ANNUAL REVIEWS

Annual Review of Biomedical Data Science

Statistical Methods in Genome-Wide Association Studies

Ning Sun and Hongyu Zhao

Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut 06520, USA; email: hongyu.zhao@yale.edu

Annu. Rev. Biomed. Data Sci. 2020. 3:265–88

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-030320-041026

Keywords

genome-wide association study, heritability, risk prediction, genome annotation, pleiotropy, genetic architecture

Abstract

Since the initial success of genome-wide association studies (GWAS) in 2005, tens of thousands of genetic variants have been identified for hundreds of human diseases and traits. In a GWAS, genotype information at up to millions of genetic markers are collected from up to hundreds of thousands of individuals, together with their phenotype information. Several scientific goals can be accomplished through the analysis of GWAS data, including the identification of variants, genes, and pathways associated with diseases and traits of interest; the inference of the genetic architecture of these traits; and the development of genetic risk prediction models. In this review, we provide an overview of the statistical challenges in achieving these goals and recent progress in statistical methodology to address these challenges.

1. INTRODUCTION

Since Hoh and colleagues (1) first identified the association between the complement factor H gene and age-related macular degeneration through a genome wide association study (GWAS), this study design has been used to find associations between tens of thousands of genetic variants and thousands of human traits and diseases. **Figure 1** shows the for the rapid increase in the number of single-nucleotide polymorphisms (SNPs) identified in the NHGRI-EBI (National Human Genome Research Institute–European Bioinformatics Institute) GWAS Catalog (**https://www.ebi.ac.uk/gwas/**) from 2008 to 2019. Following their success in identifying these replicable association signals, researchers have recently expanded their efforts to finely map the implicated chromosomal regions to identify disease-causing variants/genes, dissect the genetic architecture of various traits, and translate these findings to improve disease prevention and treatment strategies. With the generation of whole-exome sequencing (WES) and whole-genome sequencing (WGS) data from up to millions of individuals over the next several years, coupled with detailed health records from these individuals, it is likely that researchers will uncover more association signals for both common variants (the focus of GWAS) and rare variants (through WES and WGS data), as well as how these variants interact together with other risk factors to impact human health. In this review, we focus on statistical methods that have been developed to address the unique challenges in the analysis of GWAS data, most notably the very large number of genetic variants that need to be studied (with the problem of very high dimensionality) and the rich information accumulated from diverse sources about the human genome that can be used in GWAS analysis (for the problem of data integration). We organize our discussion around three scientific goals of GWAS: the identifications of genetic association signals, the inference of genetic architecture of complex traits, and the development of genetic risk prediction models. For each goal, we introduce the biological problems, describe the statistical challenges, and review statistical methods that have been proposed to jointly analyze millions of genetic variants and integrate various data sources to achieve the goals. **Figure 2** provides an overview of a typical GWAS analysis pipeline; we cover some of these methods in the following sections.

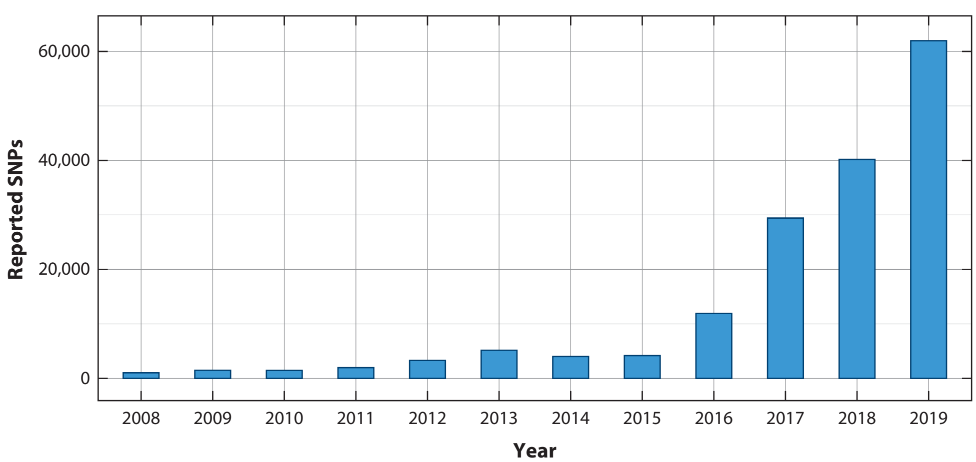

Figure 1

Number of reported significant trait–SNP (single-nucleotide polymorphism) associations in the NHGRI-EBI (National Human Genome Research Institute–European Bioinformatics Institute) GWAS Catalog (**https://www.ebi.ac.uk/gwas/**) from 2008 to 2019.

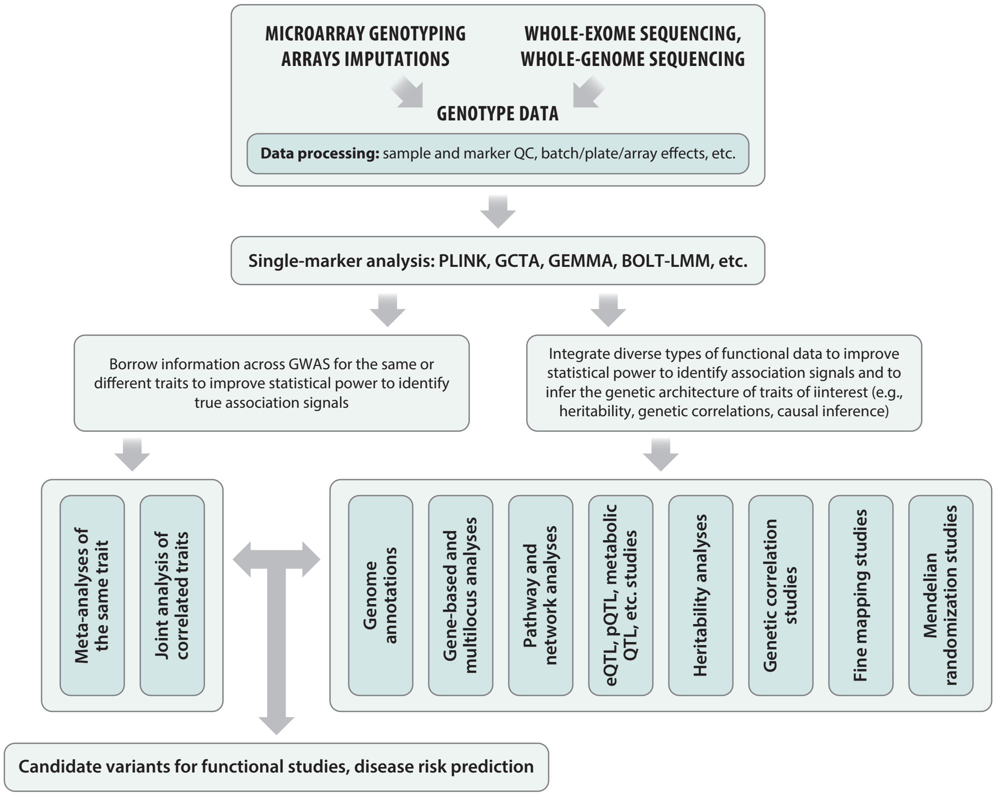

Figure 2

A typical GWAS pipeline for genome-wide association analysis. Abbreviations: eQTL, expression QTL; pQTL, protein QTL; QC, quality control; QTL, quantitative trait loci.

2. IDENTIFICATIONS OF GENETIC ASSOCIATION SIGNALS

In a typical GWAS analysis pipeline, several quality control steps are taken before performing association analysis. These steps help identify and remove low-quality genetic markers (such as those with a high missing genotype rate across the samples, a significant departure from Hardy–Weinberg equilibrium, low minor allele frequency, and low imputation quality) and low-quality samples (such as those with a high missing genotype rate across the markers, chromosomal aberrations, related samples, Mendelian errors, and mismatched sex) (2). Principal component analysis (PCA) is routinely used to infer population structure in the studied samples, and the leading principal components are used as covariates in association analysis to control for population stratification. In the following discussion, we assume that appropriate measures have been taken to minimize false-positive findings due to these potential confounding factors from data quality issues and genetic background heterogeneity across samples.

With appropriate control of confounding factors, the major challenge for detecting true association signals in GWAS analysis is the very large number of genetic variants to be studied. Because GWAS considers millions of markers, researchers use stringent thresholds for statistical significance to control for false positives. For example, if there are one million SNPs studied, a statistical significance threshold of 5×10^{-8} is needed if the goal is to control the overall family-wise false-positive rate at 5% with Bonferroni correction. At this stringent significance level, only

SNPs with large effect sizes can be detected with adequate statistical power, and tens of thousands of samples are likely needed for uncovering SNPs with small to modest effect sizes. There are several statistical tools that are commonly used to analyze GWAS data, including PLINK (3), GCTA (4), GEMMA (5), and BOLT-LMM (6, 7), among others.

Realizing the need for very large sample sizes to detect SNPs truly associated with traits, researchers have formed many large international consortia for common diseases to increase the sample size to meet the rigorous demand for statistical significance. Indeed, as sample sizes for GWAS continue to grow, the number of identified associations has continued to increase, but the effect sizes of newly identified loci are mostly very weak (8). Boyle and colleagues recently summarized these observations using a unified, omnigenic model (9). Another observation that was made early on and has proved to be the case is that most associated SNPs are located in noncoding regions of the human genome (10). Aside from increasing sample sizes, one approach that has received considerable attention to address this challenge is the incorporation of prior knowledge or data sources in the analysis of GWAS data to both identify association signals and interpret the results. Some tools have been developed, e.g., FUMA (11), to help interpret the GWAS results. In the following, we first describe several data sources that can be integrated in GWAS analysis, and we then discuss how these resources have been used to prioritize SNPs/genes for follow-up studies.

2.1. Data Sources that are Informative for GWAS Analysis

In this section, we discuss several data sources that have proved informative for the analysis of GWAS data.

2.1.1. Genome annotations. Recent years have seen major efforts to annotate SNPs, including sequence conservation across species near a SNP, genomic features (e.g., whether a SNP is in a coding region), population genetics characteristics [e.g., a SNP's minor allele frequency and its linkage disequilibrium (LD) with nearby SNPs], epigenetic information, and transcription profiles for genes near the SNP, among others. Large consortia such as the Encyclopedia of DNA Elements (ENCODE) (12), the Roadmap Epigenomics Project (13), and the Genotype-Tissue Expression (GTEx) project (14, 15) have generated vast amounts of transcriptomic and epigenetic data that can be used to infer the functional roles of SNPs. Many computational and statistical frameworks have been developed to synthesize these data into concise and interpretable annotations.

If a coding variant disrupts the function of the protein that this gene encodes, conceivably the variant may be more likely to be functional than a variant that does not affect the protein product. For example, PolyPhen (16) is commonly used to assess whether a candidate missense mutation is damaging, and several methods have been developed to predict whether a coding variant may lead to loss of function (e.g., 17).

Due to advances in both genotyping technology and imputation methods, GWAS analyses now typically include more than 10 million SNPs, most of which are noncoding. The high conservation of a DNA segment in a noncoding region based on sequence alignment across multiple species may suggest a strong purifying selection for the segment in the region, thereby hinting at its functionality. Several methods have been developed to quantify the degree of conservation, and thereby the potential for functionality, such as GERP (18, 19) and phyloP (20). Conserved DNA regions are strongly enriched for heritability of complex diseases (21), suggesting the importance of considering conservation in GWAS analysis. However, only 4.5% of the human genome is conserved across mammals (22), which is much smaller than the percentage of transcriptomic and epigenetic annotations. Because conservation is derived from sequence comparisons across species at the organism level, it does not have tissue specificity and is less likely to be relevant to late-onset diseases such as Alzheimer's disease (23) or human-specific traits such as substance dependence (24).

Recently there have been major efforts to annotate the human genome through epigenetic information, including ENCODE (12, 25) and the Roadmap Epigenomics Project (13). These projects have assayed many epigenetic marks, such as chromatin accessibility, DNA methylation, histone modifications, and transcription factor binding activities for many cell lines and human tissues. It is well known that these epigenetic marks are associated with regulatory activities (12, 26, 27), and they are highly informative for GWAS downstream analyses. Because these epigenetic markers are specific to tissue and cell type, they allow researchers to identify tissues and cell types most relevant to a disease/trait. RegulomeDB (28) provides annotations of variants in noncoding regions that include high-throughput results from ENCODE and other data. ANNOVAR (29), which is often used to assess the functional relevance of candidate SNPs, compiles annotation information from many sources. GWAS hits in noncoding regions are enriched for DNase I hypersensitive sites (30, 31). When three annotation categories including genic and regulatory features, conservation and evolutionary signatures, and chromatin states were jointly considered, Kindt et al. (32) found that SNPs annotated with all three annotations were eight times more likely to be trait associated than those with none of the annotations.

Because supervised machine learning methods have been successful at predicting deleteriousness for SNPs in protein-coding regions, it is natural to apply the same approaches (e.g., random forests and support vector machines) to annotate variants in the noncoding regions for genomic features, transcriptional activities, epigenetic marks, and DNA conservation. However, one major challenge for annotating noncoding SNPs is the lack of gold standard training data. Nevertheless, several methods have been developed, including CADD (33) and GWAVA (34), to annotate noncoding SNPs. For training, CADD compared the variants that are almost fixed in humans with simulated ones and used a support vector machine classifier to distinguish these two classes of variants. In contrast, the training data from GWAVA were defined by treating the regulatory variants in the Human Gene Mutation Database as the positive set, and three sets of negative variants were defined by considering variants with different levels of proximity to those in the positive set. With a pair of positive and negative sets of variants, a modified random forest algorithm classifier was used for prediction.

To address the issue of insufficient and potentially biased training data, researchers developed DeltaSVM (35) and DeepSea (36) to predict regulatory activities using short DNA segments as predictive features. This was made possible because of the rich data collected from ENCODE. Although these methods do not rely on labeled training data, predefined regulatory activities in the genome are required. Prediction results based on different epigenetic marks (e.g., binding sites of different transcription factors) could be substantially different (36).

In addition to supervised methods, several unsupervised methods have been proposed to address the potential biases in labeled training data to annotate the SNPs in the noncoding regions. These unsupervised methods simultaneously consider multiple data sources (features) and cluster SNPs into different categories based on patterns learned from these features. Several methods were developed along this line using epigenetic marks (e.g., histone modifications), including ChromHMM (37, 38) and Segway (39). ChromHMM used a hidden Markov model, whereas Segway adopted a dynamic Bayesian network model. Both ChromHMM and Segway inferred different chromatin states from joint analysis of genomic features (e.g., transcription and heterochromatin) using ENCODE data (40). One challenge for the use of annotation results of ChromHMM and Segway is that they may infer different numbers of chromatin states depending on the training data. Moreover, it is not easy to interpret some of the inferred states due to a lack of understanding of the features associated with these states.

In comparison, only two latent states were considered by some methods to improve robustness and interpretability of the annotation results. For example, GenoCanyon (41) jointly modeled conservation and epigenetic information through a naïve Bayes model. It used the expectation-maximization algorithm in model fitting and provided the posterior probability for the functionality of each nucleotide in the genome. In contrast, Eigen (42) explicitly considered correlations among the epigenetic marks, conditioning on the (unknown) functional state, and used a spectral method (43) to derive the annotations. Because Eigen included variant-specific information such as allele frequencies, its annotation is at the SNP rather than nucleotide level.

The methods discussed above can provide annotations for the whole organism without reference to a specific tissue or cell type. However, it may be more informative to have tissue- and cell type–specific annotations, both because of the availability of tissue- and cell type–specific transcriptomic and epigenetic data and because of the importance of tissue and cell type context for a trait and disease of interest. Because there is relatively little information on tissue-specific functional and nonfunctional variants in noncoding regions, unsupervised learning methods are more useful. As the methods discussed above are generic, the annotation results obtained using these methods will be tissue specific if tissue-specific data are used as input. In fact, when ChromHMM and Segway were introduced, they were trained using cell line–specific data from ENCODE. As for GenoCanyon and Eigen, both methods have been extended for tissue- and cell type–specific annotations. GenoSkyline (44) was proposed for tissue-specific annotations using the same framework as GenoCanyon, but it uses tissue-specific data. It was further extended to GenoSkylinePlus by providing cell type–specific annotations for each of the 127 tissue and cell types in the Roadmap Epigenomics Project (45). Similarly, FUN-LDA (46) provides tissue-specific annotations.

In addition to tissue- and cell type–specific annotations, disease-specific information has also been integrated to improve the specificity of annotations by, e.g., Phevor (47) and Phen-Gen (48), which incorporated ontology information for annotations involving protein-coding variants. As for noncoding variants, DIVAN (49) and PINES (50) incorporated SNP–disease association information to develop disease-specific annotations. However, one challenge in using these annotation results for post-GWAS analyses is that GWAS results were already used in annotating the variants.

2.1.2. Gene and pathway information. After covering annotations for individual SNPs/nucleotides, in this section we discuss the annotations for genes and pathways. Much is known about the coding regions in the genome and the SNPs can be assigned to genes based on their genomic coordinates. It is conceivable that multiple SNPs at a disease-associated gene may be functional, and joint analysis of all the SNPs at or near this gene may be more powerful than analyzing individual SNPs, especially for those with low allele frequencies. Furthermore, multiple genes in the same biological pathway may be involved in disease etiology, and joint analysis of all pathway genes may better identify disease-associated pathways, especially when individual genes only exert modest effects on disease onset. There are rich resources in the public domain for pathway annotations, including KEGG (Kyoto Encyclopedia of Genes and Genomes; **http://www.genome.jp/kegg/**), Reactome (**www.reactome.org/**), and BioCarta (**http://www.biocarta.com/**). Although genes in a pathway may be simply collected as a gene list, their detailed relationships as shown in many pathway databases may offer additional information to identify disease-associated genes and pathways. For example, genes close to each other in a pathway may have more similar biological functions, and therefore association signals. In addition to annotated pathways from different databases, we can construct coexpression networks from gene expression profiles collected from many individuals, e.g., those with or without a disease. The network structure may be informative for GWAS analysis. In addition, network modules within the overall coexpression network, i.e., a group of genes with very similar expression profiles, may be enriched with genes associated with a disease/trait (e.g., Reference 51 for autism). Moreover, instead of considering one overall network based on expression data from all the samples, we may construct separate networks (e.g., one network from diseased individuals and another network from normal individuals) and consider the differential (rewired) network to better identify GWAS signals (52).

2.1.3. Protein interaction. Protein interaction data are another useful annotation source for relationships among genes. Protein–protein interactions (PPIs) are essential for many biological functions. There are several databases for PPIs, such as BIND (Biomolecular Interaction Network Database; 53), HPRD (Human Protein Reference Database; 54), and BioGRID (Biological General Repository for Interaction Datasets; 55). The information in these databases is curated from the literature, high-throughput experiments, and computational predictions. A PPI network thus obtained may be treated similar to a gene coexpression network, and network modules (or subnetworks) may be identified that are enriched for GWAS association signals.

2.1.4. Expression quantitative trait loci data. Gene expression profiles can be used to annotate activity in the genome. Many large consortia, such as GTEx (14, 15), CommonMind (56), and STARNET (Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task; 57), have generated gene expression data for different tissues and cell lines together with genotype information. Among these consortia, GTEx provides comprehensive gene expression data, genotype data, and other data for dozens of tissue types from hundreds of donors (14, 15). GTEx data have been extensively analyzed to study the regulatory effects of SNPs, i.e., expression quantitative trait loci (eQTL), on gene expression across human tissues, and the results have proved informative for GWAS analysis. For example, regions that are transcribed in disease-related tissues are more likely to harbor risk variants for that disease (58). Nicolae et al. (59) found that GWAS hits across different diseases are 1.5 times as likely to be eQTLs. Mehta et al. (60) showed that the eQTLs identified from whole blood are highly robust and reproducible across studies.

2.2. Statistical Methods to Incorporate Different Data Types to Detect Association Signals

2.2.1. Incorporating single-nucleotide polymorphism annotations. As discussed above, several computational tools, e.g., GenoCanyon, annotate a SNP to be either functional or nonfunctional through joint analysis of multiple data types. In addition, results from prior studies may also be used to group SNPs into those likely involved in a disease and those less likely involved. In this case, different weights may be assigned to SNPs with different annotations to improve the statistical power to identify SNPs truly associated the disease. For example, if two SNPs have the same level of statistical significance from a GWAS, the SNP annotated to be functional should be prioritized over the SNP that is annotated to be nonfunctional. This idea was first proposed by Roeder et al. (61) to incorporate prior linkage results to weigh different SNPs, and the authors showed that this method can effectively improve statistical power while appropriately controlling the overall false-positive rate. This general approach can be adopted with other annotations of SNPs, such as coding versus noncoding regions or functional regions from annotation tools (62). The improvement in signal detection through this approach was studied by Hou et al. (31). Other strategies can also be adopted, such as prioritized subset analysis proposed by Lin & Lee (63).

With multiple annotations available, a simple scoring scheme was proposed by Saccone et al. (64) to rank a list of candidate SNPs based on a weighted linear score from the SNP annotations, where each annotation category was assigned a corresponding weight. In a regression setting, Chen & Witte (65) proposed a hierarchical model to analyze GWAS data, using a regression model to estimate the effect size of each SNP in the first stage, and then regressing the effect size estimates against a set of annotations for the markers in the second stage. Heron et al. (66) considered a similar model for binary traits. Lewinger et al. (67) proposed an empirical Bayes method in a similar effort to incorporate prior information. For this method, it was assumed that the test statistics follow a two-component mixture distribution, with one component corresponding to the disease-associated markers and the other component corresponding to the non-disease-associated markers. Functional annotation information was used to quantify the likelihood that a marker is disease associated. Fridley et al. (68, 69) developed a full Bayesian model along this line. More recently, a latent sparse mixed model to integrate functional annotations with GWAS data was proposed that is scalable to millions of SNPs and hundreds of functional annotations through an efficient variational expectation-maximization algorithm (70).

2.2.2. Gene-based association analysis. Gene-based analysis has been advocated since the early days of GWAS to borrow information from all the SNPs at or close to a gene in order to investigate whether there is an overall association signal between the SNP set and the disease/trait of interest. This is motivated by the hypothesis that more than one SNP may be disease associated and joint analysis will improve statistical power by simultaneously considering all the SNPs. A naïve approach proposed by Wang & Bucan (71) first analyzed all the SNPs for a gene individually and used the SNP with the most significant result to represent the gene-level association signal for this gene. It is clear that this method will favor genes with more SNPs because, even in the absence of association with a disease, genes with more SNPs will likely have more statistically significant findings simply by chance. Although it is natural to apply a multivariate regression to study the relationship between a disease/trait and a set of SNPs, where the response variable is the disease status or trait value and the independent variables are all the SNPs in this set, the strong dependency (i.e., LD) among the SNPs may present statistical and computational challenges for regression models. One approach to addressing this issue is through regressing the disease status or trait value on the principal components of the SNP genotypes. This approach first calculates the principal components of the genotype matrix, where each row corresponds to a study subject and each column corresponds to a SNP, and the entry in this matrix is the coded genotype score, e.g., the number of minor alleles, for the corresponding individual and SNP. After the calculations of the principal components, a (logistic) regression model is used to study the association between disease outcome and the leading principal components. In a systematic comparison of different gene-based association methods, Ballard et al. (72) found that the PCA-based analysis had an overall good performance compared to other methods. This method was implemented in MAGMA (73). Jointly considering both phenotype and genotype information in dimensional reduction, Chun et al. (74) proposed a sparse partial least-squares method for gene-based analysis to further improve statistical power. When there are many (hundreds of) markers in a gene, the methods considered by Ballard et al. (72) may not be optimal. Kernel-based methods such as SKAT have proved useful in this context (75, 76).

While the above methods need individual genotype data to perform gene-based association analysis, some methods have been proposed to use summary statistics from individual SNPs to derive gene-based tests, including VEGAS (77) and GATES (78), which only require marginal *p*-values without individual genotype data. To improve computational efficiency, Bacanu (79) proposed a two-stage procedure that first identifies interesting regions and then performs more refined gene-based analysis in these interesting regions.

2.2.3. Pathway-based (gene set) analysis. To demonstrate the benefit of incorporating pathway information in GWAS analysis, Dinu et al. (80) aimed to identify additional genes associated with age-related macular degeneration using the GWAS data from Klein et al. (1). This analysis focused on the complement pathway because this pathway was implicated in the genome-wide significant finding. Although none of the SNPs in this pathway passed statistical significance individually, multiple SNPs in this pathway showed marginal association signals and there was an overall statistically significant enrichment of marginal signals for this pathway. Moreover, there was good correspondence between marginal signals and prior linkage results. Ballard et al. (81) focused on GWAS results for Crohn's disease and showed that many pathways were enriched for genes with marginal association signals, although most of these genes did not pass statistical significance on their own. Similarly, pathway-based analysis of height GWAS results implicated several biological pathways (82) related to height. As for statistical methods to identify disease/trait-associated pathways, K. Wang et al. (71) adapted the gene set enrichment analysis method that was originally developed to analyze gene expression data for the analysis of GWAS data. A hierarchical Bayesian model was proposed by Shahbaba et al. (83) to aggregate information from multiple genes in a pathway. A similar approach developed by L. Wang et al. (84) and Zhang et al. (85) studied how domains are associated with different diseases. More recent developments on gene set- or pathway-based analysis include MAGMA (73) and a method based on the generalized Berk–Jones statistic (86).

2.2.4. Topology-based analysis. Although the methods discussed in the previous section are effective at integrating information from multiple genes in a pathway to identify disease-associated pathways, they do not utilize detailed relationships, also called topological information, among genes in a pathway. Several methods have been proposed to better utilize topological information. Erten et al. (87) discussed three types of topological information that can be used, including network connectivity, information flow, and topological similarity. According to the network connectivity principle, genes with more connections to genes known to be disease associated should be prioritized. To recover potential information loss due to indirect connections, the information flow principle connects candidate genes to the seed genes that are disease associated. The topological similarity principle is built on the idea that if a gene interacts with a group of genes with similar functions, then this gene may also have similar functions.

There are many established databases for PPIs. We can represent a PPI network by an unweighted, undirected graph with each node denoting a protein (gene) and each edge denoting the interaction between two proteins (genes). If some genes in this network are disease associated, we can assess the importance of a candidate gene based on its topographical relationships with disease-associated genes in the network. This idea was realized by J. Chen et al. (88) by adopting three algorithms proposed by White & Smyth (89) to rank web pages. The ToppGene Suite (90) used the *k*-step Markov method, one of White & Smyth's three algorithms, to prioritize candidate genes using PPI information. After comparing several gene prioritization methods using PPI information, Navlaka & Kingsford (91) concluded that the combination of some of these methods may lead to improved performance. Further developments were carried out by Guney & Oliva (92), who considered the node (gene) properties in analysis, and by Jia et al. (93), who considered dense modules enriched for GWAS signals when there is no known disease-associated genes.

Similar to pathway-based analysis utilizing topology information, M. Chen et al. (94) found that genes with marginal evidence of association were also likely to be neighbors. Based on this observation, they proposed a Markov random field (MRF) model incorporating topological information in order to better identify disease-associated genes by jointly modeling network information and GWAS results (94). They showed that more disease-associated genes can be identified using this MRF modeling approach. This approach has been extended by Hou et al. (52) to jointly model gene coexpression networks where the focus was the rewired network, i.e., the differences between the network constructed from the diseased individuals and that constructed from the healthy controls.

2.2.5. Expression quantitative trait loci. eQTL naturally connects SNPs, genes, and traits. PrediXcan (95) and FUSION (96) were recently introduced to incorporate eQTL results into GWAS analysis. For these methods, a gene expression level prediction model (commonly called an imputation model in this context) is first trained from an eQTL dataset (e.g., GTEx) to predict gene expression in a given tissue (e.g., whole blood) using genotype data. Most of these methods only consider nearby SNPs (i.e., *cis*-SNPs) to reduce the model search space. After the imputation model is developed for each gene in a given tissue, the expression level for an individual in a given tissue can be imputed based on his or her genotype data. Then we can assess whether there is an association between the imputed gene expression levels and observed trait values. Genes with statistically significant results will be considered trait associated. Although the initial methods were developed for individual genotype data, it was later shown that this general approach can also be applied to summary statistics data, where the test statistic can be approximated using the parameters in the imputation model, GWAS summary statistics, and a reference genotype panel (96, 97). Therefore, gene-level association tests can be performed without having access to individual-level genotype and phenotype data. These imputation-based gene-level association methods have gained much popularity in recent years and been applied to many diseases and traits (98). Compared to gene-level association tests discussed above in Section 2.2.2 that do not use eQTL information, PrediXcan and FUSION can better leverage eQTL data by assigning more informed weights to SNPs than those purely based on LD information.

Several methods have been proposed to improve the performance of the above methods (e.g., 99). To improve prediction accuracy, Nagpal et al. used a nonparametric Bayesian method that assumes a data-driven nonparametric prior for *cis*-eQTL effect sizes (100). However, there are several limitations to these methods. First, the sample sizes for some tissues, such as brain tissues, are more limited than others, leading to difficulty in downstream analysis and interpretations. Second, the relevant tissue may not even be in the training dataset; therefore, cross-tissue analysis is needed to best utilize the correlation among tissues and increase statistical power. Third, when single-tissue analysis is performed, it is nontrivial to combine results for an integrative organism-level analysis. Recently, a cross-tissue transcriptome-wise association study framework named UTMOST was introduced to address these limitations (101). The method first imputes gene expression levels in multiple tissues through a penalized regression model, and then performs a joint tissue association test by combining single-tissue association statistics through the generalized Berk–Jones test. Previous work on multitissue analysis focused on inferring eQTLs (102) instead of gene expression imputation. There has also been work to expand these methods to incorporate other types of annotations such as splicing quantitative trait loci (103). Finally, fine-mapping methods have been proposed to identify biologically relevant genes among coregulated gene candidates (104).

2.2.6. Joint GWAS analysis from multiple traits. Pleiotropy is the phenomenon that the same variant/gene may have an effect on more than one trait. Pleiotropy has been found to be widespread in human genetics. Even before the GWAS era, phenotype similarities were used to cluster diseases and identify genes for a query disease (105). Many genes have been found to be associated with multiple autoimmune disorders (106). Interested readers are referred to References 107 and 108 for more general reviews. To leverage shared genetics across different phenotypes, Zhou & Stephens (109) introduced a multivariate linear mixed effects model when individual-level genotype data and multivariate phenotypes are available. A comprehensive analysis was provided by Stephens (110). When only marginal SNP association significance levels are available from GWAS for different phenotypes, a statistical framework, called GPA (genetic analysis incorporating pleiotropy and annotation), was proposed to borrow information across different GWAS to improve association signal detections (111). When analyzing results from two traits, GPA assumes that each SNP has four possible association statuses, (0,0), (0,1), (1,0), and (1,1), where (0,0) means that the SNP is not associated with either trait, (0,1) means that the SNP is associated with the second trait but not the first trait, etc. For each association status, GPA further assumes that the observed statistical significance follows a specific distribution. Then GWAS summary statistics from both studies are jointly analyzed to infer the joint association status of each SNP with the two traits. More recently, MTAG was introduced to jointly analyze summary statistics from multiple traits to improve statistical power (112).

3. DISSECTION THE GENETIC ARCHITECTURE OF COMPLEX TRAITS

In the previous section, our focus was on the identifications of SNPs/genes associated with traits of interest. Although most GWAS papers only report statistically significant SNPs, there is often an enrichment of SNPs with small *p*-values, although many do not pass genome-wide statistical significance. It is likely that many of these marginally significant SNPs are disease/trait associated. Therefore, there is a need to understand the genetic association signal information across all the SNPs, not just those that are significant. In a seminal GWAS paper for schizophrenia and bipolar disorder, Purcell and colleagues could not identify any marker passing genome-wide statistical significance (113). However, they showed that the disease risk score derived from a subset of the data was correlated with the disease status in an independent set, suggesting the presence of association signals in the data. The authors performed extensive simulations under various disease models and showed that when there are hundreds of SNPs having weak associations with disease, it is likely there may not be any genome-significant findings but we can still derive significant risk prediction scores. This paper represents the first attempt to use results from genome-wide SNPs to infer the underlying genetic model for diseases. There are many aspects of the genetic architecture of complex traits, including the number of genes/variants affecting a trait, their effect sizes, allele frequencies, relevant tissues and cell types, and correlations among different traits, among others. In the following, we focus on three aspects of the genetic architecture that have received great attention in the GWAS literature, including heritability, tissue/cell type specificity, and genetic correlation between traits.

3.1. Heritability

Although many genetic variants have been identified for many traits, in combination they only account for a small proportion of phenotype variation. For example, with more than 700,000 people and over 3,200 near-independent association signals for height, SNPs only account for 24.5% of the variation in height (114), which differs substantially from the estimated 80% heritability for height. The so-called "missing heritability" in the literature refers to the observation that association signals from early GWAS results often only explain a small proportion of overall heritability (115). In a seminal paper (116), Yang et al. showed that although the trait prediction model developed from significant (or top) SNPs may only be modestly predictive of the trait in the general population, SNPs on the genotyping platforms can in fact account for a significant proportion of the trait variation. The authors reached this conclusion by adopting the random effects model to connect phenotypes with genotypes. Under the random effects modeling framework, the genotypes at each SNP are first standardized to have mean 0 and variance 1, and the effects of all the SNPs are assumed to follow a normal distribution with mean 0 and variance h^2/m, where m is the total number of SNPs. Hence the overall genetic contribution of the SNPs is h^2. This is more formally defined as

$$Y = X\beta + \varepsilon,$$
$$\beta \sim N\left(0, \frac{h^2}{m} I\right),$$
$$\varepsilon \sim N(0, (1 - h^2)I),$$

where Y are the standardized trait values, X is the standardized genotype matrix, β are the genetic effects, h^2 is the overall heritability, and m is the number of SNPs. The residual maximum likelihood (REML) estimator can be used to infer the model parameters when individual genotype and phenotype data are available. The heritability estimated with this approach is often called chip-based heritability, as it does not account for variants that cannot be captured by the SNPs on the genotyping platform. One potential issue with this model is that it is unlikely that all the SNPs will contribute to the observed trait; therefore, the model is likely mis-specified. A more realistic model may be that only a small proportion of the SNPs are associated with traits. In this case, the genetic model would assume that the effect size distribution is a mixture of a point mass of 0 for most SNPs and a normal distribution for those SNPs having some effects. Jiang et al. (117) showed that the traditional REML estimator has excellent robustness when the model is mis-specified, justifying its application to GWAS data analysis.

For broader applications to GWAS, the REML estimate has two limitations even with its good robustness: It requires individual-level data, which may not be accessible for the general research community due to privacy and other concerns, and the computational complexity makes it difficult to handle biobank-level data with hundreds of thousands of individuals. However, summary statistics from GWAS results are more readily available (118) and may be used to infer chip-based heritability.

The LD score regression method was first introduced as a method to distinguish polygenicity from unadjusted confounding in GWAS analysis (119). LD score regression only requires GWAS summary statistics and externally estimated LD as inputs. It is based on the same random effects model defined above. When there is no unadjusted confounding in the model, that the following can be shown:

$$E\left(z_j^2\right) = \frac{nh^2}{m} l_j + 1,$$

where z_j is the z-score of the j-th SNP, n is the sample size, and the LD score for the j-th SNP, l_j, is defined as the sum of LD scores between the j-th SNP and all other SNPs:

$$l_j = \sum_{k=1}^{m} r_{jk}^2.$$

When z_j^2 is regressed on LD scores l_j, the weighted least-squares estimator for the regression coefficient can be used to estimate heritability, h^2.

The LD score regression method has become one of the most commonly used approaches for estimating heritability because it only needs summary statistics and LD information from a reference panel as inputs. Other methods proposed for estimating heritability using summary statistics may be statistically more efficient (120), and there is also concern about confounding

(121). However, the original LD score regression method is still the predominantly used method in the literature.

One limitation of the random effects model discussed so far is that all the SNPs are assumed to have the same effect distributions. Balding and colleagues have proposed an alternative LDAK model where the trait contribution from a SNP is a function of a number of factors, including its allele frequency and associations with nearby markers (122–124). Based on their model specifications, they showed that the estimated heritability is higher for most traits than the standard random effects model.

3.2. Tissue and Cell Type Specificity

Instead of using the heritability estimation approaches discussed above with all the SNPs in the human genome, we can focus on a selected set of SNPs, such as those on a specific chromosome or with minor allele frequency below a threshold, e.g., 1%, to ask how much phenotype variation can be explained by markers on a specific chromosome or by less common variants. With this idea, it was shown that each chromosome's contribution to the overall height heritability is proportional to its length, suggesting a polygenic genetic architecture for height (125). Further, it was also shown that heritability for height and body mass index is enriched in variants with lower minor allele frequencies, hinting at selection effects (126).

When only summary statistics are available, the LD score regression method can be extended to estimate annotation-dependent heritability (21). With K functional annotations, the original random effects model can be generalized as

$$
\begin{aligned}
Y &= \sum_{i=1}^{K} X_i \beta_i + \varepsilon, \\
\beta_i &\sim N\left(0, \frac{h_i^2}{m_i} I\right), \\
\varepsilon &\sim N\left(0, \left(1 - \sum_{i=1}^{K} h_i^2\right) I\right),
\end{aligned}
$$

where X_i is the genotype matrix for m_i SNPs having the i-th functional annotation, and h_i^2 is the proportion of phenotypic variance explained by SNPs with the i-th annotation. Under this model, we have

$$
E\left(z_j^2\right) = \sum_{i=1}^{K} \frac{n h_i^2}{m_i} l_j^{(i)} + 1,
$$

where $l_j^{(i)}$ is the annotation-stratified LD score, defined as

$$
l_j^{(i)} = \sum_{k \in A_i} r_{jk}^2.
$$

If we annotate the human genome based on tissue- and cell type–specific information, we can ask which tissue or cell type is relevant for the trait of interest by performing heritability enrichment analysis based on LD score regression. This can be accomplished by comparing the proportion of heritability explained by the SNPs in a functional annotation with the proportion of the genome covered by these SNPs. Such analysis has become routine in GWAS because, when tissue-specific functional annotations are used, such enrichment analysis can identify disease-related tissue and cell types (21, 44, 45, 58). For example, in the analysis of GWAS results from Alzheimer's and the Parkinson's diseases, among 65 tissue and cell types, SNPs that are potentially functional in CD14$^+$ monocytes explain most of the heritability for both diseases, suggesting that innate immunity is involved in neurodegeneration and that these two diseases share genetics through a common neuroinflammation pathway (45).

3.3. Genetic Correlation Between Traits

When individual-level data are available, the REML estimator can be used to study shared genetics across multiple complex traits (127, 128). When only summary statistics are available, LD score regression can be extended in a similar way (129) to study genetic correlation. The cross-trait LD score regression is based on the following model:

$$Y_1 = X\beta + \varepsilon, \qquad \beta \sim N\left(0, \frac{h_1^2}{m}I\right), \qquad \varepsilon \sim N\big(0, (1 - h_1^2)\, I\big),$$

$$Y_2 = Z\gamma + \delta, \qquad \gamma \sim N\left(0, \frac{h_2^2}{m}I\right), \qquad \delta \sim N\big(0, (1 - h_2^2)\, I\big),$$

where Y_1 and Y_2 are the standardized phenotypes for the two traits with respective heritabilities h_1^2 and h_2^2, and β and γ are the respective effect sizes for the m SNPs defined in standardized genotype matrices X and Z. The effects on two different traits are assumed to be correlated:

$$E(\beta\gamma^T) = \frac{\rho_g}{m}I,$$

where ρ_g is the genetic covariance between traits Y_1 and Y_2. In the simple case where there are no shared samples between the two GWAS for these two traits, we have

$$E\Big((z_1)_j(z_2)_j\Big) = \frac{\sqrt{n_1 n_2}\rho_g}{m} l_j,$$

where n_1 and n_2 are the sample sizes for the two traits, respectively. Similar to single-trait analysis, regression coefficients can be used to estimate genetic covariance, or a closely related but more interpretable metric, genetic correlation:

$$\text{corr} = \frac{\rho_g}{h_1 h_2}.$$

The results can be generalized to the case where there are sample overlaps between the two GWAS. Under the same modeling framework, GNOVA (130) used an estimator based on the method of moments to estimate genetic covariance. Moreover, GNOVA can estimate annotation-stratified genetic covariance through the following model

$$Y_1 = \sum_{i=1}^{K} X_i\beta_i + \varepsilon,$$

$$Y_2 = \sum_{i=1}^{K} Z_i\gamma_i + \delta,$$

$$E(\beta_i\gamma_i^{\mathrm{T}}) = \frac{\rho_i}{m_i}I,$$

when there are K functional annotations, and the parameters ρ_i ($i = 1, \ldots, K$) quantify the genetic covariance components for each functional annotation.

Summary statistics–based methods are now commonly used to infer genetic correlations among many complex diseases and traits (131, 132). Several methods have been developed to estimate local genetic correlation for specific risk loci (133) and trans-ethnic genetic correlation (134), as well as to improve genetic correlation estimates for case-control studies (135). Moreover, online servers have been developed for researchers to estimate genetic correlations between their data and hundreds of traits with publicly accessible summary statistics (136).

4. GENETIC RISK PREDICTION

Achieving accurate disease risk prediction using genetic information is a major goal in human genetics research and precision medicine. Accurate prediction models can improve disease prevention and early treatment strategies (137). Consider a simple additive model for quantitative traits. There are two key aspects of a good prediction model: the selections of SNPs for prediction and the estimated effect sizes for these SNPs. One naïve approach would be to only include statistically significant SNPs in the model and use the marginal effect size estimate. However, considering only statistically significant SNPs may lose out on potentially useful information in the other markers, and there is also the potential issue of winner's curse for those significant SNPs if their marginal effect sizes are used in prediction. To improve upon this naïve approach, researchers have adopted various approaches that utilize genome-wide data, including those based on individual genotype data and those based on summary statistics.

Methods proposed for when individual genotype and phenotype data are available include those based on machine learning (138), Bayesian sparse linear mixed models (139), an improved best linear unbiased prediction method (140), and a kernel machine method (141). Methods have also been proposed to leverage information from genetic correlations among traits (138–143). Methods proposed for when only summary statistics are available include those based on pruning and thresholding, Bayesian priors that incorporate LD information (144), empirical Bayes estimates (145), penalized regression (146), and Bayesian regression with continuous shrinkage priors (147). Despite the potential information loss in summary data, summary statistics–based approaches have been widely adopted since summary statistics for large-scale association studies are often easily accessible. However, prediction accuracies for most complex diseases remain moderate, which is largely due to the challenges in both identifying all the functionally relevant variants and accurately estimating their effect sizes in the presence of LD (148).

AnnoPred (149) is a principled framework to integrate functional annotation information to improve polygenic risk prediction accuracy. A key idea in the AnnoPred framework is to utilize functional annotation information to accurately estimate SNPs' effect sizes. Using this framework, researchers can estimate the enrichment for GWAS associations in a prespecified list of functional annotations and acquire an empirically estimated informative prior of SNPs' effect sizes based on annotation assignment and signal enrichment. In general, SNPs located in annotation categories that are highly enriched for GWAS signals receive a higher effect size prior. Through its applications to real GWAS data, it was shown that AnnoPred can effectively incorporate annotation information to improve risk prediction (149). PleioPred is a method that can incorporate GWAS summary statistics from multiple traits to further improve risk prediction accuracy (150). It was shown that the improvement in the risk prediction accuracy is proportional to the genetic correlation between traits. More recently, wt-SBLUP (151) was introduced to jointly analyze summary statistics from multiple traits to improve risk prediction.

Table 1 Common terms and approaches used in genome-wide association studies and associated tools that have been developed

Term/approach	Objective	Selected resources and tools
Genome annotations	Prioritize genes, regions, and variants based on their potential functions	Protein-coding region: PolyPhen Sequence conservation: GERP, phyloP Supervised methods: CADD, GWAVA, DeltaSVM, DeepSea Unsupervised methods: ChromHM, Segway, GenoCanyon, Eigen Tissue-specific annotations: GenoSkyline, FUN-LDA Integration of disease information: Phevor, Phen-Gen, PINES Integrated resources: RegulomeDB, ANNOVAR
Meta-analysis	Integrate results from different studies	Stata, METAL
Pleiotropic analysis	Integrate information from genetically correlated traits	GPA, MTAG
Multilocus or gene-based analysis	Integrate information from different markers	PCA, MAGMA, SKAT, VEGAS, GATES
Pathway- and network-based analysis	Integrate information from multiple related genes	MAGMA, ToppGene, MRF-based analysis
Fine mapping	Localize functional genes and variants	PAINTOR, CAVIAR, FINEMAP
TWAS	Integrate information from eQTL studies	PrediXcan, FUSION, UTMOST
Heritability	Estimate the overall genetic contribution from a set of markers to a trait	GCTA, LDSC, GNOVA
Genetic correlation	Estimate the genetic correlations between traits for a set of markers	GCTA, LDSC, GNOVA
Polygenic risk score	Predict disease risk based on genetic information	P+T, LDPred, AnnoPred, PleioPred, PRS-CS
Mendelian randomization	Infer the causal relationship among traits	IVW, MR-Egger

Abbreviations: eQTL, expression quantitative trait loci; IVW, inverse variance weighted; MRF, Markov random field; P+T, pruning and thresholding; PCA, principal component analysis; PRS-CS, polygenic risk score with continuous shrinkage prior; TWAS, transcriptome-wide association studies.

Table 1 summarizes the statistical methods discussed in Sections 3 and 4. Some of the corresponding data resources are listed in **Table 2**, including results from both GWAS and functional genomics studies.

5. DISCUSSION

This review has focused on the analysis of GWAS results where most of the SNPs are common variants. With the decrease in sequencing cost and investment from many funding agencies, WES and WGS data will become more common. Compared to GWAS data, WES and WGS data offer us the opportunity to study the effects of rare variants. However, due to their low to very low allele frequencies and relatively smaller sample size, it is more difficult to estimate their effect sizes. Although statistical methods are being developed to jointly analyze both common and rare variants to infer chromosome loci associated with traits, we may benefit from extending these methods to incorporate accurate and informative annotations. An interesting question to study when both common and rare variants are available would be how these two types of variants interact to affect disease outcome (152).

Table 2 Data resources for genome-wide association studies

Resource	Available data types	URL
dbGaP (Database of Genotypes and Phenotypes)	Genetic data: SNP, CNV, WES, WGS Expression data: microarray, RNA-seq Epigenomic data: ChIP-seq, ATAC-seq, etc. Demographic data, clinical data, exposure information	**https://www.ncbi.nlm.nih.gov/gap/**
UK Biobank	Genetic data: SNP, CNV, WES Demographic data, clinical data, exposure information, wearable device data, imaging data	**http://biobank.ctsu.ox.ac.uk/showcase/index.cgi**
GTEx (Genotype-Tissue Expression) Project	Genetic data: SNP, CNV, WES, WGS Expression data: RNA-seq Demographic information	**https://www.gtexportal.org/home/**
GEO (Gene Expression Omnibus)	Expression data: microarray, RNA-seq Sample information	**https://www.ncbi.nlm.nih.gov/geo/**
ENCODE (Encyclopedia of DNA Elements)	RNA-seq, ChIP-seq, DNase-seq, ATAC-seq, methylation, RIP-seq, ChIA-PET, 5C, Hi-C	**https://www.encodeproject.org/**
Roadmap Epigenomics Project	RNA-seq, ChIp-seq, DNase-seq, methylation	**https://www.roadmapepigenomics.org/**

Abbreviations: 5C, chromosome conformation capture carbon copy; ATAC-seq, assay for transposase-accessible chromatin using sequencing; ChIA-PET, chromatin interaction analysis by paired-end tag sequencing; ChIP-seq, chromatin immunoprecipitation and sequencing; CNV, copy number variation; DNase-seq, DNase I–hypersensitive sites sequencing; Hi-C, high-throughput chromosome conformation capture; RIP-seq, RNA immunoprecipitation and sequencing; RNA-seq, RNA sequencing; SNP, single-nucleotide polymorphism; WES, whole-exome sequencing; WGS, whole-genome sequencing.

One emerging technology that will prove informative for GWAS analysis is single-cell data. Although transcriptome analysis has been well integrated in GWAS analysis (e.g., eQTL discussed above), having single-cell transcriptome data may open new doors to study the genetic effects of variants, such as cell type compositions, cell heterogeneity, and others. We also note that other types of data, e.g., ATAC-seq (153) and Hi-C (154), are being generated and curated, and such data will undoubtedly offer useful information for the analysis and interpretation of GWAS results.

With an ever-increasing sample size and better phenotyping from large-scale GWAS, such as UK Biobank (155), the All of Us Research Program (**https://allofus.nih.gov**; 156), the Million Veteran Program (157), the BioBank Japan Project (158), and many studies at the dbGaP (Database of Genotypes and Phenotypes; **https://www.ncbi.nlm.nih.gov/gap/**), a large number of phenotypes together with individual genotypes are available from millions of samples across these studies. There is no doubt that powerful statistical methods implemented in computationally efficient tools will lead to the novel identification of tens of thousands of associations between genetic variants and traits. Moreover, the availability of data from electronic medical records together with environment exposure data will present great opportunities and challenges for statisticians. With many association signals at hand, the future focus of GWAS analysis will be on understanding the molecular mechanisms of disease onset, the causal relationships among different traits, and how genetics can inform disease heterogeneity. For example, GWAS results may point to the development of new therapeutic strategies or repositioning of approved drugs for new indicators, as demonstrated by Sham and colleagues (159). Another area that has seen very active research is the inference of causal relationships among different traits, as well as gene expression traits, through various versions of Mendelian randomization methods (160, 161). Moreover, although we have discussed how GWAS results can help identify high-risk individuals for more effective prevention and treatment, risk prediction models to date do not take an individual's health records into account, and it would be much more informative if such information were fully integrated in disease risk prediction via access to electronic health records. In addition, with better understanding of the functional roles of both common and rare variants, we will be able to truly benefit from the rich information in WES and WGS data. It is critical to develop more comprehensive and robust statistical methods to accomplish these goals.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was supported in part by NIH (National Institutes of Health) grant R01 GM122078 and NSF (National Science Foundation) grants DMS 1713120 and DMS 1902903.

LITERATURE CITED

1. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* 308:385–89
2. Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, et al. 2010. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol.* 34:591–602
3. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81:559–75
4. Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88:76–82
5. Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44:821–24
6. Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, et al. 2015. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* 47:284–90
7. Loh PR, Kichaev G, Gazal S, Schoech AP, Price AL. 2018. Mixed-model association for biobank-scale datasets. *Nat. Genet.* 50:906–8
8. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, et al. 2017. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101:5–22
9. Boyle EA, Li YI, Pritchard JK. 2017. An expanded view of complex traits: from polygenic to omnigenic. *Cell* 169:1177–86
10. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *PNAS* 106:9362–67
11. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. 2017. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* 8:1826
12. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74
13. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* 518:317–30
14. Ardlie KG, Deluca DS, Segrè AV, Sullivan TJ, Young TR, et al. 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348:648–60
15. Aguet F, Ardlie KG, Cummings BB, Gelfand ET, Getz G, et al. 2017. Genetic effects on gene expression across human tissues. *Nature* 550:204–13
16. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. 2010. A method and server for predicting damaging missense mutations. *Nat. Methods* 7:248–49

17. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, et al. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 335:823–28
18. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*. 15:901–13
19. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLOS Comput. Biol.* 6:e1001025
20. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*. 20:110–21
21. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, et al. 2015. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* 47:1228–35
22. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478:476–82
23. Lambert J-C, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, et al. 2013. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* 45:1452–58
24. Furberg H, Kim Y, Dackor J, Boerwinkle E, Franceschini N, et al. 2010. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat. Genet.* 42:441–47
25. Akbarian S, Liu C, Knowles JA, Vaccarino FM, Farnham PJ, et al. 2015. The PsychENCODE project. *Nat. Neurosci.* 18:1707–12
26. Lawrence M, Daujat S, Schneider R. 2016. Lateral thinking: how histone modifications regulate gene expression. *Trends Genet*. 32:42–56
27. Trynka G, Sandor C, Han B, Xu H, Stranger BE, et al. 2013. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* 45:124–30
28. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, et al. 2012. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*. 22:1790–97
29. Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 38:e164
30. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337:1190–95
31. Hou L, Ma T, Zhao H. 2014. Incorporating functional annotation information in prioritizing disease associated SNPs from genome wide association studies. *Sci. China Life Sci.* 57:1072–79
32. Kindt AS, Navarro P, Semple CA, Haley CS. 2013. The genomic signature of trait-associated variants. *BMC Genom*. 14:108
33. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46:310–15
34. Ritchie GR, Dunham I, Zeggini E, Flicek P. 2014. Functional annotation of noncoding sequence variants. *Nat. Methods* 11:294–96
35. Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, et al. 2015. A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* 47:955–61
36. Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* 12:931–34
37. Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* 28:817–25
38. Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* 9:215–16
39. Chan RCW, Libbrecht MW, Roberts EG, Bilmes JA, Noble WS, Hoffman MM. 2018. Segway 2.0: Gaussian mixture models and minibatch training. *Bioinformatics* 34:669–71
40. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, et al. 2012. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res*. 41:827–41
41. Lu Q, Hu Y, Sun J, Cheng Y, Cheung K-H, Zhao H. 2015. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci. Rep.* 5:10576

42. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. 2016. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* 48:214–20
43. Parisi F, Strino F, Nadler B, Kluger Y. 2014. Ranking and combining multiple predictors without labeled data. *PNAS* 111:1253–58
44. Lu Q, Powles RL, Wang Q, He BJ, Zhao H. 2016. Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies. *PLOS Genet*. 12:e1005947
45. Lu Q, Powles RL, Abdallah S, Ou D, Wang Q, et al. 2017. Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset Alzheimer's disease. *PLOS Genet*. 13:e1006933
46. Backenroth D, He Z, Kiryluk K, Boeva V, Pethukova L, et al. 2018. FUN-LDA: a latent Dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation: methods and applications. *Am. J. Hum. Genet.* 102:920–42
47. Singleton MV, Guthery SL, Voelkerding KV, Chen K, Kennedy B, et al. 2014. Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *Am. J. Hum. Genet.* 94:599–610
48. Javed A, Agrawal S, Ng PC. 2014. Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nat. Methods* 11:935–37
49. Chen L, Jin P, Qin ZS. 2016. DIVAN: accurate identification of non-coding disease-specific risk variants using multi-omics profiles. *Genome Biol*. 17:252
50. Bodea CA, Mitchell AA, Bloemendal A, Day-Williams AG, Runz H, Sunyaev SR. 2018. PINES: phenotype-informed tissue weighting improves prediction of pathogenic noncoding variants. *Genome Biol*. 19:173
51. Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, et al. 2011. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 474:380–84
52. Hou L, Chen M, Zhang CK, Cho J, Zhao H. 2014. Guilt by rewiring: gene prioritization through network rewiring in genome wide association studies. *Hum. Mol. Genet.* 23:2780–90
53. Bader GD, Betel D, Hogue CW. 2003. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*. 31:248–50
54. Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, et al. 2013. The BioGRID interaction database: 2013 update. *Nucleic Acids Res*. 41:D816–23
55. Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, et al. 2004. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*. 32:D497–501
56. Fromer M, Roussos P, Sieberts SK, Johnson JS, Kavanagh DH, et al. 2016. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* 19:1442–53
57. Franzén O, Ermel R, Cohain A, Akers NK, Di Narzo A, et al. 2016. Cardiometabolic risk loci share downstream cis-and trans-gene regulation across tissues and diseases. *Science* 353:827–30
58. Finucane HK, Reshef YA, Anttila V, Slowikowski K, Gusev A, et al. 2018. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* 50:621–29
59. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. 2010. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLOS Genet*. 6:e1000888
60. Mehta D, Heim K, Herder C, Carstensen M, Eckstein G, et al. 2013. Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. *Eur. J. Hum. Genet.* 21:48–54
61. Roeder K, Bacanu SA, Wasserman L, Devlin B. 2006. Using linkage genome scans to improve power of association in genome scans. *Am. J. Hum. Genet.* 78:243–52
62. Lu Q, Yao X, Hu Y, Zhao H. 2016. GenoWAP: GWAS signal prioritization through integrated analysis of genomic functional annotation. *Bioinformatics* 32:542–48
63. Lin WY, Lee WC. 2012. Improving power of genome-wide association studies with weighted false discovery rate control and prioritized subset analysis. *PLOS ONE* 7:e33716
64. Saccone SF, Saccone NL, Swan GE, Madden PA, Goate AM, et al. 2008. Systematic biological prioritization after a genome-wide association study: an application to nicotine dependence. *Bioinformatics* 24:1805–11

65. Chen GK, Witte JS. 2007. Enriching the analysis of genomewide association studies with hierarchical modeling. *Am. J. Hum. Genet.* 81:397–404
66. Heron EA, O'Dushlaine C, Segurado R, Gallagher L, Gill M. 2011. Exploration of empirical Bayes hierarchical modeling for the analysis of genome-wide association study data. *Biostatistics* 12:445–61
67. Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC. 2007. Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet. Epidemiol.* 31:871–82
68. Fridley BL, Serie D, Jenkins G, White K, Bamlet W, et al. 2010. Bayesian mixture models for the incorporation of prior knowledge to inform genetic association studies. *Genet. Epidemiol.* 34:418–26
69. Fridley BL, Iversen E, Tsai YY, Jenkins GD, Goode EL, Sellers TA. 2011. A latent model for prioritization of SNPs for functional studies. *PLOS ONE* 6:e20764
70. Ming J, Dai M, Cai M, Wan X, Liu J, Yang C. 2018. LSMM: a statistical approach to integrating functional annotations with genome-wide association studies. *Bioinformatics* 34:2788–96
71. Wang K, Li M, Bucan M. 2007. Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.* 81:1278–83
72. Ballard DH, Cho J, Zhao H. 2010. Comparisons of multi-marker association methods to detect association between a candidate region and disease. *Genet. Epidemiol.* 34:201–12
73. Chun H, Ballard DH, Cho J, Zhao H. 2011. Identification of association between disease and multiple markers via sparse partial least-squares regression. *Genet. Epidemiol.* 35:479–86
74. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. 2015. MAGMA: generalized gene-set analysis of GWAS data. *PLOS Comput. Biol.* 11:e1004219
75. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89:82–93
76. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. 2013. Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Genet.* 92:841–53
77. Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, et al. 2010. A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* 87:139–45
78. Li M-X, Gui H-S, Kwan JSH, Sham PC. 2011. GATES: a rapid and powerful gene-based association test using extended Simes procedure. *Am. J. Hum. Genet.* 88:283–93
79. Bacanu SA. 2012. On optimal gene-based analysis of genome scans. *Genet. Epidemiol.* 36:333–39
80. Dinu V, Miller PL, Zhao H. 2007. Evidence for association between multiple complement pathway genes and AMD. *Genet. Epidemiol.* 31:224–37
81. Ballard D, Abraham C, Cho J, Zhao H. 2010. Pathway analysis comparison using Crohn's disease genome wide association studies. *BMC Med. Genom.* 3:25
82. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, et al. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467:832–38
83. Shahbaba B, Shachaf CM, Yu Z. 2012. A pathway analysis method for genome-wide association studies. *Stat. Med.* 31:988–1000
84. Wang L, Jia P, Wolfinger RD, Chen X, Grayson BL, et al. 2011. An efficient hierarchical generalized linear mixed model for pathway analysis of genome-wide association studies. *Bioinformatics* 27:686–92
85. Zhang W, Chen Y, Sun F, Jiang R. 2011. DomainRBF: a Bayesian regression approach to the prioritization of candidate domains for complex diseases. *BMC Syst. Biol.* 5:55
86. Sun R, Hui S, Bader GD, Lin X, Kraft P. 2019. Powerful gene set analysis in GWAS with the Generalized Berk-Jones statistic. *PLOS Genet*. 15:e1007530
87. Erten S, Bebek G, Koyutürk M. 2011. Vavien: an algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks. *J. Comput. Biol.* 18:1561–74
88. Chen J, Aronow BJ, Jegga AG. 2009. Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinform*. 10:73
89. White S, Smyth P. 2003. Algorithms for estimating relative importance in networks. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 266–75. New York: Assoc. Comput. Mach.

90. Chen J, Bardes EE, Aronow BJ, Jegga AG. 2009. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res*. 37:W305–11
91. Navlakha S, Kingsford C. 2010. The power of protein interaction networks for associating genes with diseases. *Bioinformatics* 26:1057–63
92. Guney E, Oliva B. 2012. Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization. *PLOS ONE* 7:e43557
93. Jia P, Zheng S, Long J, Zheng W, Zhao Z. 2010. dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics* 27:95–102
94. Chen M, Cho J, Zhao H. 2011. Incorporating biological pathways via a Markov random field model in genome-wide association studies. *PLOS Genet*. 7:e1001353
95. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, et al. 2015. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* 47:1091–98
96. Gusev A, Ko A, Shi H, Bhatia G, Chung W, et al. 2016. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* 48:245–52
97. Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, et al. 2018. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* 9:1825
98. Mancuso N, Shi H, Goddard P, Kichaev G, Gusev A, Pasaniuc B. 2017. Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *Am. J. Hum. Genet.* 100:473–87
99. Xu Z, Wu C, Wei P, Pan W. 2017. A powerful framework for integrating eQTL and GWAS summary data. *Genetics* 207:893–902
100. Nagpal S, Meng X, Epstein MP, Tsoi LC, Patrick M, et al. 2019. TIGAR: an improved Bayesian tool for transcriptomic data imputation enhances gene mapping of complex traits. *Am. J. Hum. Genet.* 105:258–66
101. Hu Y, Li M, Lu Q, Weng H, Wang J, et al. 2019. A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat. Genet.* 51:568–76
102. Flutre T, Wen X, Pritchard J, Stephens M. 2013. A statistical framework for joint eQTL analysis in multiple tissues. *PLOS Genet*. 9:e1003486
103. Raj T, Li YI, Wong G, Humphrey J, Wang M, et al. 2018. Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nat. Genet.* 50:1584–92
104. Mancuso N, Freund MK, Johnson R, Shi H, Kichaev G, et al. 2019. Probabilistic fine-mapping of transcriptome-wide association studies. *Nat. Genet.* 51:675–82
105. Freudenberg J, Propping P. 2002. A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics* 18(Suppl. 2):S110–15
106. Lettre G, Rioux JD. 2008. Autoimmune diseases: insights from genome-wide association studies. *Hum. Mol. Genet.* 17:R116–21
107. Visscher PM, Yang J. 2016. A plethora of pleiotropy across complex traits. *Nat. Genet.* 48:707–8
108. van Rheenen W, Peyrot WJ, Schork AJ, Lee SH, Wray NR. 2019. Genetic correlations of polygenic disease traits: from theory to practice. *Nat. Rev. Genet.* 20:567–81
109. Zhou X, Stephens M. 2014. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* 11:407–9
110. Stephens M. 2013. A unified framework for association analysis with multiple related phenotypes. *PLOS ONE* 8:e65245
111. Chung D, Yang C, Li C, Gelernter J, Zhao H. 2014. GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. *PLOS Genet*. 10:e1004787
112. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, et al. 2018. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* 50:229–37
113. Int. Schizophr. Consort., Purcell SM, Wray NR, Stone JL, Visscher PM, et al. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460:748–52
114. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, et al. 2018. Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. *Hum. Mol. Genet.* 27:3641–49

115. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. 2009. Finding the missing heritability of complex diseases. *Nature* 461:747–53
116. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42:565–69
117. Jiang J, Li C, Paul D, Yang C, Zhao H. 2016. On high-dimensional misspecified mixed model analysis in genome-wide association study. *Ann. Stat.* 44:2127–60
118. Pasaniuc B, Price AL. 2017. Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* 18:117–27
119. Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, et al. 2015. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* 47:291–95
120. Zhou X. 2017. A unified framework for variance component estimation with summary statistics in genome-wide association studies. *Ann. Appl. Stat.* 11:2027–51
121. Holmes JB, Speed D, Balding DJ. 2019. Summary statistic analyses can mistake confounding bias for heritability. *Genet. Epidemiol.* 43:930–40
122. Speed D, Hemani G, Johnson MR, Balding DJ. 2012. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* 91:1011–21
123. Speed D, Cai N, UCELB Consort., Johnson MR, Nejentsev S, Balding DJ. 2017. Reevaluation of SNP heritability in complex human traits. *Nat. Genet*. 49:986–92
124. Speed D, Balding DJ. 2019. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.* 51:277–84
125. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, et al. 2011. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* 43:519–25
126. Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AA, et al. 2015. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* 47:1114–20
127. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. 2012. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* 28:2540–42
128. Lee H, Ripke S, Neale B, Faraone S, Purcell S, et al. 2013. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* 45:984–94
129. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, et al. 2015. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* 47:1236–41
130. Lu Q, Li B, Ou D, Erlendsdottir M, Powles RL, et al. 2017. A powerful approach to estimating annotation-stratified genetic covariance via GWAS summary statistics. *Am. J. Hum. Genet.* 101:939–64
131. Brainstorm C, Anttila V, Bulik-Sullivan B, Finucane HK, Walters RK, et al. 2018. Analysis of shared heritability in common disorders of the brain. *Science* 360:eaap8757
132. Tylee DS, Sun J, Hess JL, Tahir MA, Sharma E, et al. 2018. Genetic correlations among psychiatric and immune-related phenotypes based on genome-wide association data. *Am. J. Med. Genet. B* 177:641–57
133. Shi H, Mancuso N, Spendlove S, Pasaniuc B. 2017. Local genetic correlation gives insights into the shared genetic architecture of complex traits. *Am. J. Hum. Genet.* 101:737–51
134. Brown BC, AGEN-T2D Consort., Ye CJ, Price AL, Zaitlen N. 2016. Transethnic genetic-correlation estimates from summary statistics. *Am. J. Hum. Genet*. 99:76–88
135. Weissbrod O, Flint J, Rosset S. 2018. Estimating SNP-based heritability and genetic correlation in case-control studies directly and with summary statistics. *Am. J. Hum. Genet.* 103:89–99
136. Zheng J, Erzurumluoglu AM, Elsworth BL, Kemp JP, Howe L, et al. 2017. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* 33:272–79
137. Chatterjee N, Shi J, Garcia-Closas M. 2016. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* 17:392–406
138. Wei Z, Wang W, Bradfield J, Li J, Cardinale C, et al. 2013. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *Am. J. Hum. Genet.* 92:1008–12

139. Zhou X, Carbonetto P, Stephens M. 2013. Polygenic modeling with Bayesian sparse linear mixed models. *PLOS Genet*. 9:e1003264
140. Speed D, Balding DJ. 2014. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res*. 24:1550–57
141. Minnier J, Yuan M, Liu JS, Cai T. 2015. Risk classification with an adaptive naive Bayes kernel machine model. *J. Am. Stat. Assoc.* 110:393–404
142. Li C, Yang C, Gelernter J, Zhao H. 2014. Improving genetic risk prediction by leveraging pleiotropy. *Hum. Genet.* 133:639–50
143. Maier R, Moser G, Chen G-B, Ripke S, Coryell W, et al. 2015. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *Am. J. Hum. Genet.* 96:283–94
144. Vilhjalmsson BJ, Yang J, Finucane HK, Gusev A, Lindstrom S, et al. 2015. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* 97:576–92
145. So HC, Sham PC. 2017. Improving polygenic risk prediction from summary statistics by an empirical Bayes approach. *Sci. Rep.* 7:41262
146. Mak TSH, Porsch RM, Choi SW, Zhou X, Sham PC. 2017. Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.* 41:469–80
147. Ge T, Chen CY, Ni Y, Feng YA, Smoller JW. 2019. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* 10:1776
148. Schrodi SJ, Mukherjee S, Shan Y, Tromp G, Sninsky JJ, et al. 2014. Genetic-based prediction of disease traits: Prediction is very difficult, especially about the future. *Front. Genet.* 5:162
149. Hu Y, Lu Q, Powles R, Yao X, Yang C, et al. 2017. Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLOS Comput. Biol.* 13:e1005589
150. Hu Y, Lu Q, Liu W, Zhang Y, Li M, Zhao H. 2017. Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. *PLOS Genet*. 13:e1006836
151. Maier RM, Zhu Z, Lee SH, Trzaskowski M, Ruderfer DM, et al. 2018. Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nat. Commun*. 9:989
152. Timberlake AT, Choi J, Zaidi S, Lu Q, Nelson-Williams C, et al. 2016. Two locus inheritance of non-syndromic midline craniosynostosis via rare SMAD6 and common BMP2 alleles. *eLife* 5:e20125
153. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. 2015. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* 21.29.1–21.29. 9
154. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326:289–93
155. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, et al. 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562:203–9
156. Collins FS, Varmus H. 2015. A new initiative on precision medicine. *N. Engl. J. Med.* 372:793–95
157. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, et al. 2016. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* 70:214–23
158. Nagai A, Hirata M, Kamatani Y, Muto K, Matsuda K, et al. 2017. Overview of the BioBank Japan Project: study design and profile. *J. Epidemiol.* 27:S2–8
159. So HC, Chau CK, Chiu WT, Ho KS, Lo CP, et al. 2017. Analysis of genome-wide association data highlights candidates for drug repositioning in psychiatry. *Nat. Neurosci.* 20:1342–49
160. Evans DM, Davey Smith G. 2015. Mendelian randomization: new applications in the coming age of hypothesis-free causality. *Annu. Rev. Genom. Hum. Genet.* 16:327–50
161. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, et al. 2016. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* 48:481–87