beyond than just two modalities- with omics data becomes relevant to amplify the synergistic value of IG studies (e.g., genomics, transcriptomics, proteomics, metabolomics, morphological MRI, DTI, or functional MRI) [93–95]. In that regard, the implementation of integrated data platforms such as the Brain Imaging Data Structure (BIDS) [96] and the BIDS genetics extension [97] facilitates data search and analytical procedures by aggregating genetic and neuroimaging features across studies.

Some studies have demonstrated the integration of gene expression data and genetic markers to facilitate the detection of markers that are both associated with brain endophenotypes and highly expressed in the brain [98]. Moreover, the integration of genetic neuroimaging methods with epigenetics, known as imaging epigenetics, promises to provide deeper insights into the causative pathways through which genes and environment interact during life and impact human brain development [99,100]. Other studies have also started to analyse proteomics and neuroimaging-based features as potential biomarkers of the basis for computing essential cell functions to identify the best proteomic model for the diagnosis, monitoring, and prediction of complex neurological disorders [101,102].

3.5. Imaging gene-environment interaction models

Research focused on multivariate modeling of gene-environment interactions has recently emerged, revealing significant interaction effects between candidate genetic variants and multiple environmental factors [103–106]. These methods may represent the starting point of designs focused on the integration of multivariate imaging gene-environment interactions open up new sources of analysis by means of which to gain an understanding of the conditional mechanisms through which genes, environment, and brain features interact to predict brain diseases and neurological conditions [107,108]. Such designs represent an opportunity, not only to integrate different omics and imaging features, but also to incorporate target environmental exposures relevant to the structure and function of the brain.

3.6. Complex prediction models

Machine learning and deep learning strategies represent a powerful alternative, allowing for heterogeneous data integration, and in turn, improving disease classification and prediction in IG studies [109]. However, this family of methods are also computationally expensive and interpretation is not straightforward. Moreover, there is not much literature reported or evidence of their performance yet. Some examples of application of these methods include predicting AD diagnosis and disease conversion [110–113]. However, more work is still required to use genetics and neuroimaging information to predict complex neurological outcomes and treatment response, as has been accomplished in other areas of psychiatric research [4,114].

CRediT authorship contribution statement

Natalia Vilor-Tejedor: Conceived the idea presented in the manuscript. Writing - original draft, Writing - review & editing, Visualization. **Diego Garrido-Martin:** Conceived the idea presented in the manuscript. Writing - original draft, Writing - review & editing, Visualization. **Blanca Rodríguez-Fernández:** Writing - original draft, Writing - review & editing and Visualization. **Sander Lamballais Yu:** Writing - review & editing. **Roderic Guigó:** Writing - review & editing. **Juan D Gispert:** Writing - review & editing and Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors apologize if relevant publications were not cited.

Author contributions

All the authors made a significant contribution to the preparation of the manuscript. NV-T and DG-M conceived the presented idea and took the lead in writing the manuscript. BR-F contributed to writing and editing the manuscript. SL, RG and JDG contributed to reviewing, and the discussion of the manuscript.

Funding

At the time of writing this review, NV-T is funded by a postdoctoral grant, Juan de la Cierva Programme (FJC2018-038085-I), Ministerio de Ciencia, Innovación y Universidades – Spanish State Research Agency. Her research is also supported by the "la Caixa" Foundation (LCF/PR/GN17/10300004), the Health Department of the Catalan Government (Health Research and Innovation Strategic Plan (PERIS) 2016–2020 grant #SLT002/16/00201), and the Alzheimer Nederland Project (WE.15–2019-09). DG-M is funded by grant number CZF2019-002436 from the Chan Zuckerberg Initiative. JDG holds a 'Ramón y Cajal' fellowship (RYC-2013–13054). All CRG authors acknowledge the support of the Spanish Ministry of Science, Innovation, and Universities to the EMBL partnership, the Centro de Excelencia Severo Ochoa, and the CERCA Programme / Generalitat de Catalunya.

References

[1] Price AL, Spencer CCA, Donnelly P. Progress and promise in understanding the genetic basis of common diseases. Proc R Soc B Biol Sci 2015;282 (1821):20151684. https://doi.org/10.1098/rspb.2015.1684.
[2] Gottesman II, Gould TD. The endophenotype concept in psychiatry: Etymology and strategic intentions. Am J Psychiatry 2003;160(4):636–45. https://doi.org/10.1176/appi.ajp.160.4.636.
[3] Glahn DC, Thompson PM, Blangero J. Neuroimaging endophenotypes: Strategies for finding genes influencing brain structure and function. Hum Brain Mapp 2007;28(6):488–501. https://doi.org/10.1002/(ISSN)1097-019310.1002/hbm.v28:610.1002/hbm.20401.
[4] Bogdan R, Salmeron BJ, Carey CE, Agrawal A, Calhoun VD, Garavan H, et al. Imaging Genetics and Genomics in Psychiatry: A Critical Review of Progress and Potential. Biol Psychiatry 2017;82(3):165–75. https://doi.org/10.1016/j.biopsych.2016.12.030.
[5] Elliott LT, Sharp K, Alfaro-Almagro F, Shi S, Miller KL, Douaud G, et al. Genome-wide association studies of brain imaging phenotypes in UK Biobank. Nature 2018;562(7726):210–6. https://doi.org/10.1038/s41586-018-0571-7.
[6] Matoba N, Love MI, Stein JL. Evaluating brain structure traits as endophenotypes using polygenicity and discoverability. Hum Brain Mapp 2020. https://doi.org/10.1002/hbm.25257.

[7] Shen Li, Kim S, Risacher SL, Nho K, Swaminathan S, West JD, 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 2010;53(3):1051–63. https://doi.org/10.1016/j.neuroimage.2010.01.042.
[8] Stein JL, Hua X, Lee S, Ho AJ, Leow AD, Toga AW, et al. Voxelwise genome-wide association study (vGWAS). Neuroimage 2010;53(3):1160–74. https://doi.org/10.1016/j.neuroimage.2010.02.032.
[9] van der Meer D, Frei O, Kaufmann T, Shadrin AA, Devor A, Smeland OB, et al. Understanding the genetic determinants of the brain with MOSTest. Nat Commun 2020;11(1). https://doi.org/10.1038/s41467-020-17368-1.
[10] Maleki F, Ovens K, Hogan DJ, Kusalik AJ. Gene Set Analysis: Challenges, Opportunities, and Future Research. Front Genet 2020;11:654. https://doi.org/10.3389/fgene.2020.00654.
[11] Wang Y, Liu A, Mills JL, Boehnke M, Wilson AF, Bailey-Wilson JE, et al. Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models. Genet Epidemiol 2015;39(4):259–75. https://doi.org/10.1002/gepi.2015.39.issue-410.1002/gepi.21895.
[12] Nathoo FS, Kong L, Zhu H. A review of statistical methods in imaging genetics. Can J Stat 2019;47(1):108–31. https://doi.org/10.1002/cjs.v47.110.1002/cjs.11487.
[13] Shen Li, Thompson PM. Brain Imaging Genomics: Integrated Analysis and Machine Learning. Proc IEEE 2020;108(1):125–62. https://doi.org/10.1109/PROC.510.1109/JPROC.2019.2947272.
[14] Vilor-Tejedor N, Alemany S, Cáceres A, Bustamante M, Pujol J, Sunyer J, et al. Strategies for integrated analysis in imaging genetics studies. Neurosci Biobehav Rev 2018;93:57–70. https://doi.org/10.1016/j.neubiorev.2018.06.013.
[15] Ma S, Dai Y. Principal component analysis based methods in bioinformatics studies. Brief Bioinform 2011;12(6):714–22. https://doi.org/10.1093/bib/bbq090.
[16] Jolliffe IT. Principal Component Analysis, Second Edition. Springer Ser Stat 2002;98:487. 10.1007/b98835.
[17] Zhang W, Gao X, Shi X, Zhu Bo, Wang Z, Gao H, et al. PCA-based multiple-trait GWAS analysis: A powerful model for exploring pleiotropy. Animals 2018;8(12):239. https://doi.org/10.3390/ani8120239.
[18] Geeraert BL, Chamberland M, Lebel RM, Lebel C, Yap P-T. Multimodal principal component analysis to identify major features of white matter structure and links to reading. PLoS ONE 2020;15(8):e0233244. https://doi.org/10.1371/journal.pone.0233244.
[19] Du K-L, Swamy MNS. Independent Component Analysis. Neural Networks Stat. Learn., London: Springer London; 2014, p. 419–50. 10.1007/978-1-4471-5571-3_14.
[20] Herault J, Jutten C, Ans B. Detection de grandeurs primitives dans un message composite par une architecture de calcul neuromimetique en apprentissage non supervise. GRETSI, Groupe d'Etudes du Traitement du Signal et des Images; 1985.
[21] Mounce J, Luo Li, Caprihan A, Liu J, Perrone-Bizzozero NI, Calhoun VD. Association of GRM3 polymorphism with white matter integrity in schizophrenia. Schizophr Res 2014;155(1-3):8–14. https://doi.org/10.1016/j.schres.2014.03.003.
[22] Meda SA, Narayanan B, Liu J, Perrone-Bizzozero NI, Stevens MC, Calhoun VD, et al. A large scale multivariate parallel ICA method reveals novel imaging-genetic relationships for Alzheimer's disease in the ADNI cohort. Neuroimage 2012;60(3):1608–21. https://doi.org/10.1016/j.neuroimage.2011.12.076.
[23] Soheili-Nezhad S, Jahanshad N, Guelfi S, Khosrowabadi R, Saykin AJ, Thompson PM, et al. Imaging genomics discovery of a new risk variant for Alzheimer's disease in the postsynaptic <scp>SHARPIN</scp> gene. Hum Brain Mapp 2020;41:3737–48. https://doi.org/10.1002/hbm.25083.
[24] Hu G, Waters AB, Aslan S, Frederick B, Cong F, Nickerson LD. Snowball ICA: A Model Order Free Independent Component Analysis Strategy for Functional Magnetic Resonance Imaging Data. Front Neurosci 2020:1005. https://doi.org/10.3389/FNINS.2020.569657.
[25] Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K. The WU-Minn Human Connectome Project: an overview. Neuroimage 2013;80:62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041.
[26] X M, A I, Z F, P K, A B, J F, et al. Multi-model Order ICA: A Data-driven Method for Evaluating Brain Functional Network Connectivity Within and Between Multiple Spatial Scales. Brain Connect 2021. 10.1089/BRAIN.2021.0079.
[27] Qi Shile, Calhoun Vince D, van Erp Theo GM, Bustillo Juan, Damaraju Eswar, Turner Jessica A, et al. Multimodal Fusion with Reference: Searching for Joint Neuromarkers of Working Memory Deficits in Schizophrenia. IEEE Trans Med Imaging 2018;37(1):93–105. https://doi.org/10.1109/TMI.2017.2725306.
[28] Pearlson GD, Liu J, Calhoun VD. 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. Front Genet 2015;6. https://doi.org/10.3389/fgene.2015.00276.
[29] Duan K, Calhoun VD, Liu J, Silva RF. ANy-way Independent Component Analysis. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, vol. 2020- July, Institute of Electrical and Electronics Engineers Inc.; 2020, p. 1770–4. 10.1109/EMBC44109.2020.9175277.
[30] Zhuang Xiaowei, Yang Zhengshi, Cordes Dietmar. A technical review of canonical correlation analysis for neuroscience applications. Hum Brain Mapp 2020;41(13):3807–33. https://doi.org/10.1002/hbm.v41.1310.1002/hbm.25090.
[31] Bai Yuntong, Pascal Zille, Hu Wenxing, Calhoun Vince D, Wang Yu-Ping. Biomarker Identification through Integrating fMRI and Epigenetics. IEEE Trans Biomed Eng 2020;67(4):1186–96. https://doi.org/10.1109/TBME.1010.1109/TBME.2019.2932895.
[32] Zille Pascal, Calhoun Vince D, Wang Yu-ping. Enforcing co-expression within a brain-imaging genomics regression framework. IEEE Trans Med Imaging 2018;37(12):2561–71. https://doi.org/10.1109/TMI.2017.2721301.
[33] Du L, Yan J, Kim S, Risacher SL, Huang H, Inlow M, et al. GN-SCCA: Graphnet based sparse canonical correlation analysis for brain imaging genetics. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9250, Springer Verlag; 2015, p. 275–84. 10.1007/978-3-319-23344-4_27.
[34] Ferreira MAR, Purcell SM. A multivariate test of association. Bioinformatics 2009;25:132–3. https://doi.org/10.1093/bioinformatics/btn563.
[35] Nath Artika P, Ritchie Scott C, Grinberg Nastasiya F, Tang Howard Ho-Fung, Huang Qin Qin, Teo Shu Mei, et al. Multivariate Genome-wide Association Analysis of a Cytokine Network Reveals Variants with Widespread Immune, Haematological, and Cardiometabolic Pleiotropy. Am J Hum Genet 2019;105(6):1076–90. https://doi.org/10.1016/j.ajhg.2019.10.001.
[36] O'Reilly Paul F, Hoggart Clive J, Pomyen Yotsawat, Calboli Federico CF, Elliott Paul, Jarvelin Marjo-Riitta, et al. MultiPhen: Joint Model of Multiple Phenotypes Can Increase Discovery in GWAS. PLoS ONE 2012;7(5):e34861. https://doi.org/10.1371/journal.pone.0034861.
[37] Yang X, Zhang S, Sha Q. Joint Analysis of Multiple Phenotypes in Association Studies based on Cross-Validation Prediction Error. Sci Rep 2019;9:1–10. https://doi.org/10.1038/s41598-018-37538-y.
[38] Escofier B, Pagès J. Multiple factor analysis (AFMULT package). Comput Stat Data Anal 1994;18(1):121–40. https://doi.org/10.1016/0167-9473(94)90135-X.
[39] Vilor-Tejedor Natàlia, Alemany Silvia, Cáceres Alejandro, Bustamante Mariona, Mortamais Marion, Pujol Jesús, et al. Sparse multiple factor analysis to integrate genetic data, neuroimaging features, and attention-deficit/hyperactivity disorder domains. Int J Methods Psychiatr Res 2018;27(3). https://doi.org/10.1002/mpr.v27.310.1002/mpr.1738.
[40] Vilor-Tejedor Natalia, Ikram Mohammad Arfan, Roshchupkin Gennady V, Cáceres Alejandro, Alemany Silvia, Vernooij Meike W, et al. Independent Multiple Factor Association Analysis for Multiblock Data in Imaging Genetics. Neuroinformatics 2019;17(4):583–92. https://doi.org/10.1007/s12021-019-09416-z.
[41] Elhaik E. Why most Principal Component Analyses (PCA) in population genetic studies are wrong. BioRxiv 2021:2021.04.11.439381. 10.1101/2021.04.11.439381.
[42] Porter HF, O'Reilly PF. Multivariate simulation framework reveals performance of multi-trait GWAS methods. Sci Rep 2017;7:1–12. https://doi.org/10.1038/srep38837.
[43] Couvy-Duchesne Baptiste, Strike Lachlan T, McMahon Katie L, de Zubicaray Greig I, Thompson Paul M, Martin Nicholas G, et al. A Fast Method for Estimating Statistical Power of Multivariate GWAS in Real Case Scenarios: Examples from the Field of Imaging Genetics. Behav Genet 2019;49(1):112–21. https://doi.org/10.1007/s10519-018-9936-9.
[44] Anderson Marti J. A new method for non-parametric multivariate analysis of variance. Austral Ecol 2001;26(1):32–46. https://doi.org/10.1046/j.1442-9993.2001.01070.x.
[45] Zapala MA, Schork NJ. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. Proc Natl Acad Sci U S A 2006;103(51):19430–5. https://doi.org/10.1073/pnas.0609333103.
[46] Price Alkes L, Zaitlen Noah A, Reich David, Patterson Nick. New approaches to population stratification in genome-wide association studies. Nat Rev Genet 2010;11(7):459–63. https://doi.org/10.1038/nrg2813.
[47] Korte Arthur, Vilhjálmsson Bjarni J, Segura Vincent, Platt Alexander, Long Quan, Nordborg Magnus. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet 2012;44(9):1066–71. https://doi.org/10.1038/ng.2376.
[48] Zhou Xiang, Stephens Matthew. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Methods 2014;11(4):407–9. https://doi.org/10.1038/nmeth.2848.
[49] Furlotte NA, Eskin E. Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model. Genetics 2015;200:59–68. https://doi.org/10.1534/genetics.114.171447.
[50] Joo JWJ, Kang EY, Org E, Furlotte N, Parks B, Hormozdiari F, et al. Efficient and accurate multiple-phenotype regression method for high dimensional data considering population structure. Genetics 2016;204:1379–90. 10.1534/genetics.116.189712.
[51] Casale Francesco Paolo, Rakitsch Barbara, Lippert Christoph, Stegle Oliver. Efficient set tests for the genetic analysis of correlated traits. Nat Methods 2015;12(8):755–8. https://doi.org/10.1038/nmeth.3439.
[52] Beasley T Mark, Erickson Stephen, Allison David B. Rank-based inverse normal transformations are increasingly used, but are they merited? Behav Genet 2009;39(5):580–95. https://doi.org/10.1007/s10519-009-9281-0.
[53] Zhang Yiwen, Zhou Hua, Zhou Jin, Sun Wei. Regression Models for Multivariate Count Data. J Comput Graph Stat 2017;26(1):1–13. https://doi.org/10.1080/10618600.2016.1154063.
[54] Nowicka Malgorzata, Robinson Mark D. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. F1000Research

2016;5:1356. https://doi.org/10.12688/f1000research10.12688/f1000research.8900.1.
[55] Shriner D. Moving toward System Genetics through Multiple Trait Analysis in Genome-Wide Association Studies. Front Genet 2012;3:1. https://doi.org/10.3389/fgene.2012.00001.
[56] Sitlani Colleen M, Rice Kenneth M, Lumley Thomas, McKnight Barbara, Cupples L Adrienne, Avery Christy L, et al. Generalized estimating equations for genome-wide association studies using longitudinal phenotype data. Stat Med 2015;34(1):118–30. https://doi.org/10.1002/sim.v34.110.1002/sim.6323.
[57] Zhang Y, Xu Z, Shen X, Pan W. Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. Neuroimage 2014;96:309–25. https://doi.org/10.1016/j.neuroimage.2014.03.061.
[58] Vounou Maria, Nichols Thomas E, Montana Giovanni. Alzheimer's Disease Neuroimaging Initiative. Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. Neuroimage 2010;53(3):1147–59. https://doi.org/10.1016/j.neuroimage.2010.07.002.
[59] Vounou Maria, Janousova Eva, Wolz Robin, Stein Jason L, Thompson Paul M, Rueckert Daniel, et al. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. Neuroimage 2012;60(1):700–16. https://doi.org/10.1016/j.neuroimage.2011.12.029.
[60] Marchini Jonathan, Howie Bryan, Myers Simon, McVean Gil, Donnelly Peter. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 2007;39(7):906–13. https://doi.org/10.1038/ng2088.
[61] Stephens Matthew, Emmert-Streib Frank. A Unified Framework for Association Analysis with Multiple Related Phenotypes. PLoS ONE 2013;8(7):e65245. https://doi.org/10.1371/journal.pone.0065245.
[62] Hartley SW, Sebastiani P. PleioGRiP: genetic risk prediction with pleiotropy. Bioinformatics 2013;29(8):1086–8. https://doi.org/10.1093/bioinformatics/btt081.
[63] Wakefield Jon. Bayes factors for genome-wide association studies: comparison with *P* -values. Genet Epidemiol 2009;33(1):79–86. https://doi.org/10.1002/gepi.v33:110.1002/gepi.20359.
[64] Hartley SW, Monti S, Liu C-T, Steinberg MH, Sebastiani P. Bayesian Methods for Multivariate Modeling of Pleiotropic SNP Associations and Genetic Risk Prediction. Front Genet 2012;3:176. https://doi.org/10.3389/fgene.2012.00176.
[65] Zhu Hongtu, Khondker Zakaria, Lu Zhaohua, Ibrahim Joseph G. Bayesian Generalized Low Rank Regression Models for Neuroimaging Phenotypes and Genetic Markers. J Am Stat Assoc 2014;109(507):977–90. https://doi.org/10.1080/01621459.2014.923775.
[66] Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE. Systematic meta-analyses of Alzheimer disease genetic association studies: The AlzGene database. Nat Genet 2007. https://doi.org/10.1038/ng1934.
[67] Jack CR, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J Magn Reson Imaging 2008;27:685–91. https://doi.org/10.1002/jmri.21049.
[68] Lu ZH, Khondker Z, Ibrahim JG, Wang Y, Zhu H. Bayesian longitudinal low-rank regression models for imaging genetic data from longitudinal studies. Neuroimage 2017;149:305–22. https://doi.org/10.1016/j.neuroimage.2017.01.052.
[69] Greenlaw K, Szefer E, Graham J, Lesperance M, Nathoo FS. A Bayesian group sparse multi-task regression model for imaging genetics. Bioinformatics 2017;33:2513–22. https://doi.org/10.1093/bioinformatics/btx215.
[70] Wang C, Parmigiani G, Dominici F. Bayesian Effect Estimation Accounting for Adjustment Uncertainty. Biometrics 2012;68:661–71. https://doi.org/10.1111/j.1541-0420.2011.01731.x.
[71] Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet 2006;7:781–91. https://doi.org/10.1038/nrg1916.
[72] Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. Nat Rev Genet 2009;10:681–90. https://doi.org/10.1038/nrg2615.
[73] Green PJ, Łatuszyński K, Pereyra M, Robert CP. Bayesian computation: a summary of the current state, and samples backwards and forwards. Stat Comput 2015;25:835–62. https://doi.org/10.1007/s11222-015-9574-5.
[74] Adams HHH, Evans TE, Terzikhan N. The Uncovering Neurodegenerative Insights Through Ethnic Diversity consortium. Lancet Neurol 2019;18:915. https://doi.org/10.1016/S1474-4422(19)30324-2.
[75] Thompson PM, Jahanshad N, Ching CRK, Salminen LE, Thomopoulos SI, Bright J, et al. ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries. Transl Psychiatry 2020;10:1–28. https://doi.org/10.1038/s41398-020-0705-1.
[76] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 2018;562:203–9. https://doi.org/10.1038/s41586-018-0579-z.
[77] Alfaro-Almagro F, Jenkinson M, Bangerter NK, Andersson JLR, Griffanti L, Douaud G, et al. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. Neuroimage 2018;166:400–24. https://doi.org/10.1016/j.neuroimage.2017.10.034.
[78] Zhu X, Feng T, Tayo BO, Liang J, Young JH, Franceschini N, et al. Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. Am J Hum Genet 2015;96:21–36. https://doi.org/10.1016/j.ajhg.2014.11.011.
[79] Cichonska A, Rousu J, Marttinen P, Kangas AJ, Soininen P, Lehtimäki T, et al. metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. Bioinformatics 2016;32:1981–9. https://doi.org/10.1093/bioinformatics/btw052.
[80] Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat Genet 2018;50:229–37. https://doi.org/10.1038/s41588-017-0009-4.
[81] Guo B, Wu B. Integrate multiple traits to detect novel trait–gene association using GWAS summary data with an adaptive test approach. Bioinformatics 2019;35:2251–7. https://doi.org/10.1093/bioinformatics/bty961.
[82] Chung J, Marini S, Pera J, Norrving B, Jimenez-Conde J, Roquer J, et al. Genome-wide association study of cerebral small vessel disease reveals established and novel loci. Brain 2019;142:3176–89. https://doi.org/10.1093/brain/awz233.
[83] Wu C. Multi-trait genome-wide analyses of the brain imaging phenotypes in UK Biobank. Genetics 2020;215:947–58. https://doi.org/10.1534/genetics.120.303242.
[84] Liu N, Xu J, Liu H, Zhang S, Li M, Zhou Y, et al. Hippocampal transcriptome-wide association study and neurobiological pathway analysis for Alzheimer's disease. PLOS Genet 2021;17:. https://doi.org/10.1371/journal.pgen.1009363e1009363.
[85] Bulik-Sullivan B, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 2015;47:291–5. https://doi.org/10.1038/ng.3211.
[86] Xu Z, Shen X, Pan W. Longitudinal Analysis Is More Powerful than Cross-Sectional Analysis in Detecting Genetic Association with Neuroimaging Phenotypes. PLoS ONE 2014;9:. https://doi.org/10.1371/journal.pone.0102312102312.
[87] Harari JH, Dláz-Caneja CM, Janssen J, Martínez K, Arias B, Arango C. The association between gene variants and longitudinal structural brain changes in psychosis: A systematic review of longitudinal neuroimaging genetics studies. Npj Schizophr 2017;3:1–12. https://doi.org/10.1038/s41537-017-0036-2.
[88] Merritt K, Luque Laguna P, Irfan A, David AS. Longitudinal Structural MRI Findings in Individuals at Genetic and Clinical High Risk for Psychosis: A Systematic Review. Front Psychiatry 2021;12:. https://doi.org/10.3389/fpsyt.2021.620401620401.
[89] Vilor-Tejedor N, Ikram MA, Roshchupkin G, Vinke EJ, Vernooij MW, Adams HHH. Aging-Dependent Genetic Effects Associated to ADHD Predict Longitudinal Changes of Ventricular Volumes in Adulthood. Front Psychiatry 2020;11. 10.3389/fpsyt.2020.00574.
[90] Wachinger C, Nho K, Saykin AJ, Reuter M, Rieckmann A. A Longitudinal Imaging Genetics Study of Neuroanatomical Asymmetry in Alzheimer's Disease. Biol Psychiatry 2018;84:522–30. https://doi.org/10.1016/j.biopsych.2018.04.017.
[91] Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol 2017;18. https://doi.org/10.1186/s13059-017-1215-1.
[92] Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. Bioinform Biol Insights 2020;14:117793221989905. 10.1177/1177932219899051.
[93] Schouten TM, Koini M, de Vos F, Seiler S, van der Grond J, Lechner A, et al. Combining anatomical, diffusion, and resting state functional magnetic resonance imaging for individual classification of mild and moderate Alzheimer's disease. NeuroImage Clin 2016;11:46–51. https://doi.org/10.1016/j.nicl.2016.01.002.
[94] Sui J, Yu Q, He H, Pearlson GD, Calhoun VD. A selective review of multimodal fusion methods in schizophrenia. Front Hum Neurosci 2012;6:27. https://doi.org/10.3389/fnhum.2012.00027.
[95] Zhu D, Zhang T, Jiang X, Hu X, Chen H, Yang N, et al. Fusing DTI and fMRI data: a survey of methods and applications. Neuroimage 2014;102(Pt 1):184–91. https://doi.org/10.1016/j.neuroimage.2013.09.071.
[96] Gorgolewski KJ, Auer T, Calhoun VD, Craddock RC, Das S, Duff EP, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Sci Data 2016;3. https://doi.org/10.1038/sdata.2016.44.
[97] Moreau CA, Jean-Louis M, Blair R, Markiewicz CJ, Turner JA, Calhoun VD, et al. The genetics-BIDS extension: Easing the search for genetic data associated with human brain imaging. GigaScience 2020;9. https://doi.org/10.1093/GIGASCIENCE/GIAA104.
[98] Liu K, Yao X, Yan J, Chasioti D, Risacher S, Nho K, et al. Transcriptome-guided imaging genetic analysis via a novel sparse CCA algorithm. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10551 LNCS, Springer Verlag; 2017, p. 220–9. 10.1007/978-3-319-67675-3_20.
[99] Lista S, Garaci FG, Toschi N, Hampel H. Imaging epigenetics in Alzheimer's disease. Curr Pharm Des 2013;19:6393–415.
[100] Hampton T. Imaging Epigenetics in the Human Brain. JAMA 2016;316:1349. https://doi.org/10.1001/jama.2016.13667.
[101] Nazeri A, Ganjgahi H, Roostaei T, Nichols T, Zarei M. Alzheimer's Disease Neuroimaging Initiative. Imaging proteomics for diagnosis, monitoring and prediction of Alzheimer's disease. Neuroimage 2014;102:657–65. https://doi.org/10.1016/j.neuroimage.2014.08.041.
[102] YAN J, RISACHER SL, NHO K, SAYKIN AJ, SHEN L, Initiative FTADN. IDENTIFICATION OF DISCRIMINATIVE IMAGING PROTEOMICS ASSOCIATIONS IN ALZHEIMER'S DISEASE VIA A NOVEL SPARSE

CORRELATION MODEL. Biocomput. 2017, vol. 22, WORLD SCIENTIFIC; 2017, p. 94–104. 10.1142/9789813207813_0010.
[103] Moore R, Casale FP, Jan Bonder M, Horta D, Heijmans BT, Peter PA, et al. A linear mixed-model approach to study multivariate gene–environment interactions. Nat Genet 2019;51:180–6. https://doi.org/10.1038/s41588-018-0271-0.
[104] Wang C, Sun J, Guillaume B, Ge T, Hibar DP, Greenwood CMT, et al. A Set-Based Mixed Effect Model for Gene-Environment Interaction and Its Application to Neuroimaging Phenotypes. Front Neurosci 2017;11:191. https://doi.org/10.3389/fnins.2017.00191.
[105] Halldorsdottir T, Binder EB. Gene × Environment Interactions: From Molecular Mechanisms to Behavior. Annu Rev Psychol 2017;68:215–41. https://doi.org/10.1146/annurev-psych-010416-044053.
[106] Ge T, Nichols TE, Ghosh D, Mormino EC, Smoller JW, Sabuncu MR. A kernel machine method for detecting effects of interaction between multidimensional variable sets: An imaging genetics application. Neuroimage 2015;109:505–14. https://doi.org/10.1016/j.neuroimage.2015.01.029.
[107] Kircher T, Wöhr M, Nenadic I, Schwarting R, Schratt G, Alferink J, et al. Neurobiology of the major psychoses: a translational perspective on brain structure and function—the FOR2107 consortium. Eur Arch Psychiatry Clin Neurosci 2019;269:949–62. https://doi.org/10.1007/s00406-018-0943-x.
[108] Gu J, Kanai R. What contributes to individual differences in brain structure? Front Hum Neurosci 2014;8:262. https://doi.org/10.3389/fnhum.2014.00262.
[109] Wang H, Nie F, Huang H, Risacher SL, Saykin AJ, Shen L. Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. Bioinformatics 2012;28:. https://doi.org/10.1093/bioinformatics/bts228i127.
[110] Ning K, Chen B, Sun F, Hobel Z, Zhao L, Matloff W, et al. Classifying Alzheimer's disease with brain imaging and genetic data using a neural network framework. Neurobiol Aging 2018;68:151–8. https://doi.org/10.1016/j.neurobiolaging.2018.04.009.
[111] Peng J, An L, Zhu X, Jin Y, Shen D. Structured sparse kernel learning for imaging genetics based alzheimer's disease diagnosis. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9901 LNCS, Springer Verlag; 2016, p. 70–8. 10.1007/978-3-319-46723-8_9.
[112] Singanamalli A, Wang H, Madabhushi A, Weiner M, Aisen P, Petersen R, et al. Cascaded Multi-view Canonical Correlation (CaMCCo) for Early Diagnosis of Alzheimer's Disease via Fusion of Clinical, Imaging and Omic Features. Sci Rep 2017;7:1–14. https://doi.org/10.1038/s41598-017-03925-0.
[113] Zhou T, Thung KH, Zhu X, Shen D. Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. Hum Brain Mapp 2019;40:1001–16. https://doi.org/10.1002/hbm.24428.
[114] Bracher-Smith M, Crawford K, Escott-Price V. Machine learning for genetic prediction of psychiatric disorders: a systematic review. Mol Psychiatry 2021;26:70–9. https://doi.org/10.1038/s41380-020-0825-2.