Informatics in Medicine Unlocked 24 (2021) 100586

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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

Novel directions in data pre-processing and genome-wide association study (GWAS) methodologies to overcome ongoing challenges

Zahra Mortezaei [*], Mahmood Tavallaei [**]

Human Genetic Research Center, Baqiyatallah University of Medical Sciences, Tehran, Iran

ARTICLE INFO

Keywords:
Genome-wide association study (GWAS)
Sequencing
Machine learning
Retrospective association analysis
Tissue signals
Rare variants

ABSTRACT

A genome-wide association study (GWAS) is a standard population-based technique for identifying the heritable genetic basis of complex diseases by discovering correlations between trait variations and allele frequencies of genetic markers. This article aims to help fill gaps in data pre-processing and GWAS methodologies by reviewing novel techniques and methodologies. Data pre-processing performed prior to a GWAS presents challenges in Hardy-Weinberg (H–W) estimation, genotyping and accounting for factors such as sample structure. Recent developments towards overcoming these challenges are presented: the likelihood ratio test for H–W estimation, sequencing for genotyping, and techniques for dealing with sample structure. Traditional statistical methods cannot provide a way to insightfully interpret the data generated from high-throughput techniques; therefore, novel directions in GWAS methodologies are reviewed using efficient statistical methods, which are flexible techniques for performing genetic association analysis when factors such as non-random sampling or population structure occur. Despite the development of these methods, genotyping costs and an increased capacity for large dataset analysis have motivated researchers to examine tissue-specific signals. This review discusses how prospective and retrospective association analyses can be used to consider binary traits, address non-random ascertainment, and increase the capacity for large dataset analysis. Importantly, for disease susceptibility, rare variants can represent a large portion of genetic markers, and this article reviews some association methods for rare variant detection. In conclusion, the recent developments in GWAS data preparation and methodologies reviewed in this article can overcome most current challenges in the field and will also address future challenges.

1. Introduction

A genome-wide association study (GWAS) is a population-based technique that identifies correlations between trait variations and allele frequencies of genetic markers throughout the genome [1]. This article aims to fill gaps in GWAS pre-processing and methodologies by reviewing novel techniques and methodologies. Initially, new genotyping approaches are reviewed that reduce time and costs and that were developed by using next-generation sequencing (NGS). Then, in data pre-processing, in cases when sample structure is unobserved, several methods are reviewed to account for ancestry or family relatedness.

Traditional statistical methods cannot be used to insightfully interpret data generated from high-throughput techniques [2]. Novel directions in GWAS methodologies are reviewed that use robust machine learning methods such as likelihood-based methods that can be used to obtain more valuable results from a GWAS. One of the objectives of machine learning can be early-stage diagnosis and predictions that can be used to discover heritability from genetic data [2].

On the other hand, in genetic association studies, a sampling design ignoring variable components, population structure, non-genetic effects, gene-environment interactions, epistasis and phenotype-based ascertainment will confer a reduced statistical power, leading to type 1 errors and phenotype model misspecification. The approaches reviewed in this article that can be used when non-random ascertainment is common are prospective and retrospective studies [3]. Since the environment and multiple genetic factors play important roles in the aetiology of complex disorders, some GWAS methods that consider tissue-specific signals are reviewed. On the other hand, rare variants contribute substantially to disease susceptibility, and some GWAS methods for detecting rare variants are reviewed here.

[*] Corresponding author.
[**] Corresponding author.
E-mail addresses: zmortezaie@gmail.com (Z. Mortezaei), tavalla.mah@gmail.com (M. Tavallaei).

https://doi.org/10.1016/j.imu.2021.100586
Received 19 February 2021; Received in revised form 8 April 2021; Accepted 27 April 2021
Available online 15 May 2021

2. Linkage analysis and association mapping

Phenotypic variants are determined by genetic dissimilarity among individuals and are thus encoded by their DNA sequences. Many phenotypes are characterized as quantitative in nature and complex in aetiology, meaning that their mutational space is very large or that multiple genetic and environmental causes contribute to their variations. Linking phenotypic diversity and the genotypic diversity captured by genetic studies can provide genotype-to-phenotype mapping. Discovering the genetic basis of phenotypic traits can provide unprecedented views into the genetic architecture of phenotypic traits and their modes of inheritance [4].

Genetic linkage analysis is a powerful tool for assessing the tendency of genetic markers to be inherited together over generations on the basis of their physical proximity in the genome. This kind of analysis can be applied for rare disorders to successfully identify contributing genetic variants [5,6]. Linkage analysis involves using genetic markers throughout the genome as well as the genotypes of affected families to reveal the segregation of genetic markers with a disease. Linkage analysis has largely been applied to detect common variants. Linkage analysis relies on only one or two generations, and when some parental genotype data are missing, type I and II errors can be increased by incorrect marker allele frequencies [3].

The advantage of linkage analysis is its ability to detect variants with large effects; its disadvantage is its poor ability to identify small-effect variations. In general, the statistical power of genetic linkage analysis is the main concern in studies employing such analysis, and factors such as genotyping, allele frequencies, the strength of genetic effects and the heterogeneity of the locus or trait are fundamental to such studies [5,6]. Locus heterogeneity occurs when a disease is linked to multiple loci independently. Disease heterogeneity means that the disease has different subtypes, stages or grades and that different genes and functional processes may be linked to distinct disease grades or subtypes [7].

Association analysis is an alternative mapping method that is population based and useful for detecting small-effect variations using case-control individuals. A difference in the number of generations is one of the most important differences between linkage and association analyses. Association analysis is performed in a population of unrelated individuals to evaluate the association between a measured phenotype and genotyped genetic polymorphisms. The GWAS method is a population-based technique that identifies correlations between trait variations and allele frequencies of genetic markers throughout the genome. The differences between linkage and association analyses summarized in Fig. 1 may lead to inaccurate estimates of rare variants when using association analysis, which will be explained in detail in the last section of this review [1]. Initially, the coefficient of the kinship matrix is computed, and then different statistical methods can be compared to select an appropriate one for the GWAS and to identify single-nucleotide polymorphisms (SNPs) significantly associated with a specific disease. Such studies are useful in personalized medicine, which aims to identify the genetic risk factors and biological underpinnings of a specific trait [8].

Based on single-SNP analysis, the GWAS approach typically uses individual SNPs to assess their association with a phenotype of interest while ignoring other SNPs. However, for most complex phenotypic traits, an unexplained proportion of heritability is attributable to the polygenic nature of traits. To resolve this problem, some multi-locus association techniques have been developed to test for joint effects of multiple genetic variants on genes, pathways and traits [9]. Initially, the GWAS approach was applied to human diseases, resulting in great progress. Additionally, GWASs have been extended to the field of animal genetics and breeding in cases in which a large number of SNPs are available [10]. For example, in domestic animal breeding, GWASs have

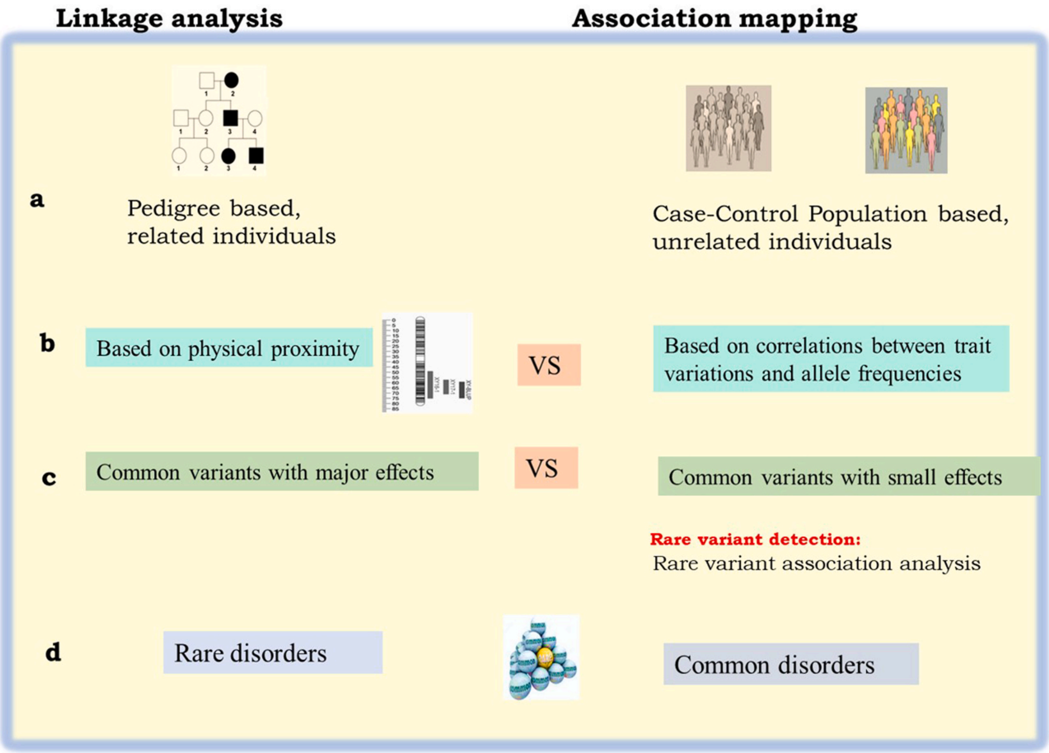
Linkage analysis
Association mapping
a
Pedigree based, related individuals
Case-Control Population based, unrelated individuals
b
Based on physical proximity
VS
Based on correlations between trait variations and allele frequencies
c
Common variants with major effects
VS
Common variants with small effects
Rare variant detection:
Rare variant association analysis
d
Rare disorders
Common disorders

Fig. 1. Linkage analysis and association mapping. a. A difference in the number of generations is one of the differences between linkage analysis and association mapping. **b**. Another difference between the two approaches is that the linkage analysis is based on physical proximity and the association study is based on the correlations between allele frequencies and trait variations. **c.** Detection of common variants with major effects or small effects is another difference between the two approaches; rare variant association analysis has been developed. **d**. Using the linkage analysis of rare disorders and the association studies of common disorders is one other difference between them.

been used to identify regions with pleiotropic effects on quantitative traits such as milk yield in dairy cattle and average daily gain in beef cattle. The GWAS approach can reveal the genetics of complex commercial traits, such as those of domestic animals, thereby allowing marker-associated selection to be performed [11].

Similar to the statistical power of linkage analysis, that of association studies aiming to detect underlying loci depends on available sample size, allele frequency, effect size, locus and disease heterogeneity and genotyping. Selection of an appropriate SNP array to increase genome coverage can result in the mapping of more relevant loci. In populations of European descent, a denser array can provide more significant SNPs in true association areas. Despite the availability of saturated GWAS arrays, using the latest Haplotype Reference Consortium panel [12] for such populations has some benefit with regard to imputing more SNPs. For both linkage and association analyses, disease heterogeneity has negative effects on statistical power and the ability to detect underlying loci because each locus might determine only a subset of cases and considering one SNP at a time in linkage or association analysis reveals only its marginal effect. One way to increase the power of linkage or association analysis and increase the marginal effect of the true locus is genetic enrichment. This consists of selecting special characteristics of the phenotype, such as cases with recurrence, early onset or a family history [1].

3. Developments in GWAS data pre-processing

Prior to performing a genetic association study, the local linkage disequilibrium (LD) structure of the whole genome needs to be quantified and assessed. To identify suitable tag SNPs and their genotypes for performing a GWAS and to study the LD pattern, a labour-intensive process of genotyping a large number of SNPs in a small subset of subjects must be carried out. This process is often performed for similar genomic areas by several researchers, resulting in a high degree of redundancy. Therefore, the International HapMap Project began in 2003 to estimate the LD between locus pairs [13]. Large-scale projects such as the 1000 Genomes Project [14] and the International HapMap Project have been used in the last 5–7 years to characterize genetic variants in different populations [15]. The efforts of the HapMap Project and selection of tag SNPs have decreased genotyping costs and have led to major improvements in related technology. Information from early proper GWASs was incorporated and used to develop SNP chips with adequate coverage for the entire genome. The resulting SNP chips have been used in recent studies and for most ethnicities to search for disease loci throughout the genome [16,17].

The HapMap Project has resulted in the most important functional genetic database and provided references for genetic association analyses and predicting the genetic causes of phenotypic traits. For example, genotype data for different populations can be downloaded from the HapMap website (http://hapmap.ncbi.nlm.nih.gov) [18–20]. Additionally, SNP allele frequencies can be obtained for a specific population and applied in comparisons and further analyses [21]. Haplotype-tagging SNPs for a specific population that can be used for association studies are also available from HapMap [22], and genetic relationships between different ethnic groups can be examined using HapMap results [23].

Estimates from association studies based on LD between SNPs can reveal correlated SNPs in an LD block that are inherited together [24]. Genetic association studies test for differences between case and control allele frequencies. To ensure that the case-control status is the only distinguishing characteristic in such studies, it is critical that all case and control samples be selected from the same population with the same ancestry. Failure to do so may lead to false positives, i.e., spurious significant associations between genetic variations and quantitative traits that originate from differences in population ancestry [1].

In addition, filtering out low-quality SNPs can minimize bias in GWASs. The minor allele frequency (MAF), missing call rate (MCR) and Hardy-Weinberg equilibrium (HWE) are common metrics used for SNP filtering. Any deviation from HWE may be the result of genotyping errors. Association studies and many other approaches rely on Hardy-Weinberg (H–W) assumptions to arrive at valuable and exact results [25]. Because the genetic markers on the X chromosome differ from those on autosomes, as applied in PLINK software [26], only females are considered for H–W proportion testing. Moreover, the "genetics" R package [27] used in population genetic studies does not distinguish between markers on autosomes and the X chromosome. Such analyses are inadequate and produce bias in testing for departure from H–W proportions. Therefore, four tests that consider both males and females and distinguish between autosomes and X chromosomes have been proposed for H–W proportions: the likelihood ratio test, the chi-square test, the permutation test and the exact test [28].

Genetic imputation is routinely used to improve the power of association studies. For common variants, the pre-filtering process usually does not have major benefits regarding the accuracy of imputation and may actually impair it because the pre-filtering process potentially weakens the strength of the LD structure, whereas imputation algorithms usually depend on the LD structure between available and missing genotypes. For example [29], investigated the performance of imputation algorithms such as MaCH and IMPUTE [30,31] while using pre-imputation filtering, and the authors concluded that very restrictive cut-offs are required for pre-filtering processes involving HWE. Prior to any downstream statistical analysis, when such restrictive pre-filtering has been applied for imputation, additional post-imputation quality control is suggested. With this information, an H–W proportion test prior to the imputation step might not be necessary but may be used as a post-imputation quality control procedure [32,33].

3.1. *Genotyping by sequencing*

The process of determining individual genetic variants is called genotyping, and different methods can be used for genotyping depending on the available resources and variants of interest. Using genotyping chips is an accurate and efficient method that searches for many common variants at once [25]. Dense arrays using chips with large overall coverage and many SNPs can be used for some populations, such as Africans, with a genome that has had more time to recombine [34]. In contrast to dense arrays, in silico genotyping of a particular region can provide a limited number of significant SNPs [12].

To study genetic variations, new alternative approaches to genotyping that reduce sequencing and time costs have been developed that utilize NGS. Modern sequencing technologies have greatly improved genetic mapping by increasing speed and resolution. To perform studies such as GWASs, genotyping by sequencing (GBS) is a low-cost procedure that can help reveal SNPs [25]. Indeed, large-scale population genetic analysis aiming to identify genetic variants is possible via GBS. Although GBS is a cost-effective method, its associated data analysis is complicated and requires complex bioinformatics because of extensive missing data. Some bioinformatics workflows have been developed for analysing GBS data, but additional packages are required to reduce the level of complexity of such data. Platforms for GBS data analysis include Fast-GBS, Stacks, GB-eaSy, IGST and Tassel-GBS [35,36].

3.2. *Sample structure*

With data pre-processing, the effectiveness of GWAS data analysis increases with the number of samples. GWAS can analyse associations between a sample's clinical conditions and single allelic variants. Association rule mining (ARM) can be used to identify multiple associations of allelic variants. As shown in Table 1, GWAS association rule mining in Spark (GARMS) is a scalable software framework comprising two steps, the pre-processing and mining association rules, for the frequent itemset [37].

Known or unknown sample structure, i.e., ancestral or familial

Table 1
GWAS Data preprocessing. Stages for GWAS data preprocessing and the methods using for them.

Preprocessing stage	Methods
Identifying SNP allele frequencies for a specific population	Using the results of the HapMap project
Sample selection	Selecting case and control samples from the same population with the same ancestry
Low-quality SNP filtering	Minor allele frequency (MAF), missing call rate (MCR)
Hardy-Weinberg equilibrium (HWE) to avoid genotyping errors	Likelihood ratio test, chi-squared test, permutation test and the exact test
Genetic imputation	MaCH, IMPUTE
Analyzing genotyping by sequencing data	Fast-GBS, Stacks, GB-eaSy, IGST and Tassel-GBS
Considering sample structures and family relatedness using principal component analysis	Generalized linear mixed model association tests (GMMATs), case-control retrospective association test (CARAT)
GWASs of binary traits	comprehensive R archive network (CRAN) package, GMMAT, learning and evaluating association patterns (LEAP) and the liability threshold-based mixed model association (LTMLM) statistic
Handle mining values and noises	(GWAS association rule mining in Spark) GARMS

relationships, is a common confounding factor in association studies. In cases in which the sample structure is unobserved, the linear mixed model (LMM) can be applied to account for ancestry or family relatedness by including principal components (PCs) based on the population structure [38]. The generalized LMM is a combination of the generalized linear model (GLM) and LMM. Generalized linear mixed model association tests (GMMATs) [39] and the case-control retrospective association test (CARAT) [40] are applicable to samples with a population structure or stratification. In the case of related samples, PC analysis must be used with care because of the occurrence of variants with low MAFs and an unstable relatedness matrix. Therefore, PC analysis based on common genetic variants can be employed to test for population structure in sequencing data [41]. Another study by Ref. [42] improved statistical methods for association studies, reducing the effect of a confounding population structure. Thus, appropriate methodologies accounting for population structure must be developed to apply sequencing data in rare variant association studies.

Because of the high computational cost, the application of GLMs in large-scale GWASs of binary traits is often limited. To overcome such problems and perform appropriate estimation, some fitting algorithms, such as the comprehensive R archive network (CRAN) package, and GMMAT [39] have been proposed. For a binary-trait GWAS, non-random ascertainment requires special attention. Several methods, such as learning and evaluating association patterns (LEAP) [43] and the liability threshold-based mixed model association (LTMLM) statistic [44], have been proposed for binary-trait GWASs dealing with case-control ascertainment. As an extension of the LTMLM method, the association statistic LT-Fam [44] can be used for case-control ascertainment in the case of a family-based design. Finally, to make the pre-processing section clearer, the methods reviewed in this section are summarized in Table 1.

4. Machine learning in GWASs

Traditional statistical methods cannot provide a way to insightfully interpret data generated from high-throughput techniques. To interpret and analyse large data, machine learning models have been developed. Machine learning has the potential to discover hidden patterns within genetic information that can help reveal disease pathogenesis. “Machine learning” is a synonymous term with “artificial intelligence” in which computers can make decisions by learning from data and with minimum human intervention. Machine learning approaches can also be applied to perform single or multiple SNP association studies to identify genotype-disease relations [45]. As shown in Table 2, machine learning approaches for GWAS range from simple regression analysis to random forest, deep learning models or other complex ensemble models [46]. On the other hand, machine learning approaches such as support vector machines can be applied to perform GWASs when combined with regression analysis [2].

When multiple causal variants occur at a locus associated with disease risk, conditional analysis can be performed. Furthermore, disease-associated variants can be prioritized using Bayesian approaches [47]. For example, previous meta-analyses of breast cancer have indicated a complex association pattern when multiple signals around a locus are involved. In these examples, conditional analysis was applied and indicated that some of the variants identified with the GWAS approach did not show similar residual associations, which may have resulted from strong links with functional variants [48,49].

A previous study by Wang et al. [50] evaluated the effect of using non-random samples for statistical inference of associations. The method included a likelihood-based statistical test using non-random samples and a conditional probability of the genotype at one locus given the genotype at another locus. For this analysis, gene segregation of marker and disease loci was assumed for randomly mating populations, with an “M” or “m” allele located at a genetic marker’s locus and an “A” or “a” allele at a disease locus; the “A” allele is the allele responsible for the disease. In case-control samples selected non-randomly, the conditional probability distribution of genotypes at two loci is approximately the same as the specific population’s genotypic distribution and is thus reliable for estimating population genetic parameters. As a result, the method developed by Wang et al. [51] yields large improvements in comparison with other approaches for evaluating LD in cases in which non-random samples are used. Some examples of such studies are discussed elsewhere [52,53].

For this kind of association analysis, the population genetic parameters that are used are the coefficient of LD (D) and the genotypic distributions at marker and disease loci, called the marker allele frequency (P) and disease allele frequency (q), respectively [51]. The conditional probability distribution of a marker genotype given a disease genotype and the conditional probability distribution of a disease genotype given a marker genotype are written in terms of population genetic parameters. In this analysis, the conditional probability distributions of marker and disease genotypes are latent variables. The expectation maximization (EM) algorithm can be used as a statistical method for estimating the maximum likelihood of unknown parameters in statistical models with latent or unobservable variables [54].

Table 2
GWAS methodologies. Different techniques to perform GWAS and some methodologies for them.

Performing GWAS	Methods
Machine learning in GWAS	Likelihood-based, random forest, deep learning, support vector machine
Retrospective association study	case-control retrospective association test (CARAT), CERAMIC, longitudinal binary-trait retrospective association test (LBRAT), retrospective generalized linear mixed model-based association test (RGMMAT)
GWAS on transcriptome	PrediXcan, Transcriptome-wide association studies (TWASs), summary Mendelian randomization (SMR)
Rare variant detection	Sequence kernel association test (SKAT), family-based SKAT (FamSKAT), burden tests, minimum p-value optimized nuisance parameter score test extended to relatives (MONSTER), pedigree disequilibrium test (PDT), variance component tests, omnibus tests, non-threshold rare (NTR) method

A likelihood-based machine learning GWAS method was developed by Wang et al. [51] and used in a study by Mortezaei et al. [55] on Parkinson's disease (PD). To further analyse the output of the GWAS, heritability and annotated genes including or close to significant SNPs have been studied. Similarly, such likelihood-based machine learning GWASs can be applied to other populations and different diseases, especially in cases in which population structure or non-random samples occur, to efficiently detect significant genetic loci in association with complex diseases.

4.1. *Heritability*

One of the objectives of machine learning can be early-stage diagnosis and predictions that can be used to discover heritability from genetic data [56]. GWASs have focused mainly on additive genetic effects, but the importance of non-additive effects in GWASs and genetic predictions have also been investigated. Increasing the accuracy of genetic studies and reducing bias are two benefits of accounting for non-additive genetic effects [57]. For example, dominance signals in association with milk yield have been identified near a candidate gene for milk production, PUNX2, in mice by using GWASs and population-based studies [58]. In humans, by using likelihood-based GWAS methods, the heritability or degree of passage of PD from parent to offspring has been compared via an additive genetic factor used as a coefficient in the likelihood formula. Additionally, dominant and recessive genetic factors have been compared between populations using likelihood-based machine learning GWAS approaches to determine the heritability of PD in specific populations [55]. As indicated in Fig. 2, similar methodologies for GWASs can be applied for different diseases and to compare heritability between populations.

5. Retrospective association analysis

In genetic association studies with a sampling design ignoring variable components, population structure, non-genetic effects, gene-environment interactions, epistasis and phenotype-based ascertainment can reduce statistical power, leading to type 1 errors and phenotype model misspecification. For example, many genetic variants have small effects on the polygenicity of a complex trait, and pleiotropy and multiple traits are affected by the same genetic variants, as shown in GWAS results. On the other hand, an active gene-environment correlation means that based on traits that are genetically influenced, individuals choose their environments. Therefore, pleiotropy can be environmentally mediated when a specific trait is influenced by genetics and then affects other traits, and it can predispose individuals to particular environments [59]. In such cases and to account for the mentioned conditions and covariates, many methods for quantitative trait analysis based on the standard LMM have recently been developed. Covariates play an important role in the association analysis of binary traits, and the LMM for binary traits is not a specified model, which can result in low performance. Other approaches that can be used when non-random ascertainment is common are prospective and retrospective studies. A retrospective study usually examines factors that may affect the study outcome by looking forward. In contrast, a prospective study tends to examine subjects for a period of time to track disease development [40,60]. When modelling genotype distribution based on covariates and the phenotype, retrospective association analysis can be applied for phenotype model misspecification [60]. However, because of the unknown trait model and strong effects of ascertainment, it is important to note that the sensitivity of retrospective association analysis to phenotype model misspecification is lower than that of prospective association analysis due to covariate-based ascertainment [40].

Common goals of retrospective binary-trait association mapping include increasing robustness to phenotype model misspecification, such as ascertainment, modelling the phenotype as binary with an appropriate variance and mean, achieving fast and accurate computation, and making appropriate corrections for various types of sample structures and related covariates. A retrospective binary trait association testing approach based on a mixed-effects quasi-likelihood framework, which includes great variability in fixed and random effects, called the case-control retrospective association test (CARAT), was proposed by Jiang et al. [40] and applied in a genome-wide analysis of Crohn's disease. The results revealed genetic regions with multiple independent association signals with Crohn's disease that may be used to identify risk factors. Another retrospective binary-trait association mapping method that helps increase power is called CERAMIC, which has been used to account for partially missing data [61].

The equation-based longitudinal binary-trait retrospective association test (LBRAT) has been proposed for genetic association studies of longitudinal binary phenotypes, and the retrospective generalized linear mixed model-based association test (RGMMAT) approach was developed as a retrospective scoring approach. Both methods have been applied for GWASs of cocaine use in a longitudinal cohort, whereby LBRAT detected loci with a significant association with cocaine use and was able to provide new insight into the genetic architecture [62]. Overall, the results from retrospective association tests can be applied to further develop such approaches and apply them to different binary

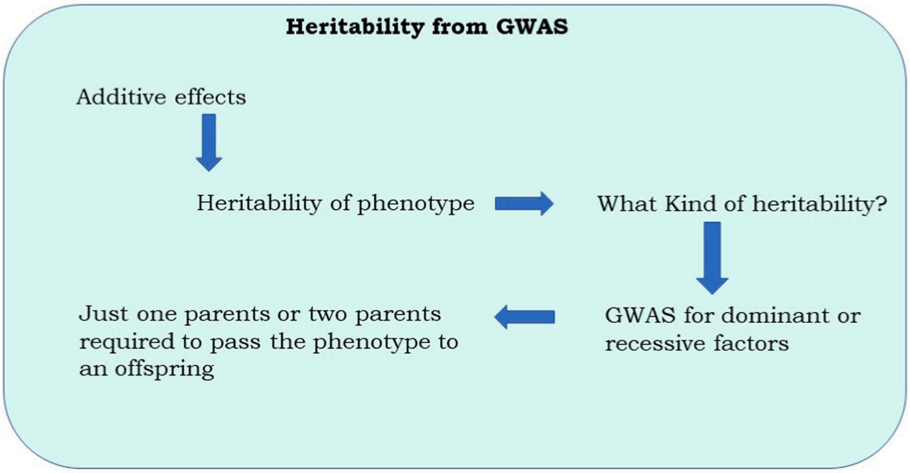
Heritability from GWAS
Additive effects
Heritability of phenotype
What Kind of heritability?
GWAS for dominant or recessive factors
Just one parents or two parents required to pass the phenotype to an offspring

Fig. 2. Heritability from GWASs. Additive genetic factors can be applied in GWASs to estimate the level of heritability for a specific phenotype. In a case in which a trait has been found to be heritable in a specific population, dominant and recessive genetic factors can be used to infer whether inheritance from one or both parents is needed for the phenotype to be acquired by offspring.

traits. In general, genotyping costs and increasing capacity for large dataset analysis have motivated researchers to examine tissue-specific signals.

6. Considering tissue-specific signals

As the environment and multiple genetic factors play important roles in the aetiology of complex disorders, to genetically study complex diseases, germline and non-germline variants are crucial and need to be considered. Non-germline genetic mutations spontaneously occurring in somatic cells during a person's lifetime are called somatic mutations. Some of these mutations can alter important cellular functions, and progressive accumulation of this type of mutation can cause complex diseases such as cancer [63]. In fact, somatic mutations have been studied most extensively in relation to cancer, but they can also cause neurodegenerative disease (ND) when they influence brain development at a different time point in life or at the pre-natal stage. For example, PD can be a result of somatic mutations in *PARK2*, *SNCA*, and the gene encoding Parkin [64].

In studies on somatic mutation burden, the number of somatic mutations per donor can be used to perform GWASs. Under a normal somatic mutation burden, the first GWAS performed by Ref. [65] identified approximately 20 sites in association with the somatic mutation burden and 2 sites with post-transcriptional modification, and cell proliferation may be promoted in a tissue-specific manner by the somatic mutations identified in that study; the study also identified some candidate genes with probable roles in initiating tumorigenesis. Such analyses have been performed before detecting the cancer phenotype when tissue-specific mutations in actively expressed genes have been identified.

At the post-transcriptional modification level, mitochondrial polymorphism information can be used to perform GWASs. GWASs of mitochondrial tRNA can be carried out to validate the somatic mutation call set, which may include signals from somatic mutations or RNA editing. Additionally, GWASs of mitochondrial tRNA can be employed for post-transcriptional modification analysis. Tumorigenesis initiation varies among tissues of the body, and this fact needs to be considered when assessing GWAS results. As the results of such an analysis can identify new oncogene or tumour suppressor candidates, GWASs can identify tRNA and search for somatic mutations as organismal-level variants. By detecting tumorigenesis mechanisms or novel oncogenes, the results of such analysis can be used for disease detection and healthy tissue differentiation [65].

Transcriptome studies have been developed to assay genotypes and expression levels for a large number of individuals [66–68]. A comprehensive cross-tissue survey called the genotype-tissue expression project (GTEX) has collected DNA and RNA sequence data from multiple tissue samples of approximately 1000 individuals to examine genetic variations at the transcript level [69,70]. To estimate phenotypic variations, the mediating effects of gene expression levels can also be tested using a gene-level association approach called PrediXcan, which was developed to shed light on the biology of complex diseases and to integrate knowledge from transcriptome studies. This is because most trait associations are tissue specific [71]. Similar to PrediXcan, transcriptome-wide association studies (TWASs) [72] and summary Mendelian randomization (SMR) analyses [73] may be used to estimate associations between phenotypes and gene expression levels. The only differences between the TWAS and PrediXcan approaches are the implementation and prediction model used. In some cases, associations between a specific trait and unexpected tissues have been detected; to discover their mechanisms, agnostic scanning of a tissue set is required [74].

Five types of omics data used for molecular-type GWASs of Crohn's disease were collected in one study by Pei et al. [75] to investigate tissues relevant to the disease and then conduct TWASs on those tissues. Tissue-specific enrichment analysis (TSEA) with the R package deTS was utilized to collect tissues in which GWAS-detected genes were expressed. In that study, the MetaXcan method was used for the TWAS and to estimate expression levels that were genetically regulated, and the results revealed the three tissues most related to Crohn's disease.

7. Rare variant detection

SNP genetic markers identified by GWASs and showing significant associations with complex traits are common variants with an MAF $\geq$5% [76]. However, some evidence indicates that rare variants represent almost 95% of genetic variants associated with complex traits [77]. Rare variants are substantially associated with disease susceptibility; therefore, interactions among SNPs using multi-locus associations have some benefits in the identification of high-risk proportions of rare variants [78]. Most methods for rare variant detection are classified as kernel association tests or burden association tests [79–81]. One of the kernel association methods that can be applied to detect rare variants is the sequence kernel association test (SKAT) [82]. Family-based SKAT (FamSKAT) [83] may also be used to test for associations. For related samples, burden tests have higher power than FamSKAT due to the causal effects of most genetic variants and the direction of the effects [25].

Other methods originally used for association tests of rare variants on unrelated samples have been extended to related samples. For example, the minimum p-value optimized nuisance parameter score test extended to relatives (MONSTER) method developed by Jiang and McPeek [84] comprises a combination of features from FamSKAT, and the burden test and has higher power than each. The pedigree disequilibrium test (PDT) [25] is a family-based association test that considers discordant sibships and nuclear families in each pedigree and can be phenotypically informative. PDT is a robust method for identifying rare variants, but its results for rare variant markers should be considered carefully.

Because rare variants are abundant in the genome but generally do not correlate with each other, a stringent threshold is required to detect these variants [85]. Such restrictions lead to power loss in the detection of rare variants, which may be mitigated by using single-variant tests when rare variants have large effect sizes and the sample size is very large [86]. Region-based analysis can be used for rare variant association studies. This kind of study identifies the joint association of genetic regions with phenotypic traits. Region-based methods include variance component tests [87], omnibus tests [79] and others [88–91]. Another region-based variant detection method, called the non-threshold rare (NTR) method, accounts for effect directions and does not use a threshold [92]. PC analysis, a pedigree-based kinship matrix or a genetic relatedness matrix can be used to address population stratification in rare variant association studies. The estimation of PCs and the genetic relatedness matrix using rare variants can be unstable because of the low MAF [93]. Regardless, further methodological developments are required to resolve this issue of population structure in rare-variant association studies.

Using population-based sequencing data can increase the resolution of microarray-based GWASs in cases where variants are not genotyped directly [94]. Single-marker analyses have essentially no power to detect rare variants in sequencing data, and joint consideration of all rare variants within a genetic region can be used instead. For sequencing data and dealing with rare variants, PDT can be implemented using collapsing methods [95,96]. In an analysis of sequence data, it is important to select and group variables into a unit and determine a proper region. One strategy is to select genes; another is to construct regions based on the number of variants or sliding windows on a particular chromosome [97].

In contrast to GWASs, methods used for rare variant sequencing studies do not include a significance threshold at the genome scale because of sequencing platforms, variant aggregation, ancestry, sample sizes and coverage depth. Based on different assumptions, the threshold range is from 1×10^{-9} to 3.75×10^{-7} for a single-variant test [98–100].

Because of the number of genes in the genome and the corresponding Bonferroni correction, gene-based tests performed at the genome scale can have a 2.5×10^{-6} threshold [101]. However, for this threshold, individual gene correlations are not considered, representing one of the limitations of this approach [102]. In rare variant analysis, the appropriateness of the significance threshold cannot be properly assessed, and how to handle multiple testing for data generated via high-throughput sequencing remains an open question. Finally, to clearly present the reviewed GWAS methodologies, the methods are summarized in Table 2.

8. Discussion

To fill gaps in data pre-processing and GWAS methodologies, novel techniques have been reviewed, and to the best of our knowledge, this is the first review article covering such important issues. For example, recent developments in HWE assessment can consider both males and females and distinguish autosomes and X chromosomes. In addition, a cost-effective GBS method has been developed to identify genetic variants via the GWAS approach. Furthermore, machine learning methods have the potential to discover hidden patterns within genetic information that can help reveal disease pathogenesis. Although typical GWASs examine individual SNPs and test for their associations with a phenotype of interest and ignore other SNPs, multi-locus association techniques have been reviewed that were recently developed to assess the joint effects of multiple genetic variants on genes, pathways, and traits. Despite the methods developed to prepare data and perform GWASs, genotyping costs and increasing capacity for large dataset analysis have recently motivated researchers to examine tissue-specific signals. Additionally, recent developments in retrospective association analysis may be employed for phenotype model misspecification based on covariates and the phenotype. On the other hand, based on the GWAS results, one way to increase the power of association analysis and the marginal effect of the true locus is genetic enrichment, which consists of selecting special characteristics of the phenotype.

9. Conclusions

In conclusion, recent developments in GWAS data pre-processing and methodologies, which are reviewed in this article, can overcome most of the current challenges in this field and help address future challenges. These methods can be applied to generate more robust GWAS results to discover heritability from genetic data when factors such as non-random sampling or population structure occur. It was concluded that in some cases, associations between a specific trait and unexpected tissues have been detected, and agnostic scanning of a tissue set is required to discover their mechanisms. In addition, for the detection of rare variants from association studies, single-marker analysis has essentially no power, and joint consideration of all rare variants in a genetic region can be used instead. In summary, recent developments and novel directions in data pre-processing and the application of GWASs have made them more cost-effective and allowed for more accurate results.

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.

References

[1] Nsengimana J, Bishop DT. Design considerations for genetic linkage and association studies. In: Elston RC, editor. Methods in mol biol. second ed., vol. 1666; 2017. p. 257–81.

[2] Maciukiewicz M, Marshe VS, Hauschild A-C, Foster JA, Rotzinger S, Kennedy JL, et al. GWAS-based machine learning approach to predict duloxetine response in major depressive disorder. J Psychiatr Res 2018;99:62–8.

[3] Ott J, Wang J, Leal SM. Genetic linkage analysis in the age of whole-genome sequencing. Nat Rev Genet 2015;16(5):275–84.

[4] Bush WS, Moore JH. Chapter 11 : genome-wide association studies. PLoS Comput Biol 2012;8(12):e1002822.

[5] Balding DJ, Bishop M, Canning C. Handbook of statistical genomics. fourth ed. John Wiley & Sons Ltd; 2019.

[6] Lobo I, Shaw K. Discovery and types of genetic linkage. Nat. Educ. 2017;1(1):139.

[7] Zan Y. Understanding the genetic basis of complex traits. In: Digital comprehensive summaries of Uppsala dissertations from the Faculty of medicine 1438. Uppsala: Acta Universitatis Upsaliensis; 2018, ISBN 978-91-513-0260-7.

[8] Hoffman GE. Correcting for population structure and kinship using the linear mixed Model : theory and extensions. PLoS One 2013;8(10):e75707.

[9] Pan W, Kwak I-Y, Wei P. A powerful pathway-based adaptive test for genetic association with common or rare variants. Am J Hum Genet 2015;97(1):86–98.

[10] Zhang H, Wang Z, Wang S, Li H. Progress of genome wide association study in domestic animals. J Anim Sci Biotechnol 2012;3(1):26.

[11] Zhang W, Gao X, Shi X, Zhu B, Wang Z, Gao H, et al. PCA-based multiple-trait GWAS analysis: a powerful moldel for exploring pleiotropy. Animals (Basel) 2018;8(12):239.

[12] Das S, Forer L, Schoenherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet 2016;48(10):1284–7.

[13] The International HapMap Consortium. The International HapMap project. Nature 2003;426(6968):789–96.

[14] The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 2015;526(7571):68–74.

[15] Zheng-Bradley X, Flicek P. Applications of the 1000 Genomes project resources. Briefings Funct Genomics 2017;16(3):163–70.

[16] Hua F, Guo Y, Sun Q, Yang L, Gao F. HapMap-based study: CYP2A13 may be a potential key metabolic enzyme gene in the carcinogenesis of lung cancer in non-smokers. Thorac Canc 2019;10:601–6.

[17] Laird NM, Lange C. The fundamentals of modern statistical genetics (statistics for biology and health). New York: Springer Publishing Company; 2011.

[18] Al-Eitan LN, Mohammad NN, Al-Mogableh HW, Hakooz NM, Dajani RB. Genetic polymorphisms of pharmacogenomics VIP variants in the Circassian subpopulation from Jordan. Curr Drug Metabol 2019;20(8):674–81.

[19] Cao B, Yang M, Kang G, Li R, Zhu X, Kang Q, et al. The relationship between gene polymorphism of miRNA regulating FGA and Schizophernia. Open Access Maced J Med Sci 2019;7(9):1436–9.

[20] Chen W, Ding H, Cheng Y, Li Q, Dai R, Yang X, et al. Genetic polymorphisms analysis of pharmacogenomic VIP variants in Bai ethnic group from China. Mol Genet Genomic Med 2019;7(9):e884.

[21] Hung CS, Huang CY, Hsu YW, Makandi PT, Chang WC, Chang YJ, et al. HSPB1 rs2070804 polymorphism is associated with the depth of primary tumor. J Cell Biochem 2020;121(1):63–9.

[22] Tang X-J, Shentu X-C, Tang Y-L, Ping X-Y, Yu X-N. The impact of GJA3 SNPs on susceptibility to age-related cataract. Int J Ophthalmol 2019;12(6):1008–11.

[23] Thomson RJ, McMorran B, Hoy W, Jose M, Whittock L, Thornton T, et al. New genetic loci associated with chronic kidney disease in an indigenous Australian. Front Genet 2019;10(330).

[24] Bambury RM, Gallagher DJ. Prostate cancer : germline prediction for a commonly variable malignancy. BJU Int 2012;110(11c):E809–18.

[25] Elston RC. Statistical human genetics: methods and protocols. Methods Mol Biol 2017:1666. Springer Nature.

[26] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81(3):559–75.

[27] Warnes G, Gorjanc G, Leisch F, Man M. Genetics: population genetics. R package version 1.3.8.1. The Comprehensive R Archive Network; 2013. https://CRAN.R-project.org/package=genetics.

[28] Graffelman J, Weir BS. Testing for Hardy-Weinberg equilibrium at biallelic genetic markers on the X chromosome. Heredity (Edinb) 2016;116(6):558–68.

[29] Roshyara NR, Kirsten H, Horn K, Ahnert P, Scholz M. Impact of pre-imputation SNP-filtering on genotype imputation results. BMC Genet 2014;15(1):88.

[30] Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. Bioinformatics 2015;31(5):782–4.

[31] Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 2012;44(8):955–9.

[32] Lert-Itthiporn W, Suktitipat B, Grove H, Sakuntabhai A, Malasit P, Tangthawornchaikul N, et al. Validation of genotype imputation in Southeast Asian populations and the effect of single nucleotide polymorphism annotation on imputation outcome. BMC Med Genet 2018;19(1).

[33] Southam L, Panoutsopoulou K, Rayner NW, Chapman K, Durrant C, Ferreira T, et al. The effect of genome-wide association scan quality control on imputation outcome for common variants. Eur J Hum Genet 2011;19(5):610–4.

[34] Makina SO, Taylor JF, van Marle-Köster E, Muchadeyi FC, Makgahlela ML, MacNeil MD, et al. Extent of linkage disequilibrium and effective population size in four South African Sanga cattle breeds. Front Genet 2015;6(337).

[35] Pavan S, Curci PL, Zuluaga DL, Blanco E, Sonnante G. Genotyping-by-sequencing highlights patterns of genetic structure and domestication in artichoke and cardoon. PLoS One 2018;13(10):e0205988.

[36] Wickland DP, Battu G, Hudson KA, Diers BW, Hudson ME. A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. BMC Bioinf 2017;18(1):586.

[37] Agapito G, Guzzi PH, Cannataro M. An efficient and scalable SPARK preprocessing methodology for Genome Wide Association Studies. In: 2020 28th Euromicro International conference on parallel, distributed and network-based processing (PDP), vol. 1; 2020. p. 369–75.

[38] Conomos MP, Laurie CA, Stilp AM, Gogarten SM, McHugh CP, Nelson SC, et al. Genetic diversity and association studies in US Hispanic/Latino populations: applications in the Hispanic community health study/study of Latinos. Am J Hum Genet 2016;98(1):165–84.

[39] Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. Am J Hum Genet 2016;98(4):653–66.

[40] Jiang D, Zhang S, McPeek MS. Retrospective binary-trait association test elucidates genetic architecture of Crohn disease. Am J Hum Genet 2016;98(2): 243–55.

[41] O'Connor D, Png E, Khor CC, Snape MD, Hill AVS, van der Klis F, et al. Common genetic variations associated with the persistence of immunity following childhood immunization. Cell Rep 2019;27(11):3241–53.

[42] Lee T, Lee I. Genome-wide association studies in Arabidopsis thaliana: statistical analysis and network-based augmentation of signals. In: Sanchez-Serrano JJ, Salinas J, editors. Arabidopsis protocols. Methods in molecular biology; 2021. 2200.

[43] Weissbrod O, Lippert C, Geiger D, Heckerman D. Accurate liability estimation improves power in ascertained case-control studies. Nat Methods 2015;12(4): 332–4.

[44] Hayeck TJ, Loh PR, Pollack S, Gusev A, Patterson N, Zaitlen NA, et al. Mixed model association with family-biased case-control ascertainment. Am J Hum Genet 2017;100(1):31–9.

[45] Nicholls HL, John CR, Watson DS, Munroe PB, Barnes MR, Cabrera CP. Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci. Front Genet 2020;11:350.

[46] Sun T, Wei Y, Chen W, Ding Y. Genome-wide association study-based deep learning for survival prediction. Stat Med 2020:1–16. 2020.

[47] Spain SL, Barrett JC. Strategies for fine-mapping complex traits. Hum Mol Genet 2015;24(R1):R111–9.

[48] Glubb DM, Maranian MJ, Michailidou K, Pooley KA, Meyer KB, Kar S, et al. Fine-scale mapping of the 5q11.2 breast cancer locus reveals at least three independent risk variants regulating MAP3K1. Am J Hum Genet 2015;96(1):5–20.

[49] Wu Y, Gao H, Li H, Tabara Y, Nakatochi M, Chiu YF, et al. A meta-analysis of genome-wide association studies for adiponectin levels in East Asians identifies a novel locus near WDR11-FGFR2. Hum Mol Genet 2014;23(4):1108–19.

[50] Wang M, Jia T, Jiang N, Wang L, Hu X, Luo Z. Inferring linkage disequilibrium from non-random samples. BMC Genom 2010;11(1):328.

[51] Wang M, Wang L, Jiang N, Jia T, Luo Z. A robust and efficient statistical method for genetic association studies using case and control samples from multiple cohorts. BMC Genom 2013;14(1):88.

[52] Gondro C, van der Werf J, Hayes B. Genome-wide association studies and genomic prediction. Humana Press. London: Springer; 2013. ISBN 978-1-62703-447-0.

[53] Park KI. Fundamentals of probability and Stochastic processes with applications to communications. Springer International Publishing AG; 2018.

[54] Yu K, Dang X, Bart H, Chen Y. Robust model-based learning via spatial EM-algorithm. IEEE Trans Knowl Data Eng 2015;27(6):1670–82.

[55] Mortezaei Z, Lanjanian H, Masoudi-nejad A. Candidate novel long noncoding RNAs, MicroRNAs and putative drugs for Parkinson's disease using a robust and efficient genome-wide association study. Genomics 2017;109(3–4):158–64.

[56] Su C, Tong J, Wang F. Mining genetic and transcriptomic data using machine learning approaches in Parkinson's disease. NPJ Parkinson's Dis 2020;6(24).

[57] Lopes MS, Bastiaansen JW, Harlizius B, Knol EF, Bovenhuis H. A genome-wide association study reveals dominance effects on number of teats in pigs. PLoS One 2014;9(8):e31825.

[58] Jiang J, Shen B, O'Connell JR, vanRaden PM, Cole JB, Ma L. Dissection of additive, dominance, and imprinting effects fpr production and reproduction traits in Holstein cattle. BMC Genom 2017;18(1):425.

[59] Avinun R. The E is in the G: gene-environment-trait correlations and findings from genome-wide association studies. Perspect Psychol Sci 2020;15(1):81–9.

[60] Jiang D, Mbatchou J, McPeek MS. Retrospective association analysis of binary traits; overcoming some limitations of the additive polygenic model. Hum Hered 2015;80(4):187–95.

[61] Zhong S, Jiang D, McPeek MS. CERAMIC: case-control association testing in samples with related individuals based on retrospective mixed model analysis with adjustment for covariates. PLoS Genet 2016;12(10):e1006329.

[62] Wu W, Wang Z, Xu K, Zhang X, Amei A, Gelernter J, et al. Retrospective association analysis of longitudinal binary traits identifies important loci and pathways in cocaine use. Genetics 2019;213(4):1225–36.

[63] Martincorena I, Campbell P. Somatic mutation in cancer and normal cells. Science 2015;349(6255):1483–9.

[64] Poduri A, Evrony GD, Cai X, Walsh CA. Somatic mutation, germline variants, and neurological disease. Science 2013;341(6141):1237758.

[65] Drubin CW, Ramu A, Rockweiler NB, Conrad DF. Somatically mutated genes under positive and negative selection found by transcriptome sequence analysis include oncogene and tumor supressor candidates. 2018. BioRxiv, https://www.biorxiv.org/content/10.1101/396739v1.

[66] Battle A, Mostafavi S, Zhu X, Potash JB, Weissmann MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA-Sequencing of 922 individuals. Genome Res 2013;24(1):14–24.

[67] Lappalainen T, Sammeth M, Friedlander MR, 't Hoen PAC, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in human. Nature 2013;501(7468):506–11.

[68] Zhang X, Joehanes R, Chen BH, Huan T, Ying S, Munson PJ, et al. Identification of common genetic variants controlling transcript isoform variation in human whole blood. Nat Genet 2015;47(4):345–52.

[69] GTEx Consortium. Genetic effects on gene expression across human tissues. Nature 2017;550(7675):204–13.

[70] The GTEx Consortium. The genotype-tissue expression (GTEx) project. Nat Genet 2013;45(6):580–5.

[71] Gamazon ER, Wheeler HE, Shah KP. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet 2015;47(9):1091–8.

[72] Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet 2016;48(3):245–52.

[73] Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet 2016;48(5):481–7.

[74] Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, et al. Exploring the phenotypic consequences of tissue specific gene expression validation inferred from GWAS summary statistics. Nat Commun 2018;9(1).

[75] Pei G, Dai Y, Zhao Z, Jia P. deTS: tissue-specific enrichment analysis to decode tissue specificity. Bioinformatics 2019;35(19):3842–5.

[76] Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 2012;337(6090):100–4.

[77] Ouyang W, Xiaofeng Z, Qin H. Detecting multiethnic rare variants. In: Elston RC, editor. (2017) Statistical human genetics: methods and protocols. Methods in molecular biology, vol. 1666; 2017. p. 527–38.

[78] Mooney M, Wilmot B, Bipolar Genome Study T, McWeeney S. The GA and the GWAS: using genetic algorithms to search for multilocus associations. IEEE ACM Trans Comput Biol Bioinf 2012;9(3):899–910.

[79] Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet 2012;91(2):224–37.

[80] Schaid DJ, McDonnel SK, Sinnwell JP, Thibodeau SM. Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. Genet Epidemiol 2013;37(5):409–18.

[81] Wang X, Zhang Z, Morris N, Cai T, Lee S, Wang C, et al. Rare variant association test in family-based sequencing studies. Briefings Bioinf 2016;18(6):954–61.

[82] Hamazaki K, Iwata H. RAINBOW: Haplotype-based genome wide association study using a novel SNP-set method. PLoS Comput Biol 2020;16(2):e1007663.

[83] Chen H, Meigs JB, Dupuis J. Sequence kernel association test for quantitative traits in family samples. Genet Epidemiol 2013;37(2):146–204.

[84] Jiang D, McPeek MS. Robust rare variant association testing for quantitative traits in samples with related individuals. Genet Epidemiol 2014;38(1):10–20.

[85] Auer PL, Lettre G. Rare variant association studies: considerations, challenges and opportunities. Genome Med 2015;7(16).

[86] Nicolae DL. Association tests for rare variants. Annu Rev Genom Hum Genet 2016;17:117–30.

[87] Zhan X, Plantinga A, Zhao N, Wu MC. A fast small-sample kernel independence test for microbiome community-level association analysis. Biometrices 2017;73 (4):1453–63.

[88] Barnett IJ, Lin X. Analytic p-value calculation for the higher criticism test in finite d problems. Biometrika 2014;101(4):964–70.

[89] Barnett I, Mukherjee R, Lin X. The generalized linear criticism for testing SNP-set effects in genetic association studies. J Am Stat Assoc 2017;112(517):64–76.

[90] Mukherjee R, Pillai NS, Lin X. Hypothesis testing for high-dimensional sparse binary regression. Ann Stat 2015;43(1):352–81.

[91] Wu Z, Sun Y, He S, Cho J, Zhao H, Jin J. Detection boundary and Higher Criticism approach for rare and weak genetic effects. Ann Appl Stat 2014;8(2):824–51.

[92] Hsieh A-R, Chen D-P, Chattopadhyay AS, Li Y-J, Chang C-C, Fann CSJ. A non-threshold region-specific method for detecting rare variants in complex diseases. PLoS One 2017;12(11):e0188566.

[93] Thornton TA. Statistical methods for genome-wide and sequencing association studies of complex traits in related samples. Curr Protoc Hum Genet 2015;84. 1.28.1-1-28.9.

[94] Gallagher MD, Chen-Plotkin AS. The post-GWAS era: from associations to function. Am J Hum Genet 2018;102(5):717–30.

[95] Kinnamon DD, Hershberger RE, Martin ER. Reconsidering association testing methods using single-variant test statistics as alternative to pooling tests for sequencing data with rare variants. PLoS One 2012;7(2):e30238.

[96] Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 2010;34(2):188–93.

[97] Sung YJ, Korthauer KD, Swartz MD, Engelman CD. Methods for collapsing multiple rare variants in whole-genome sequence data. Genet Epidemiol 2014;38 (S1):S13–20.

[98] Fadista J, Manning AK, Florez JC, Groop L. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. Eur J Hum Genet 2016;24(8):1202–5.

[99] Pulit SL, de With SAJ, de Bakker PIW. The multiple testing burden in sequencing-based disease studies of global populations. bioRXiv; 2016. https://doi.org/10.1101/053264. https://www.biorxiv.org/content/10.1101/053264v1.

[100] Sobota RS, Shriner D, Kodaman N, Goodloe R, Zheng W, Gao YT, et al. Addressing population-specific multiple testing burdens in genetic association studies. Ann Hum Genet 2015;79(2):136–47.

[101] Moutsianas L, Agarwala V, Fuchsberger C, Flannick J, Rivas MA, Gaulton KJ, et al. The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. PLoS Genet 2015;11(4):e1005165.

[102] Greenwood CMT, Xu C, Ciampi A. Significance thresholds for rare variant signals. In: Zeggini E, Morris A, editors. Assessing rare variation in complex traits: design and analysis of genetic studies. New York: Springer-Verlag; 2015. 2015.