Quantitative Biology 2021, 9(2): 168–184
https://doi.org/10.15302/J-QB-021-0249

REVIEW

Advances and challenges in quantitative delineation of the genetic architecture of complex traits

Hua Tang[1,*], Zihuai He[2,3,*]

[1] Department of Genetics, Stanford University, Stanford, CA 94305, USA
[2] Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA 94305, USA
[3] Quantitative Sciences Unit, Department of Medicine, Stanford University, Stanford, CA 94305, USA
* Correspondence: huatang@stanford.edu, zihuai@stanford.edu

Received September 30, 2020; Revised February 14, 2021; Accepted February 22, 2021

***Background*: Genome-wide association studies (GWAS) have been widely adopted in studies of human complex traits and diseases.**
***Results*: This review surveys areas of active research: quantifying and partitioning trait heritability, fine mapping functional variants and integrative analysis, genetic risk prediction of phenotypes, and the analysis of sequencing studies that have identified millions of rare variants. Current challenges and opportunities are highlighted.**
***Conclusion*: GWAS have fundamentally transformed the field of human complex trait genetics. Novel statistical and computational methods have expanded the scope of GWAS and have provided valuable insights on the genetic architecture underlying complex phenotypes.**

Keywords: genome-wide association study; heritability; rare variants; biobank; colocalization; eQTL; polygenic risk scores; transcriptome-wide association study

Author summary: Genome-wide association studies have identified a large number of genotype-phenotype associations for a wide variety of complex traits and diseases. Here we provide an overview of recent developments in quantitative methods that make use of GWAS discoveries for probing into the biology underlying specific associations, for depicting the genetic architecture of complex phenotypes, as well as for genetic risk predictions. Methodological challenges are highlighted in the hope to inspire innovation.

INTRODUCTION

Genome-wide association studies (GWAS) have become an essential approach for studying complex traits and diseases. To date, 185,864 statistical associations of single nucleotide polymorphisms (SNPs) with 4,554 complex traits have been summarized in the US National Human Genome Research Institute (NHGRI)–European Bioinformatics Institute (EBI) GWAS Catalog [1] (https://www.ebi.ac.uk/gwas/, accessed May 28, 2020). The past ten years mark a transition into the post-GWAS era, during which studies have expanded both in scale and in scope, catalyzed by the biobank-scale study populations, high-throughput functional genomic data and novel analytic tools. This new phase of research activities is marked by three paradigm shifts. First, stimulated by the mounting evidence that many complex traits are highly polygenic, that is, they are influenced by a large number of genes and variants, recent GWAS have gone beyond gene-phenotype association and strived to depict the overall genetic architecture, broadly defined to include such parameters as the number of trait loci, the distribution of allelic effects, and the mode of actions of these alleles. Second, the observation that most association findings reside in non-coding regions emphasizes the urgent need for a principled framework that informs the

biological processes underlying genotype-phenotype associations. Third, whereas early GWAS were powered to identify trait variants that are relatively common, the expanding cohorts and decreasing cost of sequencing have made it feasible to uncover low frequency ($1\% <$ minor allele frequency, MAF $< 5\%$) and rare variants (MAF $< 1\%$). Altogether, these studies have yielded valuable insights regarding the genetic architecture of complex phenotypes in general and have generated testable hypotheses related to the biological mechanisms underlying specific phenotypes [2]. Furthermore, the ability to predict phenotype based on an individual's genotypes through polygenic risk scores (PRS) holds promise for implementing genomic-based medicine through disease risk stratification, prevention and intervention. This review aims to provide an overview of recent developments in analytic approaches for GWAS in these areas and to highlight challenges in translating GWAS results to biological knowledge and clinical utilities. Other important areas, such as phenome-wide association studies (PheWAS) and causal inference are reviewed elsewhere [3–6].

To set the stage, we summarize basic models and key concepts that will be used subsequently. Detailed explanations can be found in two previous reviews [7,8]. In essence, GWAS seek genomic locations where genetic variation is statistically associated with phenotypes. The vast majority of GWAS methods and analyses consider an underlying generative model, in which a quantitative phenotype (Y) is determined by the genotype of a set of causal variants, G, as well as non-genetic risk factors and covariates, X:

$$Y = \mu + \Sigma\beta_m G_m + \gamma X + \varepsilon, \quad (1)$$

where β represents the allelic genetic effect. This linear additive model can be readily extended into generalized linear models for non-Gaussian phenotypes, such as dichotomized disease status [9].

Initial GWAS focused on common variants, typically referring to bi-allelic SNPs with minor allele frequency above 5% [10]. Denote $G = \{0,1,2\}$ for individuals carrying 0,1, or 2 copies of the reference alleles at a SNP, respectively, the most basic yet still the most routinely performed GWAS analysis tests the statistical association between Y and G through a linear model for continuous phenotypes:

$$Y = \mu + \beta G + \gamma X + \varepsilon, \quad (2)$$

or a logistic model (for disease status). Under this single-variant model, the statistical inference focuses on testing the null hypothesis of no association, $H_0 : \beta = 0$. The test is applied to each SNP in turn. To correct for multiple testing, a p-value of 5×10^{-8} is considered genome-wide significant, based on a heuristic of testing one million independent tests genome-wide [11]. While multi-variants models have been developed and applied, the single-variant model remains an essential component of GWAS studies, partly owing to the ease with which summary-level statistics from this analysis can be shared across studies, usually in the form of the estimated allelic effect, the standard error and the p-value for each tested variant. Taking advantage of the much larger sample sizes available, summary-statistics-based analytic methods have been developed for conditional analysis [12], fine mapping [13–15], colocalization [16–20] and partition of heritability by functional categories [21]. Non-additive effects, representing gene-gene (GxG) interaction (also referred to as epistasis [22]) and gene-environment (GxE) interaction can be tested using a similar model that includes both the main effects and the interaction terms. For example, robust gene-sleep and gene-smoking interactions have been identified that are associated with lipid concentration [23,24]. However, genome-wide search for GxG and GxE interaction faces both computational challenges, as well as limited statistical power due to the combinatorial search space and increased multiple testing burden [25]. Instead of testing interaction between every pair of variants, several recent studies have focused on testing interactions between a variant and the rest of the genome (termed marginal epistasis [26]) or groups of genes implicated in biologically related pathways [27]. In the remaining of this review, we focus on the identifying, characterizing and interpreting the main effects of genetic loci.

The dramatically increased throughput, accompanied by decreased cost, of next-generation sequencing technology has enabled moderate- to large-scale sequencing-based GWAS. Initial sequencing studies focused on protein-coding regions in the genome, referred to as whole exome sequencing studies (WES) [28–30]; more recent work, however, has shifted towards whole genome sequencing (WGS), which allows unbiased interrogation of the entire genomic DNA sequence and provides the most comprehensive characterization of both protein-coding and non-coding genomic regions. These sequencing studies have markedly increased the catalogue of human genetic variation, of which a majority have low frequency (MAF $< 5\%$). Statistical methods for analyzing sequencing data and rare variants will be discussed in a subsequent section.

Linkage disequilibrium (LD) refers to the correlation of genotype between genetic variants. Physically closely located variants tend to be in strong LD, although many other factors, such as the demographic history, the local recombination rate and the age of mutation, contribute to LD structure [31]. As a result, the LD pattern varies considerably across the genome, as well as across

populations at a given genomic location. LD is a double-edged sword in GWAS. On the one hand, the correlation between the biological causal variant and many non-causal variants nearby (often referred to as markers or tagSNPs) enables us to identify genomic regions, or loci, that harbor trait variants without having to sequence every base pair of the genome. Taking advantage of LD, genotyping arrays (SNP chips) have been optimized for coverage, a measure of the ability to impute variants not directly genotyped on the array on the basis of neighboring genotyped tagSNPs [32–35]. Rationale, statistical methods and reference panel resources for genotype imputation can be extensively discussed and reviewed [36–39]. On the other hand, however, LD complicates the interpretation of the association: the significant associations cannot be interpreted as causal; rather, they are proxy variants highly correlated with neighboring causal variants. In regions of high LD, it is often infeasible to distinguish between the causal and proxy variants solely based on GWAS data, even if every base is sequenced. A number of methods have been developed aiming to refine the putative causal variants by integrating external biological knowledge.

Although the covariates, X, in Eq. (2) can include any factors relevant to the phenotype, by far the most commonly adjusted covariates are designed to eliminate spurious association due to population stratification (PS). PS can occur when the study population consists of discrete subpopulations, or when study participants represent continuous variation in ancestry proportions. Both forms of PS are expected to occur in biobanks. As one of the earliest examples of genetic association due to population stratification, Knowler *et al*. reported an association between an HLA haplotype and type 2 diabetes in a Native American population that has experienced recent admixture with Europeans. This association arose because both the HLA haplotype and type 2 diabetes occur at higher frequency among Indigenous Amerindian populations compared to in European populations; the association disappears when stratifying on ancestry proportion [40]. As the cohort sample size increases, even more subtle population structure within a continent can give rise to spurious associations. Such a possibility was demonstrated by an association between height and a SNP near the LCT gene, rs4988235, in Europe, where both height and the allele frequency of rs4988235 vary geographically between the Northern and Southern European populations [41]. To eliminate the spurious associations, effective methods have been developed and routinely incorporated in GWAS analyses. Currently the most commonly used approaches for accounting for population structure fall into two groups. One group of methods use genotype data to infer population structure, either by explicitly modeling of discrete subpopulations and continuous variation of genetic admixture, or by using a non-model-based method such as principal component analysis (PCA). The estimated ancestry proportions or leading PCs are then included in Model (2) as fixed-effects covariates, along with other covariates relevant to the phenotype. In contrast, a second group of methods turn Model (2) into a mixed-model by introducing genetic relationship between pairs of individuals through a random effects term in the form of a genetic relationship matrix (GRM) [42–45]. Compared to methods that adjust PC or ancestry proportions, these mixed models incur a higher computational cost but enjoy the advantage of additionally accounting for population structure represented by cryptic relatedness. Both groups of methods are currently widely used in GWAS. Genomic control and LDscore are two popular methods that allow users to assess, *post hoc*, potential residual population stratification based on genome-wide summary-level statistics [46,47].

Heritability refers to the proportion of the phenotypic variation attributable to genetic variation. Prior to the GWAS era, the heritability of a trait was estimated using correlation between biological relatives [48,49]. Under the generative model of Eq. (1) and assuming the identity of all trait variants and their corresponding allelic effects are known, the additive (or narrow sense) heritability is defined as $h^2 = \frac{\mathrm{var}(\Sigma \beta_m G_m)}{\mathrm{var}(Y)}$. An enigma arose in early GWAS that was referred to as missing heritability: for many traits, a striking gap exists between the pedigree-based heritability estimates and the phenotypic variance explained by the known height loci [50]. Consider adult height, a classical quantitative trait, which has an estimated heritability of 0.80 in family-based studies [51]; yet the 27 genome-wide significant SNPs – discovered in a cohort of nearly 30,000 – only account for ~3.7% of the phenotypic variance [52]. As explained in the next section, through a series of elegant analyses, we now appreciate that for a wide range of traits, the heritability is not missing but simply hidden under the polygenic architecture. For dichotomous disease phenotypes, the widely-adopted liability-threshold model stipulates that the disease status is determined by a latent quantitative trait – the liability – and an unobservable threshold, beyond which an individual will manifest the disease. In GWAS setting, the heritability estimate for a disease trait represents the genetic contribution to the variation in liability [53]. This quantity differs from the heritability defined based on co-occurrence of the disease in relatives [54].

POLYGENIC ARCHITECTURE

The polygenic genetic architecture was first articulated in

a seminal paper by Visscher and colleagues, who pioneered applying a variance component approach to estimate additive heritability in GWAS studies of unrelated individuals [55]. Instead of examining the effect of each SNP individually as in Model (2), this approach estimates the chip heritability, or the total contribution by all SNPs to a trait. The main conclusion is that a large proportion of heritable variation in height and other complex phenotypes can be attributed to common genetic variation. This finding is all the more remarkable because, at the time of publication, a total of 54 genome-wide significant height loci have been reported that, together, account for a mere 5% of the phenotypic variance [52,56,57]. Thus, it provides a quantitative explanation for the missing heritability phenomenon by supporting an infinitesimal model, in which the phenotypic variation of the trait represents the aggregated effects of a large number of variants, each with modest effects [58]. It predicts that, as the sample size increases, GWAS will continue to uncover additional trait-influencing variants that fill the missing heritability gap. This prediction has been empirically validated through subsequent GWAS for height and other complex traits: in a latest meta-analysis of height totaling more than 700,000 individuals, 3290 near-independent SNPs reach genome-wide significance and together explain nearly 25% of phenotypic variance [59]. While estimating variance-component models would normally require individual-level data, it is now possible to perform this type of analysis and estimate heritability using summary-level statistics, using methods such as LD score regression [46].

The understanding of polygenic architecture has a profound impact on the design and interpretation of GWAS studies. First, the expectation that each individual variant or locus explains a tiny fraction of the phenotypic variance, and therefore a large sample size is required to uncover these loci, has incentivized the formation of consortia and the share of summary-level statistics, which enable better powered meta-analyses. Second, much effort has been devoted, bearing much fruit, characterizing the distribution of trait-loci, both with respect to physical locations and to functional categories. Extending the variance component model, Yang *et al*. (2011) demonstrate that, for most traits, the phenotypic variance that can be explained by GWAS SNPs on each of the 22 human chromosomes is roughly proportional to the length of the chromosomes [55]. In other words, genes influencing a given trait are scattered throughout the genome rather than clustering together. This conclusion is strengthened in Loh *et al*. (2015), which stipulates that more than 70% of 1-Mb genomic regions harbor at least one schizophrenia risk variant [60].

Taking the polygenic model to its extreme has inspired the recent proposal of an omnigenic model [61]. which stipulates that essentially all genes can influence a disease or trait with infinitesimal effect, but only a modest number of *core genes* do so directly through trait-relevant pathways while the remaining *peripheral genes* only affect the phenotype indirectly through perturbing the expression of the core genes. A key corollary of the omnigenic model stipulates that heritability is predominantly attributable to the peripheral genes by virtue of the sheer number of peripheral genes, while the contribution of core genes to phenotypic variance are likely low How to define core genes and how to identify these units for a specific trait awaits further development, and whether the distinction captures the complexity of the underlying process has been a subject of debate [62]. Regardless, omnigenic model serves as a timely reminder that genes and variants do not act in isolation; instead, their phenotypic consequences are dependent through tightly connected and intricately regulated networks. With expanding understanding of these networks, GWAS analyses in the next ten years will likely to shift focus from gene discovery to delineating the pathways and hierarchy among the myriads of variants that influencing a phenotype.

PARTITIONING OF HERITABILITY

While variants associated with a polygenic trait are distributed throughout the genome, the location of these variants with respect to genomic context is not random. The variance component model, applied to increasingly available large-scale GWAS and sequencing data (discussed in greater details in a subsequent section), has enabled informative dissection of the contributions from different types of variants to complex traits. In particular these analyses address two long-standing debates regarding the relative contributions of common versus rare variants [63], and the role of protein-coding sequences versus the non-coding genome [64].

Common vs. rare variants. Although the rationale of GWAS is rooted in the common disease-common variant hypothesis, the missing heritability problem has inspired the proposal of common disease-rare variant hypothesis, which argues that a large number of rare variants can have a substantial impact on the genetic susceptibility to common diseases [65,66]. The rare variants can be identified by genome sequencing technology. One common observation from association analysis of sequencing data is that the effect size of rare variants is generally larger than common variants [67], although this could be primarily due to a statistical power issue as the required sample size for detecting trait-associated rare variants can be significantly larger than for common variants of the

equal effect size [8]. To overcome this limitation, studies have sought to quantify the aggregated contributions of common or rare variants. Take height as an example, the array based chip heritability is 0.45 with common variants [55], while the estimate based on whole genome sequencing – including both common and rare variants – has reached 0.79 [68], much closer to the estimates of 0.73–0.81 based on the twin studies [51]. The results suggest that additional heritability can be potentially explained by rare variants that are not genotyped or imputed in the array-based studies [69]. We will return to discuss statistical methods for detecting rare variant association in a later section.

Coding vs. non-coding variants. While coding variants often exhibit a stronger effect on the trait of interest, they only represent 1%–2% of the genome. The remaining ~98% of the human genome is non-coding and harbors elements that dictate when, where, and how much protein-coding genes are transcribed or translated [70–73]. Among significantly associated variants from hundreds of GWAS, ~ 89% lie in non-coding regions [74]. Gusev *et al*. (2014) estimated the heritability of different functional categories across 11 common diseases [75]. Although protein coding regions are significantly enriched in trait-associated variants, as expected, these coding variants, together, only explain about 10% of the total heritability explained by all variants. The remaining heritability is attributed to non-coding elements such as DNaseI hypersensitivity sites, promoter and intronic regions. Specifically, for schizophrenia, coding variants are found to explain 21% of the chip heritability (11% from common variants; 10% from rare variants) while non-coding variants account for the remaining 79%. Using a summary-statistics based heritability partitioning algorithm, stratified LD score regression, Finucane *et al.* (2015) found strong enrichment of many non-coding functional annotation in GWAS associated regions across 17 complex diseases and traits [21].

INTEGRATIVE ANALYSIS FOR TRAIT MAPPING

A hallmark of GWAS is the agnostic interrogation of the entire genome. All variants are treated as equally likely to be relevant to the phenotype, *a priori*. However, GWAS identify statistical association, but do not reveal biologically causative genes or variants. This limitation is exacerbated by the LD between markers and by the empirical evidence that most variants implicated in GWAS do not change protein sequences; consequently, the causal variants and the mechanism of action remain unknown for most GWAS discoveries. Yet, as discussed above, variants are not created equal, a realization that has driven systematic efforts to annotate the genome based on functional and evolutionary properties, referred to as functional annotations [76–78]. Leveraging high-throughput technologies, projects such as ENCODE, Roadmap Epigenomics and Gene-Tissue-Expression (GTEx) produce catalogs of genome function at single nucleotide resolution [72,79,80]. These data open up unprecedented possibilities for elucidating the statistical associations using biological knowledge. As an example, a recent GWAS of clinically diagnosed Alzheimer's disease (AD) and AD-by-proxy (based on parental diagnoses) integrates array-based data from 71,880 cases and 383,378 controls from four independent cohorts: Alzheimer's disease working group of the Psychiatric Genomics Consortium (PGC-ALZ), the International Genomics of Alzheimer's Project (IGAP), the Alzheimer's Disease Sequencing Project (ADSP), and AD-by-proxy data from the UK Biobank (see below) [81]. The meta-analysis identified 29 risk loci, implicating 215 associated genes. These associated genes are strongly expressed in immune-related tissues and cell types (spleen, liver, and microglia) and implicate shared biology between AD and multiple health-related outcomes.

A versatile and widely adopted framework to integrate functional annotations into trait mapping introduces a hidden variable, $\delta_m \in \{0,1\}$, which indicates if a locus, or a variant, "causally" affects the phenotype. Here the term "causal" is used loosely to mean that locus or variant explains the observed statistical association. The key insight is that both the GWAS test statistics and the annotation at a locus inform about δ, and the two quantities are conditionally independent given δ. In other words, $P(S_m, A_m|\delta_m) = P(S_m|\delta_m)P(A_m|\delta_m)$, where S_m and A_m generically refer to the GWAS test statistics and functional annotations, respectively. $P(S_m|\delta_m)$ is specified by the GWAS model and is assumed known under the null and alternative hypotheses; $P(A_m|\delta_m)$ describes the degree of *enrichment* of specific functional annotation among causative locus/variants, and can be modeled using a logistic regression for binary annotations. The model parameters representing enrichment of each annotation and δ_m can be estimated using either a likelihood or a Bayesian approach [13–15]. An appealing feature of these methods is that they require only the association test statistics, thus enabling these methods to leverage the summary-level statistics from the much better powered meta-analyses. Another innovative feature of these methods is the estimation of the enrichment parameters on the basis of genome-wide data; this is in sharp contrast to previous candidate gene approaches, which rely on prior domain knowledge to choose the

appropriate annotation. Therefore, these methods can consider multiple annotation features simultaneously and adaptively integrate the phenotype-specific, relevant annotations, while ignoring non-informative annotations. As an example, the analysis of Pickrell (2014) considers 450 genomic annotations that include maps of DNase-I hypersensitivity in a variety of primary cell types and cell lines [14]. Consistent with biological expectations, GWAS SNPs are depleted at repressed chromatin regions and enriched at enhancers in cell types that are plausibly related to the traits.

LD between neighboring markers presents a thorny issue in jointly modeling GWAS and functional enrichment. Intuitively this is because the causality indicator, δ_m, affects the association statistic at SNP m, as well as neighboring SNPs through LD. This problem is especially challenging because a GWAS region may harbor more than one causative variant and because summary-level statistics do not convey the LD pattern. A number of methods have made significant progress through two strategies: first, the LD between SNPs can be approximated by using an external reference population, such as individuals of matching ethnicity from the 1000 genomes project [82], and second, explicitly modeling the multivariate test statistic of all SNPs at each locus, R, whose mean depends on the vector of hidden causative indicators and whose covariance depends the approximate LD matrix computed on the external reference population [83–85]. However, applications of these methods to integrate high-dimensional annotation features face computational challenges; methods that improve statistical and computational efficiency and biological interpretability are under active development.

In the methods reviewed thus far, the annotation variables are assumed to be observed. These variables often indicate specific attributes, such as "non-synonymous variant" or "repressed chromatin in hepatocyte". A special class of annotations are derived from large-scale expression-QTL (eQTL) experiments, which use gene expression, measured by microarray or through RNA-seq, as phenotype [86,87]. These studies, such as GTEx, are building an ever-expanding compendium of variants that affect RNA abundance in a variety of human tissues [80,88]. The decreasing cost of single-cell RNA sequencing will enable the characterization of gene expression at cell-type resolution in the near future. The trove of eQTL datasets present unprecedented opportunity for elucidating biological actions underlying GWAS associations, as empirical evidence demonstrates that a significant proportion of GWAS SNPs fall on or close to eQTLs [88]. Thus, the enrichment approaches described above can be applied with annotations indicating if a variant is an eQTL in a specific tissue. However, integrating eQTL evidence as annotation also presents unique challenges. Like GWAS, eQTL studies identify statistical associations and thus face the same limitations: causal regulatory variants may not reach statistical significance due to insufficient statistical power, and the precise regulatory variants are not unambiguously identified due to LD. These unique challenges have inspired the so-called colocalization methods, which aim to determine if the same set of variants give rise to the observed association with both gene expression and GWAS phenotypes [16–20]. Applications of these methods have not only refined putative causal variants underlying GWAS hits, but have also pointed to candidate tissues in which the actions of these eQTL variants may take place. Fully accounting for the LD pattern in GWAS and eQTL studies and jointly modeling eQTLs from multiple tissues will further enhance colocalization analysis but these remain open problems.

In parallel to colocalization analysis, the ability to systematically interrogate the regulatory wiring of gene expression through eQTL studies has inspired the transcriptome-wide association study (TWAS) framework [89–92]. These methods consist of three steps: first, for every gene, a genotype-to-expression prediction model is built using the eQTL data; second, these predictive models are applied to genotype data in a GWAS, predicting the unmeasured gene expression levels in GWAS individuals; last, the association between the predicted gene expression levels and the GWAS phenotype is tested, in a model similar to Eq. (2), but substituting genotype with the expression of a single gene. The improved statistical power of TWAS are due to two factors: first, for genes whose RNA expression is associated with the GWAS trait, TWAS leverages biological insight to aggregate the effects of multiple variants into a single variable, whose association with the trait is expected to be stronger than the corresponding association between the trait and each variant individually. Second, testing at the gene level dramatically reduces the multiple testing burden; thus, while a GWAS typically corrects for one million tests, the transcriptome-wide significance level for TWAS is based on testing 20,000 genes. Despite its considerable success, two open challenges in TWAS invite novel methodology. First, the predictive accuracy of expression is low for many genes, and it is reasonable to expect that the power of TWAS can be substantially improved with a more accurate gene expression prediction model. The prediction will improve as the eQTL sample increases; at the same time, novel ideas such as UTMOST have shown promise by leveraging information across multiple tissues [90]. Second, an appealing feature of TWAS is interpretability: TWAS implicitly assume that the genetic effect on

a phenotype is mediated through gene expression, and therefore genes identified in TWAS hint at the functional pathways and biological processes. However, expression-phenotype association can arise without *bona fide* biological causality as explained in [93]. Statistical frameworks that facilitate rigorous interpretation of TWAS association and estimation of causal effects are under active development [94].

GENETIC RISK PREDICTION AND POLYGENIC RISK SCORE

An initial expectation of GWAS was that variants identified through these studies could be used to predict an individual's genetic predisposition for various complex diseases. Polygenic risk score (PRS), which aggregates genetic risk variants into a single score to predict disease, can be computed at birth and, in combination with other non-genetic risk factors, has the potential to guide the prevention, diagnosis, and treatment of diseases, as well as for life planning. It should be emphasized that the risks of most complex traits are influenced by an array of non-genetic factors, and therefore PRS alone cannot be expected to achieve accurate prediction at the individual level. Nonetheless, recent work has demonstrated remarkable success using genome-wide PRS for population-level risk stratification: for coronary artery disease (CAD), PRS identified 8% of the population with at least three-fold elevated risk; as a comparison, rare monogenic mutations in familial hypercholesterolemia, which confers comparable risk for CAD, has a prevalence of 0.4% of the population [95]. Likewise, even though the current PRS for schizophrenia does not achieve the sensitivity and specificity for it to be used for individual prediction, the odds ratios of manifesting the disease between individuals in the top and bottom decile range between 7–20 in three large cohorts [96].

In its simplest form, PRS is defined as the weighted sum of risk alleles with the weights given by the allelic effects, $\Sigma_{m=1}^{M}\beta_m G_m$. For a quantitative trait, PRS can be interpreted as the predicted deviation of an individual from the population mean. For the risk of dichotomous disease outcomes, PRS can be thought of as the predicted deviation, from the population mean, on the liability scale. Building PRS is fundamentally a high-dimensional prediction problem, in which the number of predictors (genetic variants) far exceeds the number of training individuals. Several additional features pose challenges for PRS. First, LD creates complex correlation structure between variants in physical proximity, and multiple variants in a region may contribute to the genetic risk. Second, the polygenic architecture suggests that the true PRS depends on thousands, if not more, of risk variants; hence methods favoring extremely sparse models are not optimal. This hypothesis is supported by the empirical evidence that the prediction improves when many variants not reaching the genome-wide significance level are included in the PRS [97]. Lastly, while typical prediction algorithms assume availability of individual-level genotype and phenotype data, it is desirable to exploit the wealth of existing GWAS summary-level statistics, for which individual-level data is not available.

A simple PRS algorithm is referred to "P + T", which selects a set of independent SNPs by pruning correlated SNPs (P) and thresholding based on the single-variant association *p*-value (T) [97,98]. Specifically, SNPs are clumped based on LD and pruned to keep only the most significant SNP in each clump. SNPs with *p*-values exceeding a pre-specified threshold are kept and the estimated allelic effects are plugged-in to compute PRS. The tuning parameters, the LD threshold for clumping and the *p*-value threshold for retaining risk variants, are chosen through a validation dataset. P + T is intuitive and easy to implement on summary level-statistics: clumping can be based on the LD pattern in a publicly available dataset, such as the 1000 Genomes project, while thresholding and the computing of PRS do not require individual-level data. However, the performance of P + T is compromised by two limitations: first, in regions of high LD, the clumping step necessarily selects only one SNP even if multiple risk variants independently contribute to the disease risk; second, since the same dataset is used to select SNPs and to estimate allelic effects, the allelic effects tend to be overestimated due to the phenomenon of "winner's curse" [99]. Subsequent research has tackled these problems through frequentist [100], empirical Bayes [101,102] and Bayesian [103–105] approaches. Additionally, several efforts have developed tools for, and demonstrated the benefits of, integrating functional annotation information [106,107] and leveraging shared genetic architecture of correlated phenotypes [108,109].

BIOBANK-SCALE DATA

The decreasing costs and increasing throughput of genomic technologies have propelled comprehensive genomic characterization of ultra-large cohorts, many of which are biobank repositories [110]. These biobanks host a rich catalog of phenotypes through participants' survey response or electronic medical record [111]. For example, UK Biobank is a recently established cohort, in which more than 500,000 participants are

genotyped at more than 820,000 SNP [112]. At the time of recruitment, an assortment of demographic, normal variation and health-related phenotypes was recorded for each participant; these participants are followed-up prospectively through linked health records. Summary statistics for more than 4,000 phenotypes are publicly available [113], facilitating integration of the UK biobank resource not only into studies of individual traits, but also characterization of shared genetic determinism between traits (pleiotropy) through PheWAS [4]. The sheer size of biobank data poses computational challenges when applying existing analysis methods. For example, the standard implementation of linear mixed model used to account for population stratification and sample relatedness requires computations of $O(MN^2)$ or $O(M^2N)$ (where M and N denote the number of SNPs and of individuals, respectively). One approach, implemented in the GENESIS package, generates a sparse, block-diagonal, GRM, which effectively reduces both the computational time and the memory requirement [114,115]. Alternatively, BOLT-LMM and SAIGE circumvent the need to compute and to perform the spectral decomposition on a GRM, by using preconditioned conjugate gradient to solve a linear systems of mixed-model equations [116–118]. In addition, these methods reduce the memory use by compactly storing raw genotypes instead of calculating and storing the GRM. Another issue arises in Biobank and EHR studies is that binary disease phenotypes often feature highly unbalanced case-control ratios. For such phenotypes, statistical significance based on the asymptotic approximations of the logistic regression may be inaccurate [119]. A number of methods, including SAIGE, ameliorate the problem by implementing a saddlepoint approximation for calibrating the score test statistics.

GENOME SEQUENCING STUDIES

Genome sequencing technology has extended the coverage of GWAS beyond array-based studies of common variants [120]. Sequencing-based GWAS studies have the advantage to investigate low frequency ($1\% < \text{MAF} < 5\%$) and rare ($\text{MAF} < 1\%$) variants, which are not practical to be directly and comprehensively measured in array-based studies (Fig. 1). Data generated from genome sequencing can also serve as a reference panel to impute unmeasured variants in array-based studies to expand the existing GWAS studies to study rare variants, although the imputation accuracy for rare variants is lower than for common variants [121–124]. Therefore, sequencing-based GWAS are particularly effective for characterizing the contribution of rare variants to complex phenotypes. However, the large number of variants, of which a majority have very low minor allele frequencies, poses analytic challenges. The current release of UK Biobank interrogated the exomes of 49,960 individuals, revealing more than four million variants [125]. As WGS replaces WES and more individuals are sequenced, the number of variants increases accordingly. The Alzheimer's Disease Sequencing Project (ADSP) Discovery Extension Study performed whole genomes sequencing in 3085 individuals of diverse genetic ancestry and identified ~85 million variants [126]. Likewise, more than 400 million single-nucleotide and insertion/deletion variants have been detected in the 53,581 genome sequences generated by the multi-ethnic Trans-Omics for Precision Medicine (TOPMed) program [127,128].

When the minor allele is very rare, single-variant association test used for analyzing common variants (Eq. (2)) suffers from low statistical power. To improve power, gene-based methods, such as burden and dispersion tests, aggregate rare variants in a gene or region [124,129–132]. Many computational techniques described above for single variant analysis have been extended to gene/region-based analysis of sequencing studies [30,118,133]. For exome sequencing data, the coding region of a gene serves as a natural unit for grouping variants. In whole genome sequencing studies, however, such an approach would leave out a large number of non-coding variants. Currently, most WGS studies adopt a heuristic strategy of sliding window: a gene-based test is applied to rare variants within a contiguous window of fixed width. Bonferroni correction is used to account for multiple testing [134,135]. This strategy is suboptimal because the test statistics are correlated due to window overlap. Moreover, the optimal window size is often unknown, and misspecification may lead to loss of statistical power. An alternative approach is to group variants by annotation, such as chromatin states, genomic context (coding, intron, promoter, UTR, and intergenic) or regulatory functions (*e.g.*, predicted enhancers, transcription factor binding and DNase-hypersensitive sites) [136]. A limitation of such approaches is that the choice of categories for testing is subjective. Recently, data-driven screening strategies, at the genome-wide scale, have been proposed to dynamically choose the window size and localize the windows in which association signals reside [137,138]. To account for the linkage disequilibrium among genetic variants and window overlap, Monte Carlo simulations can be used to establish an empirical significant threshold, which controls for the genome-wise type I error rate [118]. Alternatively, He *et al*. (2019) proposed an analytical estimate of the significance threshold for WGS window-based analysis using extreme-value distribution (*e.g.*, the Gumbel distribution) [119].

In addition to the single nucleotide variants, WGS have

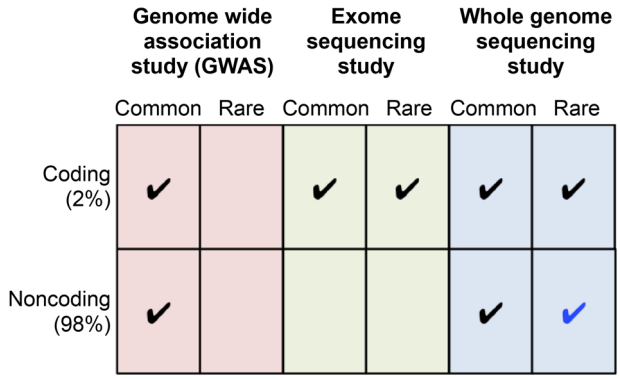

Figure 1. Genetic variants directly measured in different types of studies. The large number of rare variants in the non-coding genome identified by whole genome sequencing studies presents particular challenges, and therefore is highlighted.

enabled the characterization of a broad range of structure variations (SVs), which refers to sequence variation larger than 50 base pairs (bp) in size and includes copy number variants (CNVs), mobile element insertions (MEIs), inversions, and complex rearrangements. Although SVs are abundant in the genome and can have pronounced phenotypic impact through disrupting coding sequence or regulation, they are difficult to genotype by array-based platforms. Recent studies have developed accurate SV calling algorithms using WGS data [139–141], as well as have identified SVs associated with gene expression and complex traits [142,143].

GENERALIZABILITY OF GWAS RESULTS ACROSS POPULATIONS AND MULTI-ETHNIC STUDIES

Until recently, our knowledge regarding complex traits has been derived primarily from populations of European ancestry. A survey published in 2019 reveals that minority populations remain severely under-represented in GWAS studies: among studies curated by the NHGRI-EBI GWAS Catalog, over 78% of the participants are Europeans; in contrast, the two largest US minority groups, African Americans and Hispanic Americans, constitute over 25% of the population but less than 5% of GWAS participants [144]. There is no doubt that GWAS cohorts with expanded genetic diversity is a priority for all individuals to benefit from the fruit of genomic medicine. Yet analyzing multi-ethnic GWAS poses special analytic challenges. Population stratification is a particular concern because a number of minority populations, such as the African American and the Hispanic populations have experienced recent genetic admixture, which acts as a confounder and may give rise to spurious genotype-phenotype association. Methods for correcting PS was discussed in an earlier section.

A second challenge faced by GWAS in minority populations is that because the power of discovering genotype-phenotype association depends on sample size, far fewer trait loci have been identified in non-European populations. Likewise, the under-representation of non-European participants hinders the inference of causal variants and the estimation of allelic effects in these populations. Therefore, although nearly all analytic strategies described above are applicable to any population, they are less effective for minority populations. Importantly, several studies have demonstrated that polygenic risk scores do not transfer well across ethnicities; in other words, the approach that uses European GWAS results to select risk variants and to estimate their allelic effects yields polygenic risk scores with poorer predictive accuracy when applied to non-European populations. Multiple factors contribute to this phenomenon. First, trait loci and risk variants that contribute substantially to the phenotypic variation in a minority population may be entirely missed in a European GWAS if the causal variants are rare, not well tagged by markers on the genotyping array, or have weak effects in the study cohort. Consider the example of *APOL1*: two African-ancestry-specific risk variants in this gene confer increased risk for specific forms of kidney diseases; the risk alleles are virtually absent in all populations with no African ancestry [145]. Thus, despite an odds ratio of 7.3 for hypertension-attributed end-stage kidney disease, the association could not have been identified in a European GWAS. Second, heterogeneity in allelic effects among populations is well documented. As an example, multiple studies support a causal role of the *APOEε*4 allele that increases the risks for Alzheimer's disease and coronary artery disease across ethnicities [146]. Yet the magnitudes of effects, for both diseases, are higher in East Asians than in Europeans and African Americans [147]. Therefore, PRS built based on European GWAS, which uses the allelic effects estimated in Europeans, would not generalize accurately to East Asian individuals. Lastly, even when a risk variant occurs in all populations and exerts identical allelic effects, the SNPs used in PRS are often tag SNPs and not the causal variants. Therefore, a PRS model built using European cohorts selects SNPs based on the LD pattern in Europeans. As the LD structure differs between populations, these SNPs are often not optimal tags for the causal variants in non-European populations; in other words, tagging SNPs selected based on European GWAS are expected to be poorer surrogates in non-European populations on average, leading to reduced predictive accuracy. For the same reasons,

integrative analyses–such as determining colocalization of a GWAS association in a minority population and eQTL evidence – are hampered because eQTL and other-omics studies have also been conducted predominantly in European populations. To fill this gap, concerted efforts are underway to expand the diversity in GWAS cohort. Consortia such as the Human Heredity and Health in Africa (H3Africa) focus on establishing infrastructure for genomic studies in historically under-represented populations [148]. Multi-ethnic cohorts offer opportunities to minority-specific and trans-ethnic GWAS for a variety of diseases and phenotypes, examples of such cohorts include ALL of Us Research Program, Resource for Genetic Epidemiology Research on Aging (GERA) [149], the Population Architecture using Genomics and Epidemiology (PAGE) Consortium [150] and the Million Veterans Program [151]. In addition to these large-scale genotyping and sequencing efforts, the TOPMed program aims to systematically assay the methylome, transcriptome, proteome and metabolome in under-represented minority populations, which will enable the characterization of biological processes that underlie complex phenotypes in a population-specific context [127,128].

Complementary to the resource-building efforts, a number of analytic innovations have been proposed to ameliorate the challenges posed by the limited existing minority genomic resources. A key insight motivating these statistical approaches is the empirical evidence supporting a model that, for many complex phenotypes, a considerable fraction – but not all – of trait loci are shared across populations [152,153]. Therefore, a comprehensive, and population-aware, delineation of the genetic architecture underlying a phenotype of interest can benefit from borrowing information across populations. Such trans-ethnic approaches, however, must be adapted for the phenotype and population under study and acknowledge minority-specific genetic component. Coram *et al*. proposed an empirical Bayes approach, XPEB, which improves the power of GWAS in a minority populations by explicitly modeling and estimating shared genetic architecture [154]. Distinct from meta-analysis approaches, which treat all populations symmetrically, XPEB focuses on one minority population and treats information from other populations as auxiliary. It is adaptive in the sense that the method automatically assesses the relevance of the auxiliary information, and only 'borrows' information across ethnicities when empirical evidence supports substantial overlap. Furthermore, XPEB acknowledges, and has power to detect, population-specific biomolecules. This is in contrast to conventional meta-analysis approaches whose goal is to improve the overall mapping power but at a sacrifice of detecting minority-specific association. Applied to an African American lipid GWAS data, XPEB more than quadrupled the discoveries compared to the standard single-population approach. Furthermore, XPEB estimated that, for matching lipid traits, $> 95\%$ loci overlap between EU and AA; at the same time African-specific lipids loci are identified. In a follow-up paper, these authors proposed a trans-ethnic approach for improving PRS for minority individuals [155]. Novel methods that more precisely account for population-specific disease loci, rare variants and LD structure are needed to maximize the value of the rich multi-omics and phenotypic data that are being generated in diverse populations.

SUMMARY AND CONCLUSIONS

During the past ten years, GWAS have not only successfully identified abundant loci that underly a variety of traits and diseases but have also provided fundamental insights into the genetic architecture of complex phenotypes. Critical to these successes are novel statistical and computational approaches. Figure 2 provides an overview of major areas discussed in this review. Much of recent advances in understanding the genetic architecture of complex traits have benefitted from the open sharing of GWAS summary-level statistics from the single-variant analysis (Fig. 2A). Methods based on mixed-effects models have enabled the estimation of heritability of a trait, as well as partitioning the heritability based on attributes of the genomic regions (Fig. 2B). Integrating the GWAS summary-statistics with prior biological annotation or functional genomic data can substantially refine the associated genomic regions, although precisely identifying causal variant(s) underlying the statistical association remains challenging (Fig. 2C). Lastly, the predictive accuracy of PRS has continued to improve, owing to the expanding GWAS cohorts (Fig. 2D). We anticipate that GWAS in the next ten years will feature participants with more extensive phenotypes and representing increasing genetic diversity. Furthermore, fueled by technological advancement, molecular phenotypes, such as epigenetics marks and gene expression, are systematically characterized in GWAS cohorts. In the near future, it may be possible to assess these molecular phenotypes across tissues and even at a single cell resolution. Exploiting this data explosion, innovation in statistical and computational strategies have the potential to elucidate the biological functions underlying the genotype-phenotype associations and to inform disease risk or intervention.

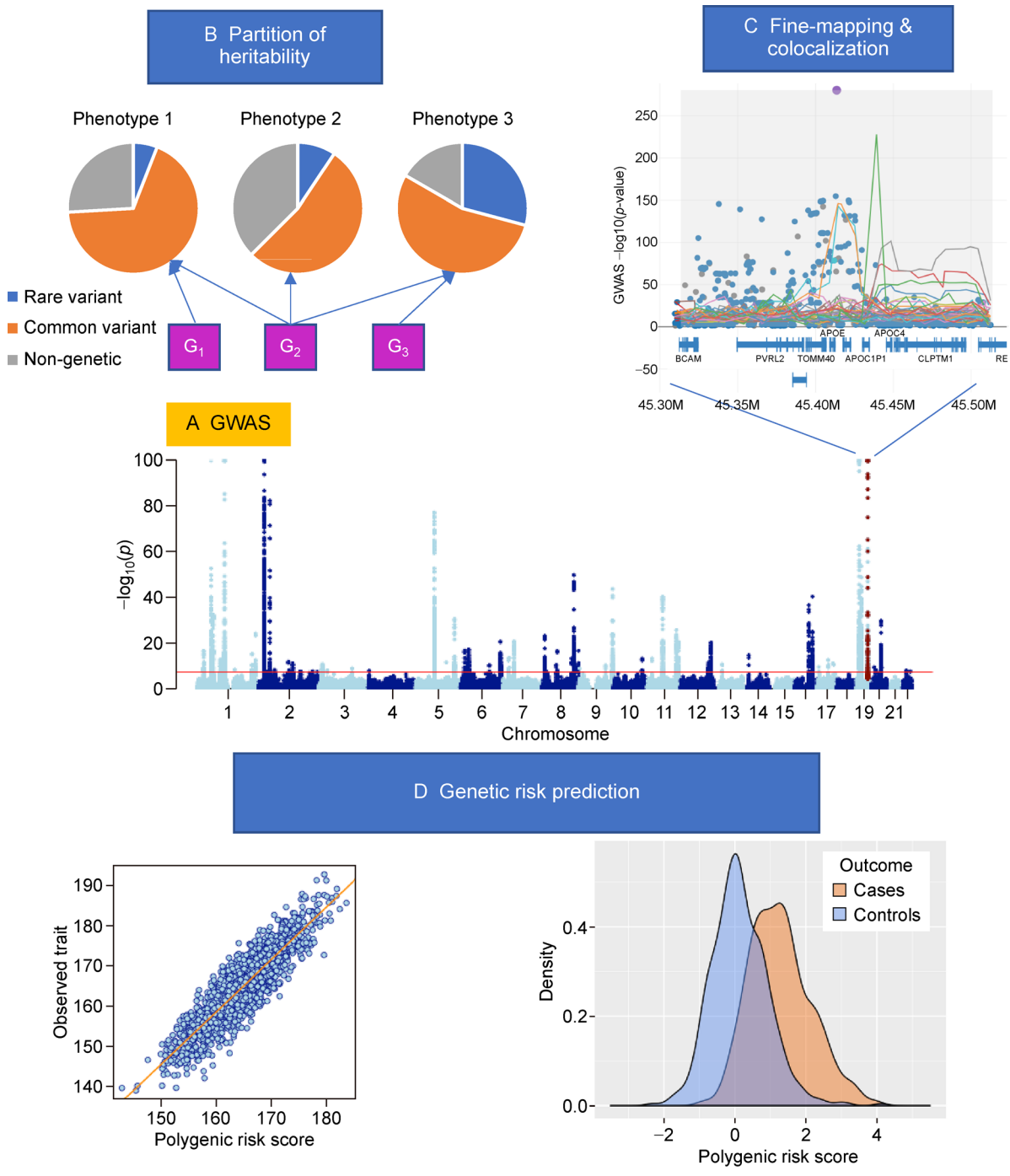

Figure 2. Overview of recent development in using GWAS results for understanding the genetic architecture of complex traits. (A) Manhattan plot summarizing a GWAS of low-density lipid (LDL) [156]. (B) Inferring genetic architecture. Pie charts represent the partition of the phenotypic variation to the contributions of common variants (orange), rare variants (blue) and non-genetic factors (gray) for three phenotypes (P1, P2, and P3). G_1, G_2 and G_3 indicate three variants; while G_1 and G_2 are associated with P1 and P3, respectively, G_2 has pleiotropic effects on all three phenotypes. (C) Colocalization analysis aiming to identify potential causal variant(s) underlying the association between LDL and the APOE region on chr19 by integrating eQTL evidence. Points are GWAS association results; colored curves are eQTL mapping results from various GTEx tissues. Plot produced using LocusFocus [157]. (D) Polygenic risk score for a simulated quantitative trait (left) and a simulated disease status (right).

ACKNOWLEDGMENTS

This work is supported by NIH R35GM127063 (HT) and NIH AG066206 (ZH).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Hua Tang and Zihuai He declare that they have no conflict of interests.

The article is a review article and does not contain any human or animal subjects performed by any of the authors.

REFERENCES

1. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., *et al.* (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res., 45, D896–D901
2. Crawford, N. G., KellyD. E. , Hansen, M. E. B., Beltrame, M. H., Fan, S., Bowman, S. L., Jewett, E., Ranciaro, A., Thompson, S., Lo, Y., *et al.* (2017) Loci associated with skin pigmentation identified in African populations. Science, 358, eaan8433
3. Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., Field, J. R., Pulley, J. M., Ramirez, A. H., Bowton, E., *et al.* (2013) Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat. Biotechnol., 31, 1102–1110
4. Bush, W. S., Oetjens, M. T. and Crawford, D. C. (2016) Unravelling the human genome-phenome relationship using phenome-wide association studies. Nat. Rev. Genet., 17, 129–145
5. Smith, G. D. and Hemani, G. (2014) Mendelian randomization: Geneticanchorsfor causal inference in epidemiological studies. Hum. Mol. Genet., 23, R89–R98
6. Pingault, J. B., O'Reilly, P. F., Schoeler, T., Ploubidis, G. B., Rijsdijk, F. and Dudbridge, F. (2018) Using genetic data to strengthen causal inference in observational research. Nat. Rev. Genet., 19, 566–580
7. Visscher, P. M., Brown, M. A., McCarthy, M. I. and Yang, J. (2012) Five years of GWAS discovery. Am. J. Hum. Genet., 90, 7–24
8. Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A. and Yang, J. (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet., 101, 5–22
9. McCullagh, P. and Nelder, J. A. (1989) Generalized Linear Models, 2nd Ed. (Monographs on Statistics and Applied Probability). Chapman and Hall/CRC
10. Wellcome, T., Case, T. and Consortium, C. (2007) Genomewide association study of 14, 000 cases of seven common diseases and 3, 000 shared controls Supplementary Information. Nature, 447, 661–78
11. Pe'er, I., Yelensky, R., Altshuler, D. and Daly, M. J. (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet. Epidemiol., 32, 381–385
12. Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Weedon, M. N., Loos, R. J., *et al.* (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat. Genet., 44, 369–375, S1–S3
13. Chung, D., Yang, C., Li, C., Gelernter, J. and Zhao, H. (2014) GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. PLoS Genet., 10, e1004787
14. Pickrell, J. K. (2014) Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am. J. Hum. Genet., 94, 559–573
15. Lu, Q., Yao, X., Hu, Y. and Zhao, H. (2016) GenoWAP: GWAS signal prioritization through integrated analysis of genomic functional annotation. Bioinformatics, 32, 542–548
16. He, X., Fuller, C. K., Song, Y., Meng, Q., Zhang, B., Yang, X. and Li, H. (2013) Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. Am. J. Hum. Genet., 92, 667–680
17. Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C. and Plagnol, V. (2014) Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet., 10, e1004383
18. Hormozdiari, F., van de Bunt, M., Segrè, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B. and Eskin, E. (2016) Colocalization of GWAS and eQTL Signals Detects Target Genes. Am. J. Hum. Genet., 99, 1245–1260
19. Giambartolomei, C., Zhenli Liu, J., Zhang, W., Hauberg, M., Shi, H., Boocock, J., Pickrell, J., Jaffe, A. E., Pasaniuc, B. and Roussos, P., *et al.* (2018) A Bayesian framework for multiple trait colocalization from summary association statistics. Bioinformatics, 34, 2538–2545
20. Wen, X., Pique-Regi, R. and Luca, F. (2017) Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. PLoS Genet., 13, e1006646
21. Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P. R., Anttila, V., Xu, H., Zang, C., Farh, K., *et al.* (2015) Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet., 47, 1228–1235
22. Cordell, H. J. (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum. Mol. Genet., 11, 2463–2468
23. Noordam, R., Bos, M. M., Wang, H., Winkler, T. W., Bentley, A. R., Kilpeläinen, T. O., de Vries, P. S., Sung, Y. J., Schwander, K., Cade, B. E., *et al.* (2019) Multi-ancestry sleep-by-SNP interaction analysis in 126,926 individuals reveals lipid loci stratified by sleep duration. Nat. Commun., 10, 5121
24. Bentley, A. R., Sung, Y. J., Brown, M. R., Winkler, T. W., Kraja, A. T., Ntalla, I., Schwander, K., Chasman, D. I., Lim, E., Deng,

X., *et al.* (2019) Multi-ancestry genome-wide gene-smoking interaction study of 387,272 individuals identifies new loci associated with serum lipids. Nat. Genet., 51, 636–648
25. Wei, W. H., Hemani, G. and Haley, C. S. (2014) Detecting epistasis in human complex traits. Nat. Rev. Genet., 15, 722–733
26. Crawford, L., Zeng, P., Mukherjee, S. and Zhou, X. (2017) Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. PLoS Genet., 13, e1006869
27. Fang, G., Wang, W., Paunic, V., Heydari, H., Costanzo, M., Liu, X., Liu, X., VanderSluis, B., Oately, B., Steinbach, M., *et al.* (2019) Discovering genetic interactions bridging pathways in genome-wide association studies. Nat. Commun., 10, 4274
28. Bis, J. C., Jian, X., Kunkle, B. W., Chen, Y., Hamilton-Nelson, K. L., Bush, W. S., Salerno, W. J., Lancour, D., Ma, Y., Renton, A. E., *et al.* (2020) Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation. Mol. Psychiatry, 25, 1859–1875
29. Marouli, E., Graff, M., Medina-Gomez, C., Lo, K. S., Wood, A. R., Kjaer, T. R., Fine, R. S., Lu, Y., Schurmann, C., Highland, H. M., *et al.* (2017) Rare and low-frequency coding variants alter human adult height. Nature, 542, 186–190
30. Zhao, Z., Bi, W., Zhou, W., VandeHaar, P., Fritsche, L. G. and Lee, S. (2020) UK Biobank whole-exome sequence binary phenome analysis with robust region-based rare-variant test. Am. J. Hum. Genet., 106, 3–12
31. Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Peltonen, L., *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. Nature, 467, 52–58
32. Hoffmann, T. J., Zhan, Y., Kvale, M. N., Hesselson, S. E., Gollub, J., Iribarren, C., Lu, Y., Mei, G., Purdy, M. M., Quesenberry, C., *et al.* (2011) Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. Genomics, 98, 422–430
33. Hoffmann, T. J., Kvale, M. N., Hesselson, S. E., Zhan, Y., Aquino, C., Cao, Y., Cawley, S., Chung, E., Connell, S., Eshragh, J., *et al.* (2011) Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. Genomics, 98, 79–89
34. Hunter-Zinck, H., Shi, Y., Li, M., Gorman, B. R., Ji, S. G., Sun, N., Webster, T., Liem, A., Hsieh, P., Devineni, P., *et al.* (2020) Genotyping array design and data quality control in the Million Veteran Program. Am. J. Hum. Genet., 106, 535–548
35. Bien, S. A., Wojcik, G. L., Zubair, N., Gignoux, C. R., Martin, A. R., Kocarnik, J. M., Martin, L. W., Buyske, S., Haessler, J., Walker, R. W., *et al.* (2016) Strategies for enriching variant coverage in candidate disease loci on a multiethnic genotyping array. PLoS One, 11, e0167758
36. Das, S., Abecasis, G. R. and Browning, B. L. (2018) Genotype imputation from large reference panels. Annu. Rev. Genomics Hum. Genet., 19, 73–96
37. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., *et al.* (2016) A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet., 48, 1279–1283
38. Howie, B., Marchini, J. and Stephens, M. (2011) Genotype imputation with thousands of genomes. G3- Genes, Genomes, Genet., 1, 457–470
39. Li, Y., Willer, C., Sanna, S. and Abecasis, G. (2009) Genotype imputation. Annu. Rev. Genomics Hum. Genet., 10, 387–406
40. Knowler, W. C., Williams, R. C., Pettitt, D. J. and Steinberg, A. G. (1988) Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am. J. Hum. Genet. 43, 520–526
41. Campbell, C. D., Ogburn, E. L., Lunetta, K. L., Lyon, H. N., Freedman, M. L., Groop, L. C., Altshuler, D., Ardlie, K. G. and Hirschhorn, J. N. (2005) Demonstrating stratification in a European American population. Nat. Genet., 37, 868–872
42. Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., Sabatti, C. and Eskin, E. (2010) Variance component model to account for sample structure in genome-wide association studies. Nat. Genet., 42, 348–354
43. Zhou, X. and Stephens, M. (2012) Genome-wide efficient mixed-model analysis for association studies. Nat. Genet., 44, 821–824
44. Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I. and Heckerman, D. (2011) FaST linear mixed models for genome-wide association studies. Nat. Methods, 8, 833–835
45. Listgarten, J., Lippert, C. and Heckerman, D. (2013) FaST-LMM-Select for addressing confounding from spatial structure and rare variants. Nat. Genet., 45, 470–471
46. Bulik-Sullivan, B. K., Loh, P. R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L. and Neale, B. M., *et al.* (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet., 47, 291–295
47. Bacanu, S. A., Devlin, B. and Roeder, K. (2000) The power of genomic control. Am. J. Hum. Genet., 66, 1933–1944
48. Falconer, D. S. and Mackay, T. F. C. (1962) Introduction to Quantitative Genetics. Benjamin-Cummings Pub Co
49. Haseman, J. K. and Elston, R. C. (1972) The investigation of linkage between a quantitative trait and a marker locus. Behav. Genet., 2, 3–19
50. Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., *et al.* (2009) Finding the missing heritability of complex diseases. Nature, 461, 747–753
51. Silventoinen, K., Sammalisto, S., Perola, M., Boomsma, D. I., Cornes, B. K., Davis, C., Dunkel, L., De Lange, M., Harris, J. R., Hjelmborg, J. V., *et al.* (2003) Heritability of adult body height: a comparative study of twin cohorts in eight countries. Twin Res., 6, 399–408
52. Gudbjartsson, D. F., Walters, G. B., Thorleifsson, G., Stefansson, H., Halldorsson, B. V., Zusmanovich, P., Sulem, P., Thorlacius, S., Gylfason, A., Steinberg, S., *et al.* (2008) Many sequence variants affecting diversity of adult human height. Nat. Genet., 40, 609–615

53. Lee, S. H., Wray, N. R., Goddard, M. E. and Visscher, P. M. (2011) Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet., 88, 294–305
54. Tenesa, A. and Haley, C. S. (2013) The heritability of human disease: estimation, uses and abuses. Nat. Rev. Genet., 14, 139–149
55. Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. Nat. Genet., 42, 565–569
56. Lettre, G., Jackson, A. U., Gieger, C., Schumacher, F. R., Berndt, S. I., Sanna, S., Eyheramendy, S., Voight, B. F., Butler, J. L., Guiducci, C., *et al.* (2008) Identification of ten loci associated with height highlights new biological pathways in human growth. Nat. Genet., 40, 584–591
57. Weedon, M. N., Lango, H., Lindgren, C. M., Wallace, C., Evans, D. M., Mangino, M., Freathy, R. M., Perry, J. R., Stevens, S., Hall, A. S., *et al.* (2008) Genome-wide association analysis identifies 20 loci that influence adult height. Nat. Genet., 40, 575–583
58. Fisher, R. A. (1918) The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinb., 53, 399–433
59. Yengo, L., Sidorenko, J., Kemper, K. E., Zheng, Z., Wood, A. R., Weedon, M. N., Frayling, T. M., Hirschhorn, J., Yang, J. and Visscher, P. M., *et al.* (2018) Meta-analysis of genome-wide association studies for height and body mass index in ~700, 000 individuals of European ancestry. Hum. Mol. Genet., 27, 3641–3649
60. Loh, P. R., Bhatia, G., Gusev, A., Finucane, H. K., Bulik-Sullivan, B. K., Pollack, S. J., de Candia, T. R., Lee, S. H., Wray, N. R., Kendler, K. S., *et al.* (2015) Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat. Genet., 47, 1385–1392
61. Boyle, E. A., Li, Y. I. and Pritchard, J. K. (2017) An expanded view of complex traits: from polygenic to omnigenic. Cell, 169, 1177–1186
62. Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. and Visscher, P. M. (2018) Common disease is more complex than implied by the core gene omnigenic model. Cell, 173, 1573–1580
63. Pritchard, J. K. and Cox, N. J. (2002) The allelic architecture of human disease genes: common disease-common variant...or not? Hum. Mol. Genet., 11, 2417–2423
64. Botstein, D. and Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat. Genet., 33, 228–237
65. Schork, N. J., Murray, S. S., Frazer, K. A. and Topol, E. J. (2009) Common vs. rare allele hypotheses for complex diseases. Curr. Opin. Genet. Dev., 19, 212–219
66. Gibson, G. (2012) Rare and common variants: twenty arguments. Nat. Rev. Genet., 13, 135–145
67. Bomba, L., Walter, K. and Soranzo, N. (2017) The impact of rare and low-frequency genetic variants in common disease. Genome Biol., 18, 77
68. Wainschtein, P., Jain, D. P., Yengo, L., Zheng, Z., TOPMed Anthropometry Working Group, Trans-Omics for Precision Medicine Consortium, Adrienne Cupples, L., Shadyab, A. H., McKnight, B., Shoemaker, B. M., *et al.* (2019) Recovery of trait heritability from whole genome sequence data. BioRxiv, doi: 10.1101/588020
69. Young, A. I. (2019) Solving the missing heritability problem. PLoS Genet., 15, e1008222
70. Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., *et al.* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. Nature, 478, 476–482
71. Khurana, E., Fu, Y., Colonna, V., Mu, X., Kang, H. M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., *et al.* (2013) Integrative annotation of variants from 1092 humans: Application to cancer genomics. Science, 342, 1235587
72. ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature, 489, 57–74
73. Altshuler, D., Daly, M. J. and Lander, E. S. (2008) Genetic mapping in human disease. Science, 322, 881–888
74. Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. and Manolio, T. A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA, 106, 9362–9367
75. Gusev, A., Lee, S. H., Trynka, G., Finucane, H., Vilhjálmsson, B. J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., *et al.* (2014) Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am. J. Hum. Genet., 95, 535–552
76. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. and Sunyaev, S. R. (2010) A method and server for predicting damaging missense mutations. Nat. Methods, 7, 248–249
77. Ng, P. C. and Henikoff, S. (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res., 31, 3812–3814
78. Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M. and Shendure, J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet., 46, 310–315
79. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., *et al.* (2015) Integrative analysis of 111 reference human epigenomes. Nature, 518, 317–330
80. The GTEx Consortium. (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science,348, 648–660
81. Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., Sealock, J., Karlsson, I. K., Hägg, S., Athanasiu, L., *et al.* (2019) Genome-wide meta-analysis identifies

new loci and functional pathways influencing Alzheimer's disease risk. Nat. Genet., 51, 404–413
82. Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., Abecasis, G. R., *et al.* (2015) A global reference for human genetic variation. Nature, 526, 68–74
83. Zhu, X. and Stephens, M. (2018) Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. Nat. Commun., 9, 4361
84. Kichaev, G., Yang, W. Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., Kraft, P. and Pasaniuc, B. (2014) Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet., 10, e1004722
85. Wen, X., Lee, Y., Luca, F. and Pique-Regi, R. (2016) Efficient integrative multi-SNP association analysis via deterministic approximation of posteriors. Am. J. Hum. Genet., 98, 1114–1129
86. Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S. and Cheung, V. G. (2004) Genetic analysis of genome-wide variation in human gene expression. Nature, 430, 743–747
87. Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature, 501, 506–511
88. GTEx Consortium (2017) Genetic effects on gene expression across human tissues. Nature, 550, 204–213
89. Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., *et al.* (2015) A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet., 47, 1091–1098
90. Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S. M., Yu, Z., Li, B., Gu, J., Muchnik, S., *et al.* (2019) A statistical framework for cross-tissue transcriptome-wide association analysis. Nat. Genet., 51, 568–576
91. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., Jansen, R., de Geus, E. J., Boomsma, D. I., Wright, F. A., *et al.* (2016) Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet., 48, 245–252
92. Barbeira, A. N., Dickinson, S. P., Bonazzola, R., Zheng, J., Wheeler, H. E., Torres, J. M., Torstenson, E. S., Shah, K. P., Garcia, T., Edwards, T. L., *et al.* (2018) Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat. Commun., 9, 1825
93. Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A. N., Knowles, D. A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., *et al.* (2019) Opportunities and challenges for transcriptome-wide association studies. Nat. Genet., 51, 592–599
94. Zhang, Y., Quick, C., Yu, K., Barbeira, A., Luca, F., Pique-Regi, R., Im, H. K. and Wen, X. (2019) Investigating tissue-relevant causal molecular mechanisms of complex traits using probabilistic TWAS analysis. bioRxiv, 808295
95. Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., Ellinor, P. T., *et al.* (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet., 50, 1219–1224
96. Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014) Biological insights from 108 schizophrenia-associated genetic loci. Nature, 511, 421–427
97. Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., Sklar, P., and the International Schizophrenia Consortium. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature, 460, 748–752
98. Stahl, E. A., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B. F., Kraft, P., Chen, R., Kallberg, H. J., Kurreeman, F. A., *et al.* (2012) Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat. Genet., 44, 483–489
99. Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. and Hirschhorn, J. N. (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat. Genet., 33, 177–182
100. Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X. and Sham, P. C. (2017) Polygenic scores via penalized regression on summary statistics. Genet. Epidemiol., 41, 469–480
101. So, H. C. and Sham, P. C. (2017) Improving polygenic risk prediction from summary statistics by an empirical Bayes approach. Sci. Rep., 7, 41262
102. Song, S., Jiang, W., Hou, L. and Zhao, H. (2020) Leveraging effect size distributions to improve polygenic risk scores derived from summary statistics of genome-wide association studies. PLOS Comput. Biol., 16, e1007565
103. Zhu, X. and Stephens, M. (2017) Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. Ann. Appl. Stat., 11, 1561–1592
104. Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P. R., Bhatia, G., Do, R., *et al.* (2015) Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am. J. Hum. Genet., 97, 576–592
105. Lloyd-Jones, L. R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K. E., Wang, H., Zheng, Z., Magi, R., Esko, T., *et al.* (2019) Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat. Commun., 10, 5086
106. Hu, Y., Lu, Q., Powles, R., Yao, X., Yang, C., Fang, F., Xu, X. and Zhao, H. (2017) Leveraging functional annotations in genetic risk prediction for human complex diseases. PLOS Comput. Biol., 13, e1005589
107. Hu, Y., Lu, Q., Liu, W., Zhang, Y., Li, M. and Zhao, H. (2017) Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. PLoS Genet., 13, e1006836
108. Li, C., Yang, C., Gelernter, J. and Zhao, H. (2014) Improving genetic risk prediction by leveraging pleiotropy. Hum. Genet., 133, 639–650

109. Maier, R. M., Zhu, Z., Lee, S. H., Trzaskowski, M., Ruderfer, D. M., Stahl, E. A., Ripke, S., Wray, N. R., Yang, J., Visscher, P. M., *et al.* (2018) Improving genetic prediction by leveraging genetic correlations among human diseases and traits. Nat. Commun., 9, 989
110. Ginsburg, G. S., Burke, T. W. and Febbo, P. (2008) Centralized biorepositories for genetic and genomic research. JAMA, 299, 1359–1361
111. Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., Sanderson, S. C., Kannry, J., Zinberg, R., Basford, M. A., *et al.* (2013) The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet. Med., 15, 761–771
112. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., *et al.* (2018) The UK Biobank resource with deep phenotyping and genomic data. Nature, 562, 203–209
113. http://www.nealelab.is/uk-biobank/. Accessed: September 1, 2020
114. Li, X., Li, Z., Zhou, H., Gaynor, S. M., Liu, Y., Chen, H., Sun, R., Dey, R., Arnett, D. K., Aslibekyan, S., *et al.* (2020) Dynamic incorporation of multiple *in silico* functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. Nat. Genet., 52, 969–983
115. Gogarten, S. M., Sofer, T., Chen, H., Yu, C., Brody, J. A., Thornton, T. A., Rice, K. M. and Conomos, M. P. (2019) Genetic association testing using the GENESIS R/Bioconductor package. Bioinformatics, 35, 5346–5348
116. Loh, P. R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., *et al.* (2015) Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet., 47, 284–290
117. Loh, P. R., Kichaev, G., Gazal, S., Schoech, A. P. and Price, A. L. (2018) Mixed-model association for biobank-scale datasets. Nat. Genet., 50, 906–908
118. Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., LeFaive, J., VandeHaar, P., Gagliano, S. A., Gifford, A., *et al.* (2018) Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet., 50, 1335–1341
119. Peng, J. and Siegmund, D. (2004) Mapping quantitative traits with random and with ascertained sibships. Proc. Natl. Acad. Sci. USA, 101, 7845–7850
120. Cirulli, E. T. and Goldstein, D. B. (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet., 11, 415–425
121. Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet., 39, 906–913
122. Li, Y., Willer, C. J., Ding, J., Scheet, P. and Abecasis, G. R. (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol., 34, 816–834
123. Browning, B. L. and Browning, S. R. (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet.,
124. Lee, S., Abecasis, G. R., Boehnke, M. and Lin, X. (2014) Rare-variant association analysis: study designs and statistical tests. Am. J. Hum. Genet., 95, 5–23
125. Van Hout, C. V., Tachmazidou, I., Backman, J. D., Hoffman, J. D., Liu, D., Pandey, A. K., Gonzaga-Jauregui, C., Khalid, S., Ye, B., Banerjee, N., *et al.* (2020) Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. Nature, 586, 749–756
126. Leung, Y. Y., Valladares, O., Chou, Y. F., Lin, H. J., Kuzma, A. B., Cantwell, L., Qu, L., Gangadharan, P., Salerno, W. J., Schellenberg, G. D., *et al.* (2019) VCPA: genomic variant calling pipeline and data management tool for Alzheimer's Disease Sequencing Project. Bioinformatics, 35, 1768–1770
127. Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., *et al.* (2019) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature, 590, 290–299
128. Kowalski, M. H., Qian, H., Hou, Z., Rosen, J. D., Tapia, A. L., Shan, Y., Jain, D., Argos, M., Arnett, D. K., Avery, C., *et al.*, (2019) Use of $>$ 100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. PLoS Genet., 15, e1008500
129. Madsen, B. E. and Browning, S. R. (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet., 5, e1000384
130. Li, B. and Leal, S. M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet., 83, 311–321
131. Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet., 89, 82–93
132. Pan, W. (2009) Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet. Epidemiol., 33, 497–507
133. Chen, H., Huffman, J. E., Brody, J. A., Wang, C., Lee, S., Li, Z., Gogarten, S. M., Sofer, T., Bielak, L. F., Bis, J. C., *et al.* (2019) Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing Studies. Am. J. Hum. Genet., 104, 260–274
134. Morrison, A. C., Huang, Z., Yu, B., Metcalf, G., Liu, X., Ballantyne, C., Coresh, J., Yu, F., Muzny, D., Feofanova, E., *et al.* (2017) Practical approaches for whole-genome sequence analysis of heart- and blood-related traits. Am. J. Hum. Genet., 100, 205–215
135. Sarnowski, C., Satizabal, C. L., DeCarli, C., Pitsillides, A. N., Cupples, L. A., Vasan, R. S., Wilson, J. G., Bis, J. C., Fornage, M., Beiser, A. S., *et al.* (2018) Whole genome sequence analyses of brain imaging measures in the Framingham Study. Neurology, 90, e188–e196
136. Werling, D. M., Brand, H., An, J. Y., Stone, M. R., Zhu, L.,

Glessner, J. T., Collins, R. L., Dong, S., Layer, R. M., Markenscoff-Papadimitriou, E., *et al.* (2018) An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. Nat. Genet., 50, 727–736

137. Li, Z., Li, X., Liu, Y., Shen, J., Chen, H., Zhou, H., Morrison, A. C., Boerwinkle, E. and Lin, X. (2019) Dynamic scan procedure for detecting rare-variant association regions in whole-genome sequencing studies. Am. J. Hum. Genet., 104, 802–814
138. He, Z., Xu, B., Buxbaum, J. and Ionita-Laza, I. (2019) A genome-wide scan statistic framework for whole-genome sequence data analysis. Nat. Commun., 10, 3018
139. Jakubosky, D., Smith, E. N., D'Antonio, M., Jan Bonder, M., Young Greenwald, W. W., D'Antonio-Chronowska, A., Matsui, H., Stegle, O., Montgomery, S. B., DeBoever, C., *et al.* (2020) Discovery and quality analysis of a comprehensive set of structural variants and short tandem repeats. Nat. Commun., 11, 2928
140. Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M. and Kamatani, Y. (2019) Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol., 20, 117
141. Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C. and Sedlazeck, F. J. (2019) Structural variant calling: the long and the short of it. Genome Biol., 20, 246
142. Jakubosky, D., D'Antonio, M., Bonder, M. J., Smail, C., Donovan, M. K. R., Young Greenwald, W. W., Matsui, H., D'Antonio-Chronowska, A., Stegle, O., Smith, E. N., *et al.* (2020) Properties of structural variants and short tandem repeats associated with gene expression and complex traits. Nat. Commun., 11, 2927
143. Chen, L., Abel, H. J., Das, I., Larson, D. E., Ganel, L., Kanchi, K. L., Regier, A. A., Young, E. P., Kang, C. J., Scott, A. J., *et al.* (2020) Association of structural variation with cardiometabolic traits in Finns. Am. J. Hum. Genet., 108, 583–596
144. Sirugo, G., Williams, S. M. and Tishkoff, S. A. (2019) The missing diversity in human genetic studies. Cell, 177, 26–31
145. Genovese, G.Friedman, D. J., Ross, M. D., Lecordier, L., Uzureau, P., Freedman, B. I., Bowden, D. W., Langefeld, C. D., Oleksyk, T. K., Uscinski Knob, A. L., *et al.* (2010) Association of trypanolytic ApoL1 variants with kidney disease in African Americans. Science, 1193032
146. Belloy, M. E., Napolioni, V. and Greicius, M. D. (2019) A quarter century of APOE and Alzheimer's disease: progress to date and the path forward. Neuron, 101, 820–838
147. Zhang, R., Wang, X., Tang, Z., Liu, J., Yang, S., Zhang, Y., Wei, Y., Luo, W., Wang, J., Li, J., *et al.* (2014) Apolipoprotein E gene polymorphism and the risk of intracerebral hemorrhage: a meta-analysis of epidemiologic studies. Lipids Health Dis., 13, 47
148. H3Africa Consortium, Rotimi, C., Abayomi, A., Abimiku, A., Adabayeri, V. M., Adebamowo, C., Adebiyi, E., Ademola, A. D., Adeyemo, A., Adu, D., *et al.* (2014) Enabling the genomic revolution in Africa. Science, 344, 1346–1348
149. Banda, Y., Kvale, M. N., Hoffmann, T. J., Hesselson, S. E., Ranatunga, D., Tang, H., Sabatti, C., Croen, L. A., Dispensa, B. P., Henderson, M., *et al.* (2015) Characterizing race/ethnicity and genetic ancestry for 100,000 subjects in the genetic epidemiology research on adult health and aging (GERA) cohort. Genetics, 200, 1285–1295
150. Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., Highland, H. M., Patel, Y. M., Sorokin, E. P., Avery, C. L., *et al.* (2019) Genetic analyses of diverse populations improves discovery for complex traits. Nature, 570, 514–518
151. Fang, H., Hui, Q., Lynch, J., Honerlaw, J., Assimes, T. L., Huang, J., Vujkovic, M., Damrauer, S. M., Pyarajan, S., Gaziano, J. M., *et al.* (2019) Harmonizing genetic ancestry and self-identified race/ethnicity in genome-wide association studies. Am. J. Hum. Genet., 105, 763–772
152. Coram, M. A., Duan, Q., Hoffmann, T. J., Thornton, T., Knowles, J. W., Johnson, N. A., Ochs-Balcom, H. M., Donlon, T. A., Martin, L. W., Eaton, C. B., *et al.* (2013) Genome-wide characterization of shared and distinct genetic components that influence blood lipid levels in ethnically diverse human populations. Am. J. Hum. Genet., 92, 904–916
153. Galinsky, K. J., Reshef, Y. A., Finucane, H. K., Loh, P. R., Zaitlen, N., Patterson, N. J., Brown, B. C. and Price, A. L. (2019) Estimating cross-population genetic correlations of causal effect sizes. Genet. Epidemiol., 43, 180–188
154. Coram, M. A., Candille, S. I., Duan, Q., Chan, K. H., Li, Y., Kooperberg, C., Reiner, A. P. and Tang, H. (2015) Leveraging multi-ethnic evidence for mapping complex traits in minority populations: An empirical Bayes approach. Am. J. Hum. Genet., 96, 740–752
155. Coram, M. A., Fang, H., Candille, S. I., Assimes, T. L. and Tang, H. (2017) Leveraging multi-ethnic evidence for risk assessment of quantitative traits in minority populations. Am. J. Hum. Genet., 101, 218–226
156. Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., Mora, S., *et al.* (2013) Discovery and refinement of loci associated with lipid levels. Nat. Genet., 45, 1274–1283
157. Panjwani, N., Wang, F., Mastromatteo, S., Bao, A., Wang, C., He, G., Gong, J., Rommens, J. M., Sun, L. and Strug, L. J. (2020) LocusFocus: Web-based colocalization for the annotation and functional follow-up of GWAS. PLOS Comput. Biol., 16, e1008336