Methods for Association Studies

5

Rebecca E. Graff, Caroline G. Tai, Linda Kachuri, and John S. Witte

Abstract

Association studies are a key approach to evaluating the relationship between genetic factors and phenotypes or traits. This chapter presents general methods for genetic association studies in unrelated humans. Topics covered include types of association studies, study design considerations, measurement of genetic information, and analytical techniques. This material provides readers with background for interpreting results from association studies and for undertaking their own studies.

5.1 Introduction

Since the first sequences of base pairs were published in the late 1960s and early 1970s (Gilbert and Maxam 1973; Wu and Kaiser 1968; Wu and Taylor 1971), our ability to investigate the human genome has advanced immensely. Genetic epidemiology largely aims to identify genetic factors that are associated with a particular phenotype or disease state. To evaluate these relationships, one essential approach used by researchers is the *genetic association study*. These studies relate germline genetic variants—or other sources of genetic variation—to some measure of phenotype, disease status, progression, and/or mortality.

Before association studies became pervasive in the journey toward deciphering the genetic basis of complex disease, *linkage analysis* was a common method for detecting genes with a major effect on phenotype (Claussnitzer et al. 2020). In the 1980s and early 1990s, many researchers undertook genetic studies that

R. E. Graff (✉) · C. G. Tai · L. Kachuri · J. S. Witte
Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, USA
e-mail: Rebecca.Graff@ucsf.edu

K. E. Lohmueller, R. Nielsen (eds.), *Human Population Genomics*,
https://doi.org/10.1007/978-3-030-61646-5_5

utilized family structures ranging from sibling pairs to large multiplex pedigrees. Such studies use families with numerous disease-affected individuals to evaluate markers spaced widely across the genome, at intervals of up to 20 million base pairs, and to examine how these markers segregate with the disease phenotype across multiple families (Botstein et al. 1980). Linkage analyses are often successful in the evaluation of rare and/or monogenetic disorders but are generally underpowered to detect genetic factors with subtle effects on complex diseases. They also have low resolution on account of the limited number of meioses from one generation to the next within families (Risch and Merikangas 1996).

Given that high-penetrance genes co-segregating in affected families have turned out to be relatively rare, association studies have become the far more common and more powerful tool to investigate genetic relationships (Claussnitzer et al. 2020). They rely on historical recombination events from millions of years of evolution and thus do not require pedigree information or controlled crosses to identify genetic variants associated with the phenotype. In addition, because most association studies leverage the phenomenon of linkage disequilibrium (LD) to localize such variants, they can detect causal loci within narrower regions and allow for genetic mapping at a finer scale than linkage studies (Xiong and Guo 1997). That is, association studies do not require the *direct* evaluation of postulated causal variants. Rather, they may utilize LD to *indirectly* evaluate genetic variants neighboring those assayed (see Chap. 2 on LD). Moreover, genome-wide association studies (GWAS) allow investigators to broadly search the genome for disease-causing variants in a manner that is relatively agnostic to previous biological knowledge.

The fundamental approach to any genetic association study is based on the following premise: compare the frequency of the genetic characteristic of interest across individuals with different values for the phenotype of interest. Consider, for example, a single-nucleotide polymorphism (SNP) with effect allele A under investigation in a standard analysis of a binary phenotype (Fig. 5.1). To determine whether or not the SNP is associated with the phenotype, one would calculate the frequency of the effect allele in cases and controls. When the frequency is greater in individuals with the phenotype than in those without it, then the effect allele is positively associated with the phenotype (as in the figure). When the opposite is true, then the effect allele is inversely associated. In GWAS, these associations are estimated for every SNP measured across the entire genome.

Genomic research traverses genetic sequence information, protein products, and the eventual expression of traits. It may also utilize a range of organisms; only one facet is the study of humans. Our focus in this chapter is on population-based genetic association studies in humans, in which data are derived from unrelated individuals. Relative to family-based association studies, population-based studies are the more common—and often more powerful—approach to the evaluation of genetic associations. In describing types of association studies (Sect. 5.2), considerations in their design (Sect. 5.3), measurement of genetic information (Sect. 5.4), and analytical techniques (Sect. 5.5), we aim to provide a basis on which readers can build their own efforts to characterize associations between genetic polymorphisms and measured phenotypes.

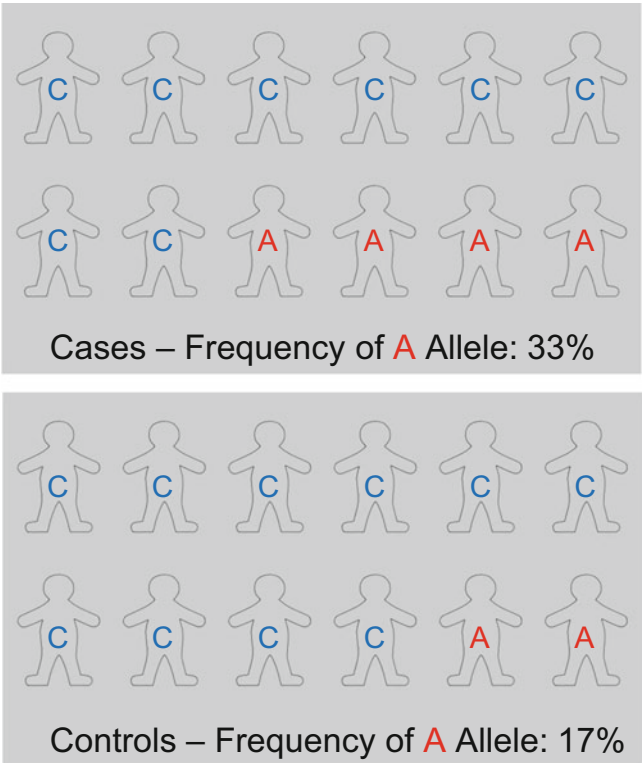

Fig. 5.1 Standard approach to a genetic association study

5.2 Types of Association Studies

Before recent technologies enabled larger-scale investigations, efforts to decipher the genetic basis of disease were predominantly supported by candidate gene studies (Claussnitzer et al. 2020). GWAS have since become an important avenue for undertaking agnostic evaluation of the association between common genetic variants and risk of disease (Claussnitzer et al. 2020). Here we describe these most common designs for genetic association studies, and Fig. 5.2 summarizes some of their differences with respect to the number of variants they can address and the sample sizes they require. In brief, as investigators shift from discovery to confirmation of associations, the number of markers investigated tends to decrease, while the number of samples should increase. Fine-mapping studies, however, do not require particularly large sample sizes as they evaluate a limited number of variants.

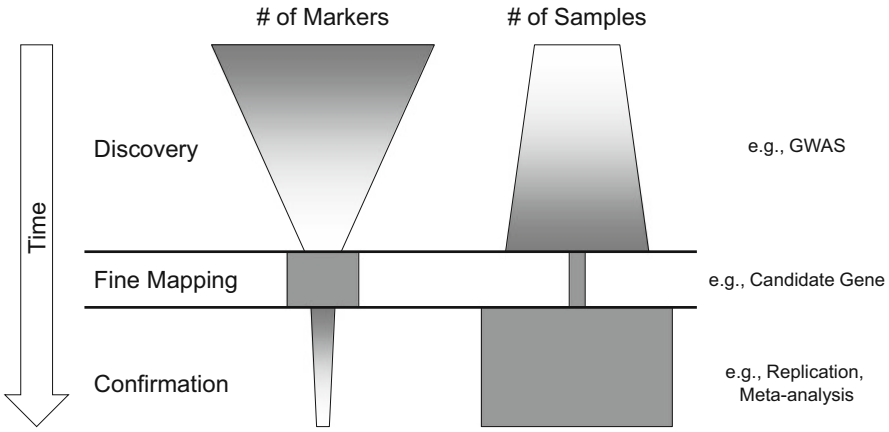

Fig. 5.2 Overview of genetic association study designs

5.2.1 Candidate Gene Studies

Candidate gene studies overcome some of the issues of linkage analysis by focusing on associations between disease and specific variants plausibly involved with the disease a priori. These studies became pervasive following the realization that genetic variants contributing to the risk of complex disease were likely to have individually weak effects (Claussnitzer et al. 2020).

Candidate gene studies generally evaluate several SNPs within a single gene under the assumption that the SNPs capture information about the underlying genetic variability of the gene (even if the SNPs are not the true causal variants). They may do so either *directly*, by evaluating postulated causal variants, or *indirectly*, by leveraging LD. Sufficiently large candidate gene studies are able to detect weak effects due to common variants, though it is important to note that they too become underpowered as variants become more rare (Risch and Merikangas 1996). In addition, their focus on particular genes means that they ignore much of the genome.

Many early candidate gene studies were underpowered, and results went largely unreplicated (Cordell and Clayton 2005). It was also unclear what should actually constitute a candidate gene. Traditionally, lists of candidate genes were compiled after an extensive manual biomedical literature review. The process to identify candidate genes then evolved to incorporate automated text-mining procedures, selection of genes belonging to specific biological pathways, and/or prioritization based on gene characteristics such as degree of conservation or proximity to known loci (Piro and Di Cunto 2012). Since GWAS have come into the picture, however, the role of candidate gene studies has become increasingly coherent as a fine-mapping approach. These studies can be targeted toward regions of the genome in which GWAS find strong hits in order to see which findings are replicable and thus more likely to be true associations.

5.2.2 Genome-Wide Association Studies

5.2.2.1 Background

Increased throughput, scalability, and speed have enabled investigators to undertake GWAS (Claussnitzer et al. 2020)—research that would have been far too complex to consider even 20 years ago. It has become possible to simultaneously measure hundreds of thousands of SNPs due to technological advances in array-based genotyping (Wang et al. 1998). The number of variants that may be assayed by these SNP arrays rapidly increased at the same time that array prices steadily decreased. At present, arrays can directly measure millions of SNPs while providing relatively high coverage of common genetic variation across the human genome (Jorgenson and Witte 2006; Lindquist et al. 2013; Nelson et al. 2013; Xing et al. 2016; Wojcik et al. 2018).

The genetic content of such arrays was facilitated by the development of technology that allows for large-scale sequencing efforts in combination with the sequencing of the human genome (Lander et al. 2001; Venter et al. 2001). Beginning in 2002, the International Haplotype Map (HapMap) Project undertook an effort to catalog the common genetic variants that occur in human beings. It was also determined that a substantial portion of this variation can be efficiently captured by a subset of "tag" SNPs via the phenomenon of LD among neighboring SNPs (Daly et al. 2001; Gabriel et al. 2002; International HapMap Consortium 2003; International HapMap Consortium 2005) and that this structure varies across ancestral populations (International HapMap Consortium 2003; International HapMap Consortium 2005; Frazer et al. 2007).

Unlike candidate gene studies, GWAS are not hypothesis-driven; they do not require a priori specification of the genes or polymorphisms that are conjectured to be associated with the phenotype of interest. Rather, they quantify DNA sequence variations from across the entire human genome in an attempt to pinpoint genetic risk factors for common diseases. In designing an array for genome-wide assessment, a primary objective should thus be to capture as much common variation in the human genome as possible.

5.2.2.2 Multistage Study Designs

When GWAS first became popular, the high cost of SNP arrays and necessity for large sample sizes (to achieve sufficient statistical power to detect the anticipated modest associations among hundreds of thousands of SNPs) (Witte et al. 2000) motivated the development and use of *multistage* GWAS designs (Thomas et al. 2005). Decreasing SNP array costs have made multistage designs for GWAS less essential, but we choose to briefly describe them for two reasons: (1) an overview is important for understanding historical studies, and (2) as we move into the post-GWAS era, next-generation sequencing of entire genomes may be sufficiently expensive to once again make multistage designs relevant.

In the initial discovery stage of a multistage design, a subset of the study sample is genotyped using genome-wide SNP arrays. Then the most strongly associated

SNPs are genotyped with a less expensive genotyping platform in the remaining samples. The procedure prioritizes the most promising SNPs for evaluation in additional stages and can pinpoint associated regions for fine mapping. The optimal division of samples across stages depends on a number of factors, but in general, the most efficient approach entails the inclusion of approximately one-third to one-half of the samples in the initial stage and the remaining samples in follow-up stages (Skol et al. 2006, 2007). The number of noteworthy SNPs that should be tested depends on the sample sizes in the respective stages, the number of false-negative results that one is willing to accept, and whether or not one wishes to incorporate SNP information (e.g., proximity to the nearest gene or likelihood of being functional) (Chen and Witte 2007; Roeder and Wasserman 2009; Roshan et al. 2011; Thomas et al. 2009). Ideally, at least 1% of the first stage SNPs should be typed in the second stage (Skol et al. 2006). One must also decide whether the early follow-up stages should be treated as part of a replication or joint analysis.

5.2.2.3 Limitations

Despite their numerous strengths, GWAS carry several notable limitations. First, it is important to note that most variants discovered via GWAS are only associated with, and not causal for, disease. Even when an association is real and statistically reproducible in other datasets, another untyped variant in LD with the associated SNP may still be the causal variant. Determining the factors underlying results can be extremely challenging and require separate fine-mapping and mechanistic studies. That many of the associations detected to date are not in gene regions can make the findings yet more complicated (Buniello et al. 2019). These issues limit our understanding of the biological basis of results and our ability to implement preventive or therapeutic measures.

Second, findings from GWAS thus far account for only a limited amount of disease heritability (Maher 2008; Nolte et al. 2017). Most SNPs detected by GWAS show a small magnitude of effect. That said, as sample sizes for GWAS are increasing, studies are detecting and replicating a larger number of trait-associated variants. That we are now also able to examine essentially the entirety of common variation across the genome (at least indirectly) allows us to explain an increasing proportion of heritability. So too does our ability to assess the contribution of rare variants. The polygenic model of heritability is becoming increasingly accepted; many risk variants with small effect sizes are thought to underlie disease risk.

Finally, GWAS have not yet sufficiently distinguished between individuals with low- and high-risk disease. In general, screening tests based on SNPs detected by GWAS to date may have low positive (and negative) predictive value for disease and thus limited utility in a diagnostic setting (Kraft et al. 2009; Ware 2006). As more SNPs are discovered, however, combining them into polygenic risk scores (PRS) efficiently summarizes individuals' genetic susceptibility profiles, thereby improving phenotypic prediction (Torkamani et al. 2018). PRS have the potential to personalize risk estimates and improve the discriminatory ability of screening tests (Mavaddat et al. 2019; Toland 2019). For example, a 2015 study created a risk score of 105 SNPs that was strongly associated with prostate cancer risk among

non-Hispanic whites (P value: 1.0×10^{-211}) (Hoffmann et al. 2015). More recently, a PRS for breast cancer based on 313 variants demonstrated strong predictive performance (AUC = 0.630) and identified 19% of women who could be eligible for early screening at age 40 (Mavaddat et al. 2019). Still, few individuals will carry large numbers of risk alleles from GWAS, though essentially all individuals will carry some risk alleles. Screening for them in the general population is thus unlikely to be cost-effective, unless individuals receive genome-wide evaluations. In addition, predictive models may have worse performance in ancestral populations other than those in which the models were discovered, because effect size estimates will be diluted when SNPs in populations with one set of LD patterns (e.g., Europeans) are applied to populations with a different set of LD patterns (e.g., African Americans) (Carlson et al. 2013). Note also that justification for genetic testing additionally depends on the existence of effective interventions.

5.2.3 Mendelian Randomization

In some instances, genetic variation can be leveraged toward evaluating causal relationships between exposures and outcomes that may be challenging to investigate in traditional observational studies. By using a genetic predictor of exposure as an instrumental variable, *Mendelian randomization* circumvents issues of confounding and reverse causation that often afflict epidemiological studies. While the method has been around for several decades (Gray and Wheatley 1991; Katan 1986; Smith and Ebrahim 2003), its use has exploded with the ever-increasing discovery of trait-associated variants and modern statistical methods for high-dimensional genetic data. In general, its implementation requires the identification of a set of genetic variants that is predictive of the exposure of interest followed by the performance of instrumental variable analyses (Burgess et al. 2013; Pierce and Burgess 2013).

As with all instrumental variable approaches, Mendelian randomization is premised on three assumptions: (1) the genetic instrument is associated with the exposure, (2) the genetic instrument shares no common causes with the outcome, and (3) the genetic instrument only affects the outcome through exposure. The first assumption is easily satisfied by selecting genetic variants that are strongly associated with the exposure of interest, such as those reaching genome-wide significance. The second assumption can be at least partially verified by assessing associations between genetic instruments and known confounders. The third assumption, however, cannot be substantiated empirically. Nevertheless, sensitivity analyses can help evaluate the consistency and robustness of observed results (Bowden et al. 2017; Haycock et al. 2016).

5.2.4 Transcriptome-Wide Association Studies

Among the more recent methodological developments in genetic association studies is the *transcriptome-wide association study* (TWAS) (Gamazon et al. 2015; Gusev et al. 2016). Without relying on directly measured expression levels, TWAS aim to identify genes associated with complex traits. By using an external reference set of individuals with genetic and transcriptomic data, one can impute gene expression levels in the target study population and evaluate associations with the outcome. Extensions of this approach allow for implementation with summary statistics rather than individual-level data, making TWAS an increasingly popular study design (Barbeira et al. 2018). Furthermore, because associations at the gene expression level often have clearer functional interpretations than associations with individual risk variants, TWAS have the potential to offer insights distinct from those offered by GWAS. Testing for associations with genes rather than SNPs also reduces the multiple testing burden, thereby improving statistical power for discovery. TWAS are, however, limited by the comprehensiveness of gene expression reference panels both across different tissues and for populations of non-European ancestry. Furthermore, although the genetic architecture of gene expression allows for reasonable imputation accuracy, gene expression can also be influenced by non-genetic, external factors.

5.2.5 Replication and Meta-analysis

Findings from a single genetic association study are not generally sufficient to instill confidence in results. Rather, results should be validated in independent samples and combined with other studies to bolster sample size.

5.2.5.1 Replication

Early genetic association studies frequently yielded results that failed to reproduce in independent samples (Hirschhorn et al. 2002; Ioannidis 2006; Ioannidis et al. 2001; Lohmueller et al. 2003). Why the surfeit of false positives? Historically, studies of candidate markers or genes often had small sample sizes, inappropriate thresholds for statistical significance, and/or low prior probabilities of true associations (Chanock et al. 2007; Hirschhorn and Altshuler 2002; Ioannidis 2005; Manolio et al. 2008; Mutsuddi et al. 2006; Wacholder et al. 2004). Even now, investigators are conscious of "winner's curse," whereby the effect estimates from initial discovery studies are consistently biased upward (Lohmueller et al. 2003; Goring et al. 2001; Huang et al. 2018). They are generally more attentive to winner's curse for genetic association studies than for other epidemiological investigations because the former most often test a large number of exposures. The gold standard for substantiating results from genetic association studies has thus become *replication* in independent samples. Replication has become important (and essentially required for publication) to externally validate the credibility of genetic associations.

Studies designed for the purposes of replication should ensure that sample sizes are sufficiently large to detect associations of the hypothesized magnitudes. In fact, sample sizes should ideally be larger than those of the initial study so as to account for overestimation in the original sample (unless one only wishes to replicate a limited number of variants). The larger the sample, the better success replication studies will have in reproducing results from and identifying false positives generated by the initial study. Replication studies should also evaluate the same ancestral population as the discovery study and, ideally, the same genetic variant with respect to the same definition of phenotype. Successful replication then entails finding the same direction of association (for the same effect allele) at a predetermined threshold for statistical significance. What that threshold should be is somewhat controversial; some investigators expect that associations be replicated at a genome-wide significance level, whereas others apply a less conservative threshold based on evaluating a smaller number of variants in the replication sample. Still others are not as concerned with the statistical significance of the replication association as they are with the significance of the joint analysis of discovery and replication.

In some cases, studies are not designed exclusively for the purposes of replication. Rather, colleagues may help one another replicate their strongest results by looking them up in independent, existing "discovery" studies. Once results are confirmed in the original target populations, investigators may also choose to evaluate associations in populations of varying ancestries. Results that replicate from these studies are often said to *generalize*, meaning that the effect is relevant to multiple human populations. In contrast to replication, studies conducted for generalization should draw from an ancestral population different from the discovery population.

It should be noted that while replication has become standard practice to corroborate genetic associations, it may not be as necessary as it once was. As genetic association studies have become increasingly sizeable and larger numbers of markers have been genotyped in large replication samples, the statistical power to detect modest effects has substantially increased. As a result, the potential for winner's curse has decreased. Still, replication inspires confidence in findings and remains customary for genetic association studies.

5.2.5.2 Meta-analysis

Results from multiple studies or even multiple stages of the same study can be combined into a single result via *meta-analysis*. Meta-analytical methods synthesize results from analyses that examine the same hypothesis without accessing individual-level data (as mega-analytical studies would). In doing so, they considerably boost the sample size and power for examining the hypothesis and thus may achieve a more precise estimate of the association of interest. Several software packages are available for the implementation of meta-analysis for GWAS, among which are METASOFT (Han and Eskin 2011), METAL (Willer et al. 2010), GWAMA (Magi and Morris 2010), PLINK (Purcell et al. 2007), and GenABEL/MetABEL (Aulchenko et al. 2007). Available features in most of these packages were summarized in a side-by-side comparison (Evangelou and Ioannidis

2013). In addition, there exist tools to meta-analyze results from populations of varying ethnicities (Hong et al. 2016; Morris 2011).

Meta-analytical methods can also be used to discover novel genetic loci with pleiotropic effects and to explore associations across phenotypes or disease subtypes. Association analysis based on subsets (ASSET) is a flexible meta-analysis framework that can evaluate associations for a given SNP across phenotypes and identify the combination of associated traits that maximizes the overall test statistic (Bhattacharjee et al. 2012). In addition to boosting power in the presence of heterogeneity, attractive features of ASSET are its ability to account for sample overlap across contributing studies and its internal correction for the multiple tests required by the subset search. ASSET has been applied to a number of traits, among which are multiple cancers (Fehringer et al. 2016) and immune-related diseases (Marquez et al. 2018; Zhu et al. 2018).

To conduct a rigorous meta-analysis, all studies should be subject to a standard quality control procedure that determines which SNPs are included in each study. It is a fundamental assumption of meta-analysis that the studies provide independent information, so it is also critical to ensure that there not be any overlap in the samples included from each study. In addition, the design of each study incorporated into a meta-analysis should ideally be similar; the measurement of covariates and phenotypes should be analogous, analytic procedures should be comparable, and covariate adjustment should be standardized (Zeggini and Ioannidis 2009). It is also important that all studies report results using the same reference allele and mode of inheritance. Imputation is often required to ensure that all studies in a meta-analysis offer data about the same SNPs (discussed further below).

The most common method to estimate an average effect across studies is fixed-effects modeling that weights each study effect based on its inverse variance. Mixed-effects models may also be used when there is substantial heterogeneity of effects across studies; their random effect parameters can help account for the heterogeneity. Regardless of the model selected for meta-analysis, it is important to quantify the differences across studies, particularly given that it is rare that studies perfectly fulfill the stringent criteria for meta-analysis. The most commonly used measures to do so are the Q statistic and I^2 index (Evangelou and Ioannidis 2013; Huedo-Medina et al. 2006; Panagiotou et al. 2013).

5.3 Design of Association Studies

The first step toward obtaining meaningful results from any genetic association study is designing it effectively. Investigators must always define appropriate phenotypes, designate a valid study population, and ensure a sufficient sample size. In this section, we outline some of these key elements that should be contemplated in conceiving new studies.

5.3.1 Quantitative Versus Qualitative Traits

There are two primary classes of phenotypes that one might wish to evaluate with a genetic association study, namely *quantitative* and *qualitative* (most often binary case-control). Quantitative traits generally have higher statistical power to detect genetic effects, and the interpretation of effects is often more straightforward. For a genetic variant that influences a quantitative trait, each allele or genotype class may be interpreted as affecting a unit change in the level of the trait. Alternatively, one might opt to study subjects at the extremes of a quantitative trait distribution to maximize power per genotyped individual for detecting associations (Huang and Lin 2007; Guey et al. 2011).

Many diseases do not have meaningful or well-established quantitative measures. In such scenarios, individuals are commonly classified as either affected or unaffected, and studies most often implement a case-control design. Frequencies of genetic variants observed in cases are compared with those observed in controls in order to evaluate whether an association between genes and disease exists. It is important to note that for a complex phenotype (e.g., metabolic syndrome) or one that is diagnosed over a long period (e.g., Alzheimer's disease), there may be some measurement error in dichotomizing individuals as cases or controls. Still, many association studies of binary traits have been extremely successful in detecting genetic variants correlated with disease (see Chap. 7 on what we have learned from GWAS).

5.3.2 Subject Selection

The most important facet of subject selection is ensuring that subjects are representative of their source population (Wacholder et al. 1992). For a case-control study in which cases with a particular disease are compared to unaffected controls, this means that controls should be individuals who, if diseased, would be cases. Whenever controls are not selected to represent the source population of the cases, spurious associations may result. Consider, for example, a scenario in which controls are selected from a different ancestral population from cases. In such a circumstance, control subjects might have fundamentally different allele frequencies in the SNPs of interest relative to cases. As a result, one is likely to find associations between these SNPs and disease even in the absence of true associations. This particular bias is called population stratification and can, if unaccounted for, confound GWAS. We will discuss methods to control for population stratification later in this chapter.

Cases are commonly recruited from a specific population, hospital, or disease registry. Depending on the study design, controls may either be unrelated (population-, hospital-, or registry-based) or family members of the cases. Controls are also commonly matched to cases with respect to ancestry, age, and sex.

Even without rigorous control selection, many GWAS have been successful at detecting highly replicated variants. Due to the high cost of subject recruitment and

genotyping, investigators sometimes use genotype information from controls who have been recruited into prior studies and that has been made publicly available to researchers (e.g., via the database of genotypes and phenotypes (dbGaP)) (Luca et al. 2008; Burton et al. 2007; Paltoo et al. 2014). The inclusion of controls from public databases can also increase statistical power without affecting costs (Ho and Lange 2010). The potential bias arising from the use of such "convenience" or "public" controls is mitigated by the low measurement error in SNP genotyping, the absence of recall bias when studying inherited variants, large sample sizes, stringent criteria for statistical significance, and rigorous replication of findings. Nevertheless, the use of convenience controls may result in the confounding of associations due to population stratification (discussed further below). One should thus address the bias analytically with genetic information (Devlin and Roeder 1999; Price et al. 2006; Pritchard and Rosenberg 1999; Mitchell et al. 2014). One must also consider the potential for batch effects due to differences in genotyping quality control procedures and phenotype misclassification in individuals not thoroughly screened for common diseases. These issues can be assessed in small subsets of the sample by comparing genotype concordance in re-genotyped individuals or by conducting sensitivity analyses of the phenotype (Mitchell et al. 2014). Restricting the use of controls to those genotyped on the same platform and from the same genetic ancestry as cases may prevent or reduce these biases (Sinnott and Kraft 2012).

5.3.3 Sample Size

As with any study, it is critical that genetic association studies include a sufficiently large sample size to ensure good statistical power. The power of a study depends on the unknown frequency and effect size of the causal genetic variant(s) for which one is searching. Whenever SNPs in LD with the true causal variant are genotyped rather than the causal variant itself, power is reduced; the sample size required will be inflated proportionally to the inverse of the correlation between the genotyped and causative markers. There undoubtedly exist genetic variants that have a small effect on disease but that have not been detected due to insufficient sample size. In general, it is rare for successful GWAS to include fewer than 1000 cases and 1000 controls, and many include substantially larger numbers of individuals. Among the largest GWAS conducted to date have investigated smoking initiation ($n = 1{,}232{,}091$) (Liu et al. 2019), educational attainment ($n = 1{,}131{,}881$), blood pressure traits ($n = 1{,}006{,}863$) (Evangelou et al. 2018), and risk tolerance ($n = 975{,}353$) (Karlsson Linner et al. 2019).

5.4 Measurement of Genetic Information

Accurate measurement of genetic information is yet another crucial component of a reliable genetic association study. The study design is likely to inform the appropriate category of measurement broadly, but each method requires nuanced

decision-making to achieve the highest quality and most relevant results. Here we discuss some of the most common tools utilized to measure genetic information and some considerations to contemplate in deciding on methods.

5.4.1 Common Variants

SNP genotyping arrays are currently the most pervasive tool utilized for evaluating genetic information. GWAS in particular typically employ microarray-based tag SNP genotyping techniques that capture common variation in the human genome. The arrays for early GWAS generally contained between 100,000 and 500,000 variants identified in databases such as HapMap. More current chips include approximately one million or more variants. No matter the set of variants, the array is then typed in a specific set of individuals. The arrays have generally been designed to measure variants at or above a minor allele frequency of 5%, though they may even miss some common variation (Jorgenson and Witte 2006). More recently, however, microarrays have been designed to detect variants down to a minor allele frequency of 1% (Hoffmann et al. 2011a, b). The platforms most often come from one of two companies: Illumina (San Diego, CA) or Affymetrix (Santa Clara, CA; now owned by Thermo Fisher Scientific) (Hindorff et al. 2009).

One consideration in designing a chip for assessment is determining the set of SNPs required to capture common variation in the population of interest. Consider, for example, the number of SNPs required to capture variation across the genome for African versus European populations. When the latter emigrated from Africa, they experienced a bottleneck that reduced the population size and resulting genetic variation. They thus have more LD than the former (see Chap. 8 on human demographic history). As a result, the chip used for a study of an African population requires more SNPs to obtain the same overall genomic coverage.

5.4.2 Rare Variants

The recent considerable expansion of the human population and negative selection of deleterious alleles over time have resulted in low allele frequencies for many disease-causing variants. Consequently, rare variants with substantial effects may remain untyped by standard genotyping assays. In addition, whole-genome sequencing is not generally cost-effective for the evaluation of rare variants, because the sample sizes required for association studies are normally much too large. More effective methods to identify rare disease-causing variants involve utilizing *exome sequencing* or *exome genotyping arrays* to investigate coding variation, even though these approaches ignore potentially important parts of the genome.

Sequencing and capture technologies are now able to accurately determine the sequence of nearly all protein-coding variants in humans (Choi et al. 2009; Gnirke et al. 2009; Ng et al. 2009; Teer and Mullikin 2010). They allow researchers to detect and genotype variants found in particular individuals without requiring that

the variants be previously ascertained and included on genome-wide genotyping arrays. Originally, such technologies were most often utilized for (and successful at) identifying genetic contributors to many Mendelian disorders. More recently, the technologies have been leveraged to assay the exome as a popular approach for evaluating associations between rare variants and complex phenotypes. Exome sequencing selects the entire set of human exons as the sequencing target (Gnirke et al. 2009; Hodges et al. 2007). In contrast, exome arrays concentrate on a fixed set of variants. Regardless of the platform used to evaluate the exome, the variants assessed have functional implications that are *relatively* easy to derive. Still, due to reduced penetrance, sample sizes required for detecting associations with complex traits are generally larger than those required for the evaluation of Mendelian disorders. As exome sequencing costs have come down, however, it has become increasingly feasible to conduct well-powered studies. It is just important that they increase sample sizes in proportion to the rarity of causal variants.

5.5 Data Analysis

Upon completing the measurement of both genotype and phenotype, one must consider the appropriate methods for the analysis of the data. In addition to thinking through the statistical methods that should be applied, one must also assess the data for their quality, consider covariates that should be accounted for, and potentially incorporate information from external sources. Below we identify some key considerations for analyzing genetic association studies.

5.5.1 Quality Control

Before analyzing the data from any study of genetic association, it is imperative that the genotyping be subject to a number of quality control checks. Samples that come from various sources may be processed in different ways or measured at different times, which can result in systematic differences across batches. One must also evaluate the proportion of samples that are successfully genotyped and test for Hardy–Weinberg equilibrium. Issues with these metrics could indicate genotyping problems that affect all of the SNPs in the sample. As such, one should remove SNPs that fail predefined quality standards from further consideration. Using SNP genotypes and external LD information on the underlying structure of a genetic region (e.g., from the TOPMed imputation reference panel), one can impute some of the untyped variants and variants that fail quality control. Doing so allows for a more thorough and powerful evaluation of potential associations across the genome (Huang et al. 2009; Marchini and Howie 2008; Marchini et al. 2007).

Note that sequencing (as opposed to genotyping) also requires appreciable quality control efforts, but their description is beyond the scope of this chapter.

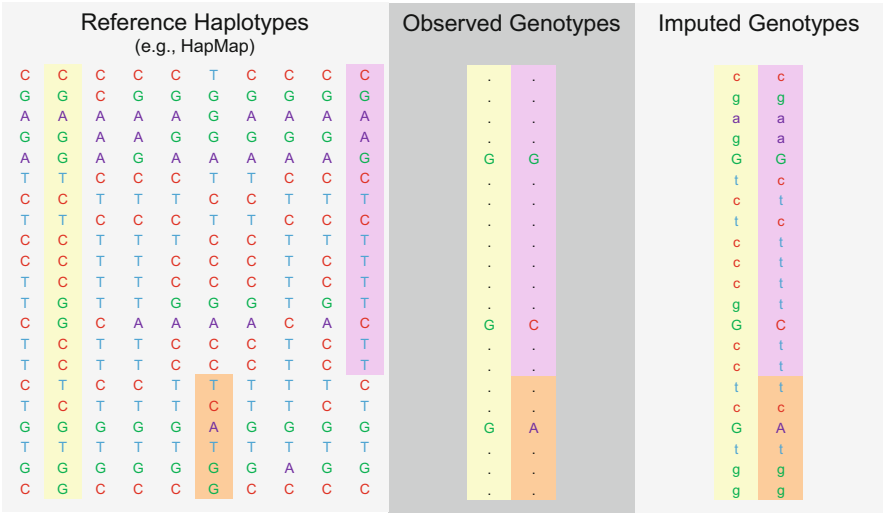

Fig. 5.3 Schematic of genotype imputation modified from Li and colleagues (2009). Observed genotypes are compared to haplotypes in a reference panel to fill in unobserved genotypes

5.5.2 Data Imputation

Above, we discussed the exploitation of LD to capture common variation in the human genome without directly genotyping every SNP. Using LD patterns and haplotype frequencies (e.g., from the 1000 Genomes Project or TOPMed imputation reference panel), it is possible to *impute* data for SNPs that are not directly genotyped (Fig. 5.3) (Li et al. 2009). First, directly genotyped markers are compared to a variant-dense reference panel that contains haplotypes drawn from the same population as the study sample. A collection of shared haplotypes is then identified, and genotypes missing from the study panel can be inferred from the matching reference haplotypes. Because the study sample may match multiple reference haplotypes, one might opt to give a score or probability for an imputed marker rather than a definitive allele. In such scenarios, uncertainty can be incorporated into the analysis of imputed data, typically with Bayesian methods (Marchini et al. 2007). A less computationally intensive method involves pre-phasing, in which haplotypes are first estimated for every individual, followed by genotype imputation using the reference panel for each haplotype. This method also makes it faster to execute the imputation step with different reference panels as they become updated, since the genotypes need only be phased once and the estimated haplotypes are saved for future use (Howie et al. 2012). Note that imputation is especially useful for meta-analyzing results across studies that rely on different genotyping platforms.

5.5.3 Analysis of Common Variants

The conventional analysis plan for genome-wide data is a series of statistical tests that examine each locus independently for an association with the phenotype of interest. In the case of binary phenotypes, the simplest approach to these tests is *contingency tables* of counts of disease status by either genotype or allele count (Clarke et al. 2011). One can use a series of χ^2 or Cochran-Armitage trend tests to evaluate the independence of the rows and columns of each table. More commonly used than contingency tables is a regression approach, linear for quantitative traits and logistic for case-control traits, with categorical predictor variables for the genotypes. Regression models are generally favored because they allow adjustment for covariates, such as principal components (PC) of genetic ancestry.

For quantitative phenotypes, one can use a linear regression model framework, $\mathbf{y} = \alpha + \mathbf{x}\boldsymbol{\beta} + \mathbf{c}\boldsymbol{\gamma}$, to model association with phenotype, where $\mathbf{x}$ is a matrix (or vector) of genotypes, $\mathbf{c}$ is a matrix of covariates such as ancestral PCs, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the corresponding vectors of regression coefficients. For binary phenotypes, one would implement a logit link function. Regardless of the link function for the outcome, the $\boldsymbol{\beta}$s are the parameters of interest, and we can test the null hypothesis of no association between $\mathbf{x}$ and $\mathbf{y}$, H_0: $\boldsymbol{\beta} = \mathbf{0}$.

In addition to maximum likelihood-based regression models, linear mixed models (LMM) have become increasingly popular in GWAS, motivated by the computational challenges of analyzing datasets with large numbers of subjects and genetic variants. One of the most attractive features of LMM is their ability to control for confounding due to population structure by directly modeling relatedness among individuals, thereby improving power relative to standard GWAS with adjustment for PCs (Yang et al. 2014; Zhang et al. 2010; Zhou and Stephens 2012). The most recent addition to this class of methods, BOLT-LMM, adopts a Bayesian perspective by imposing a prior distribution on SNP effect sizes. It does not require computing or storing a genetic relationship matrix, which substantially reduces computational time compared to other methods (Loh et al. 2015). However, applying BOLT-LMM to case-control data can be problematic since genetic effects are estimated on the observed 0–1 scale rather than the odds ratio scale. As a result, transformations are required to make LMM-based results for binary traits comparable with logistic regression (Lloyd-Jones et al. 2018).

Regardless of the structure of the phenotypic data, there are several ways in which one might code the genotype data for association tests; the choice made should reflect the assumed mode of inheritance and genetic effect. In GWAS, the genotypes are usually coded as 0, 1, or 2 to reflect the number of effect alleles. This coding assumes that each additional copy of the variant allele increases the phenotype or log risk of disease by the same amount. The approach is fairly robust to incorrect assumptions about the mode of inheritance and has reasonable power to detect both additive and dominant genetic effects. It may, however, be underpowered if the true mode of inheritance is recessive (Lettre et al. 2007). If one believes the mode of inheritance to be recessive, then one may use an alternative genetic model

that assumes that two copies of the risk allele are necessary to result in phenotype susceptibility. For such models, the heterozygote and wild-type homozygote are collapsed into a single category, and the genotypic exposure is treated as binary. The genotypic exposure is also treated as binary for models that assume a dominant mode of inheritance, but the heterozygote and mutant homozygote are collapsed separately from the wild-type homozygote. These dominant models assume that a single copy of the risk allele is expected to result in phenotype susceptibility. Both recessive and dominant models force heterozygotes to have the same risk or mean phenotype as one of the homozygotes. If investigators do not have an a priori hypothesis as to the mode of inheritance, they may choose to assess several different genetic models. Doing so, however, requires additional corrections for multiple testing. Another option when there is no a priori hypothesis would be to avoid any assumptions about how the risk for heterozygotes compares with both homozygotes. In such codominant models, maintaining the three distinct genotype classes requires two degrees of freedom, while the other models require only one, thereby making the latter more attractive if the genetic effect approximately follows one of their modes of inheritance.

To visualize the results from analyses of common variants, particularly from GWAS, investigators often generate *Manhattan plots* (e.g., Fig. 5.4). The *x*-axis of these scatter plots is a chromosomal position, and the y-axis shows the *P* value for association with the phenotype. Each point on the plot represents a single SNP, and the height of each point depicts the strength of association between the SNP and the phenotype. Manhattan plots with genome-wide significant results often exhibit clear peaks where SNPs in LD show comparable signals. Those with points seemingly scattered at random should be viewed with some skepticism.

Quantile-quantile (Q-Q) plots are another important visualization tool to evaluate potential bias or quality control problems in GWAS results (e.g., Fig. 5.5). These plots present the expected distribution of association test statistics for all SNPs on the *x*-axis against the observed values on the *y*-axis. Deviation from the $x = y$ line suggests a systematic difference between cases and controls across the whole genome (such as that which might occur in the presence of population stratification). One should rather hope to see the plotted points fall on the $x = y$ line until a curve at the very end representing any true associations.

5.5.4 Analysis of Rare Variants

While many common variants that contribute to complex diseases have been identified, the majority of variants contributing to disease susceptibility have yet to be described. Rare variants, which are unlikely to be captured by GWAS focusing on common SNPs, undoubtedly contribute to phenotype as well (Frazer et al. 2009; Gorlov et al. 2008). Unfortunately, detecting associations between individual rare variants and phenotypes can be difficult, even with large sample sizes; the low frequency of rare variants in the population results in low power (Gorlov et al. 2008; Altshuler et al. 2008; Li and Leal 2008). To increase power, researchers have

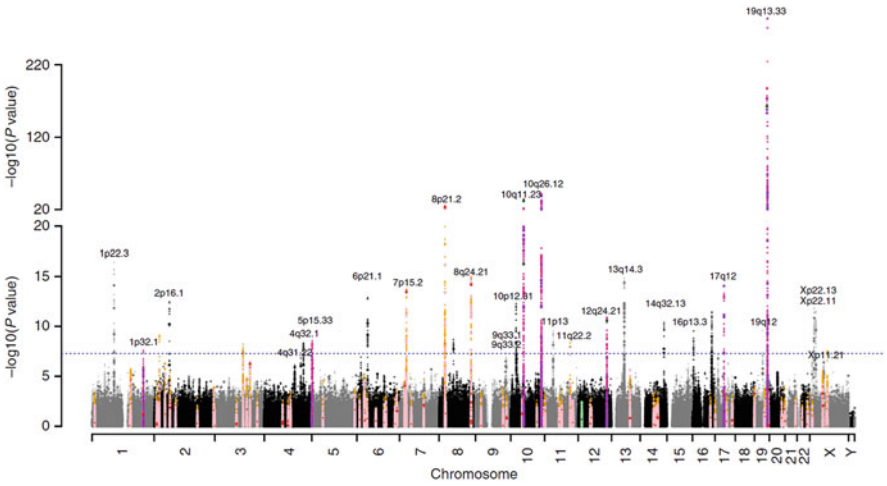

Fig. 5.4 Example Manhattan plot from a recent GWAS of prostate-specific antigen (PSA) levels (Hoffmann et al. 2017). *P* values are for variant associations with log-transformed PSA levels, adjusted for age and ancestry PCs using a linear regression model. Black and grey peaks indicate novel findings. Dark purple and magenta indicate previously reported PSA level-associated genotyped and imputed hits, respectively, and light purple and magenta indicate those within 0.5 Mb of previously reported hits that were replicated at genome-wide significance. Dark pink and red points denote previously reported prostate cancer SNPs genotyped and imputed, respectively, and pink and orange indicate those within 0.5 Mb of previously reported prostate cancer SNPs genotyped and imputed. Dark blue and green points denote the previously reported genotyped and imputed, respectively, SNPs associated with PSA levels only (and not prostate cancer), and light blue and green those within 0.5 Mb previously reported hits. Circles denote genotyped SNPs, and triangles represent imputed SNPs

developed a number of methods to evaluate the collective effect of multiple rare variants within and across genomic regions (Li and Leal 2008; Asimit and Zeggini 2010; Morgenthaler and Thilly 2007; Larson et al. 2017; Santorico and Hendricks 2016).

The two primary approaches to rare variant analyses are burden tests (Morgenthaler and Thilly 2007; Asimit et al. 2012; Li et al. 2012; Madsen and Browning 2009; Morris and Zeggini 2010; Zawistowski et al. 2010) and variance component tests (Neale et al. 2011; Pan 2009; Wu et al. 2010; Wu Michael et al. 2011). The simplest burden approach collapses rare variants into a single group by counting individuals who possess *at least* one rare variant in the genomic region under study and then tests for frequency differences across phenotypic groups. A limitation of burden tests is their assumption that all alleles have the same direction of effect; in the presence of both protective and deleterious variants, power can be substantially reduced. Burden tests also have reduced power in regions with a large number of non-causal variants. These limitations are addressed by variance component tests, the most common of which is the sequence kernel association test (SKAT) (Wu Michael et al. 2011). SKAT aggregates genetic information across variants using a

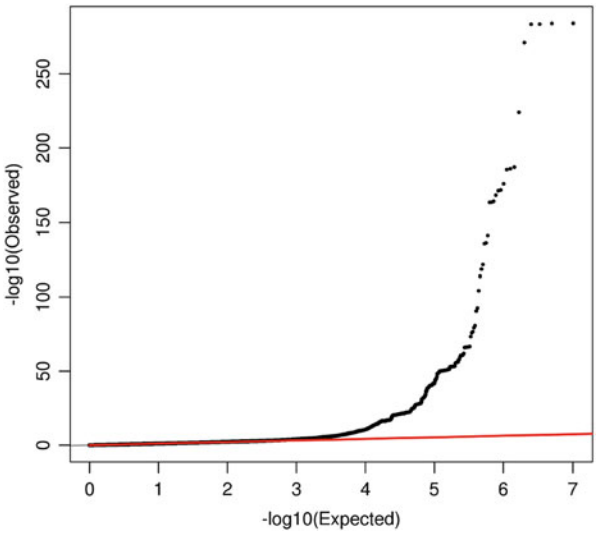

Fig. 5.5 Example Q-Q plot of results from a recent GWAS of PSA levels (Hoffmann et al. 2017). Because there were so many positive results, we see a substantial curve representing true associations at the end

kernel function. To test the null hypothesis that a set of rare variants does not impact the phenotype, one can compute the variance component score statistic Q, which is equal to $(\mathbf{y} - \overline{\mathbf{y}})'\mathrm{K}(\mathbf{y} - \overline{\mathbf{y}})$, where $\overline{\mathbf{y}}$ is the predicted mean of $\mathbf{y}$ under the null hypothesis of no association, adjusting for covariates $\mathbf{c}$, and the kernel K is an $n \times n$ matrix that defines the genetic similarity among individuals. The SKAT framework has expanded to create a family of tests accommodating a range of scenarios (Wu et al. 2013; Lee et al. 2012), including combination tests for common and rare variants (Ionita-Laza et al. 2013), time-to-event models (Chen et al. 2014), and multiple phenotypes (Dutta et al. 2019). Most recently, rare variant tests based on generalized linear mixed models have been proposed (Chen et al. 2019), as have flexible sliding-window approaches that account for LD structure (Li et al. 2019).

5.5.5 Incorporating External Information into Association Study Analyses

5.5.5.1 Gene Set Analysis

Analyses of data from GWAS can test multi-marker combinations of SNPs. Such *gene set analyses* can be used to determine whether groups of functionally related genes defined a priori are associated with a phenotype. Given that complex disease may result from a sum of changes across genes in a biological pathway, it makes sense to evaluate genes in a pathway as a set. These analyses aim to identify gene

sets with coordinated expression changes that would not be detected by single variant methods (e.g., testing one SNP at a time).

Gene set analysis generally consists of four steps: (1) determining the gene sets to be tested, (2) selecting an appropriate set of hypotheses, (3) carrying out corresponding statistical tests, and (4) evaluating the statistical significance of said tests. Regarding step 2, there are three standard null hypotheses against which investigators most often test (Dinu et al. 2009; Nam and Kim 2008; Tian et al. 2005). The first is the competitive null hypothesis, which states that the genes in a set have the same association with phenotype as genes in the rest of the genome. The second is the self-contained null hypothesis, which asserts that the genes in a set are not associated with the disease phenotype. The third option, the mixed null hypothesis, declares that none of the sets under consideration is associated with the disease.

The set of hypotheses selected largely informs the tests that should be used for analysis. To obtain a test statistic for the competitive null, a measure of association should first be computed for each gene and the phenotype of interest. For genes in a given set, the association measures should then be combined. To evaluate the statistical significance of the combined test statistic, it should be compared against the distribution under the null hypothesis, obtained by permuting the association measures (Tian et al. 2005). The procedure is similar to obtaining a test statistic for the self-contained null hypothesis, but the null distribution should be generated by permuting the phenotypes across samples (Tian et al. 2005). Regardless of the test statistic, larger magnitudes indicate increasing significance, and the sign indicates the direction of change in phenotype.

The gene sets that are deemed significant are likely to depend on the choice of methods implemented to analyze gene set associations (Elbers et al. 2009a, b). Oftentimes, gene set analyses lack sufficient statistical power to detect gene sets consisting of markers only weakly associated with disease, and they are prone to several sources of bias, among which are gene set size, LD patterns, and overlapping genes (Elbers et al. 2009b; Cantor et al. 2010; Hong et al. 2009; Wang et al. 2011; Sun et al. 2019). It is important to consider and address all of these limitations when interpreting results from gene set analyses.

5.5.5.2 Hierarchical Modeling

Hierarchical modeling leverages the abundance of bioinformatic data characterizing the structural and functional roles of common variants analyzed for GWAS (Cantor et al. 2010; Wang et al. 2010). It aims to incorporate a priori biological knowledge via Bayesian methods (Cardin et al. 2012), stabilize effect and variance estimation of SNP associations (Aragaki et al. 1997; Evangelou et al. 2014), and improve the selection of SNPs for further evaluation (Witte 1997; Witte and Greenland 1996). It also addresses issues of multiple comparisons in analyses of GWAS. Rather than perform traditional single-locus analyses, hierarchical models output "knowledge-based" estimates of SNP effects, thereby improving the ranking of results from GWAS.

The hierarchical modeling approach uses higher-level "priors" to model the parameters of interest as random variables with a joint distribution that is a function of hyperparameters (Witte 1997). In addition to information about **x** and **y** (as defined above), one also utilizes information about similarities among the components of β. For example, one might assume that associations corresponding to markers that are located near one another on a particular chromosome might be similar. Conditional on this additional information, one may fit a second-stage generalized model for the expectation of β: $f_2(\beta \mid \mathbf{Z}) = \delta + \mathbf{Z}\pi$. According to this model, f_2 is a strictly increasing link function, and **Z** is a second-stage design matrix expressing the similarities among the β. Hierarchical (i.e., posterior) estimates are obtained by combining results from the different level models (Witte 1997).

5.5.6 Interactions

GWAS present an opportunity to go beyond single-locus analyses and into the realm of *gene-gene interactions* throughout the genome. Given the number of SNPs generally evaluated in GWAS, it would prove intractable to evaluate all pairwise combinations. Instead, one can reduce the set of SNPs to further investigate via one of several methods (McAllister et al. 2017). The first is to select an arbitrary significance threshold for the set of single-locus analyses. One can then evaluate all pairwise interactions between SNPs falling below the threshold, or between such SNPs and all other SNPs. Implementing this method, however, will preclude the discovery of combinations of markers that affect a significant change in disease risk even when the individual markers' marginal effects are statistically undetectable. An alternative approach is to restrict the analysis of interactions to SNPs with an established biological function or within a particular protein family. A general comment regarding all analyses of interaction is that the scale (additive or multiplicative) on which they are evaluated will impact the results.

The evaluation of gene-gene interactions is not limited to model-based methods. Multifactor dimensionality reduction (MDR) was developed to reduce the dimensionality of multilocus data so as to improve the ability to detect gene-gene interactions. MDR pools genotypes into high-risk and low-risk groups, thereby reducing data to a single dimension. The method is nonparametric and model-free—one need not make hypotheses regarding the values of any parameters or assume any particular mode of inheritance (Motsinger and Ritchie 2006). The details of MDR analyses have been well described (Hahn et al. 2003; Ritchie et al. 2001, 2003).

The study of *gene-environment interactions* is another critical component of understanding the biological mechanisms of complex disease, heterogeneity across studies, and susceptible subpopulations (Dick et al. 2015). Until recently, gene-environment interaction studies have been largely carried out using candidate approaches. Such studies require the identification of genes with related biological functionality as well as knowledge of the mode of action through which environmental factors affect the genes of interest (Rava et al. 2013).

With the advent of high-throughput technology, investigators are exploring gene-environment interactions at the genome-wide level. They are also realizing some of the fundamental challenges of doing so. The genome-wide approach does not make use of prior knowledge of biological processes and pathways. In addition, the stringent significance threshold required due to the number of statistical tests may preclude the identification of significant interactions.

Analysis approaches can focus on environmental interactions with single genes, multiple genes, and/or biological pathways (Thomas 2010). Alternatively, one can utilize available biomarker data that may reflect intermediate phenotypes to establish informative priors in a hierarchical model framework (Li et al. 2012). Regardless, it bears recognition that statistical interaction does not conclusively indicate causality. Variants showing interaction may not be causal, and interactions may be significant for reasons other than true association (as is true for any other type of association). Still, the identification of gene-environment interactions with respect to disease risk is of fundamental public health relevance.

5.5.7 Incorporating Covariates

5.5.7.1 Population Stratification

Population-based association studies are susceptible to a form of confounding known as *population stratification*. It occurs whenever the gene of interest shows pronounced variation in allele frequency across ancestral subgroups of the population, and these subgroups differ in their baseline risk of disease. In extreme scenarios, population stratification can result from *cryptic relatedness*, wherein some individuals in an ostensibly unrelated population are actually related.

In a sample with population stratification in any form, SNPs with large allele frequency differences across groups are likely to be associated with the trait under study. The first step toward dealing with the bias is to ensure that cases and controls are well-matched in the study design phase. One can then evaluate the extent of residual population stratification via *Q*–*Q* plots and their associated inflation factor, lambda (λ). The latter is defined as the ratio of the median of the observed test statistics relative to the expected median and reflects the excess false-positive rate. When the value of lambda is inflated, one can adjust the test statistics by dividing them by lambda, thereby reducing them, and then recalculating the associated *P* values (Devlin and Roeder 1999).

In recent years, the more common approach to the management of population stratification has been the measurement of the ancestry of each sample in the dataset using PC methods (Price et al. 2006; Falush et al. 2003). These methods cluster individuals together based on their ancestral populations, often by comparing them with an external reference population such as the 1000 Genomes Project or TOPMed imputation reference panel. With the results, one can then exclude samples that are extremely different from the main clusters of individuals and then include the top 10 or so PCs as covariates in association analyses. A criticism of PCs is that they are unable to differentiate between true signal due to polygenicity and confounding

due to population stratification. An alternative adjustment method can be used that requires LD score regression, which quantifies the association between LD and test statistics within a GWAS using a reference panel, in order to calculate a correction factor for genomic control in GWAS analysis (Bulik-Sullivan et al. 2015).

Methods that account for cryptic relatedness specifically tend to be more complex and are largely beyond the scope of this chapter. Software packages such as KING (Manichaikul et al. 2010) can be implemented to identify closely related individuals, who can subsequently be removed from the analytical population. There also exist approaches to deal with more distant relatedness as part of data analysis (Price et al. 2010).

5.5.7.2 Addressing Other Confounding

Genetic association studies are distinct from traditional epidemiological studies in that behavioral and environmental factors are unlikely to confound the associations of interest. Despite this, to draw clinically relevant conclusions, it is critical to properly account for patient-level covariates that may confound the associations under investigation. For example, associations may be confounded by sex whenever allele frequencies differ between the sexes (i.e., for sex-linked traits) (Clayton 2009). Other associations may be confounded by age whenever tag SNPs are in LD with both longevity SNPs and causal SNPs for a phenotype that most commonly occurs in late-life. If one does not adjust for the necessary covariates, one might find spurious associations due to sampling artifacts or bias in the study design. It is important to note that covariate adjustment may reduce statistical power because it requires additional degrees of freedom.

5.5.7.3 Improving Precision

Covariates may also be included in tests of association in an attempt to improve the precision of estimates. If a behavioral or environmental factor is associated with a quantitative phenotype under study independently of the genes of interest, then its inclusion is often beneficial. The covariate can explain some of the variability in the outcome, thereby reducing noise and increasing power (Mefford and Witte 2012). For binary traits, the story is more complicated. Inclusion of a covariate associated only with the outcome may actually reduce power for case-control association studies (Pirinen et al. 2012). There do exist methods, however, that leverage information about covariates to increase power in association studies of binary traits (Zaitlen et al. 2012).

5.5.8 Multiple Testing

Genetic association studies generally test hundreds of thousands of associations and may also examine multiple phenotypes and/or the results from various genetic models and covariate adjustments. The enormous number of resulting hypothesis tests must be adjusted for multiple comparisons, lest a large number of false-positive associations be detected. One approach to management of this issue of *multiple*

testing is to limit the number of association tests executed. For example, in a study of multiple SNPs and a single phenotype, one might consider running only a single test per SNP. If the number of SNPs is sufficiently large, however, one will have to further adjust for the number of tests performed. Doing so is helpful to address issues of multiple comparisons but it is also important to note that the significance cutoffs described below are somewhat arbitrary and do not reflect the potential clinical or biological importance of an association (Witte et al. 1996).

5.5.8.1 Bonferroni and Number of Effective Independent Tests

The most straightforward and the most stringent means by which to correct for multiple testing is the *Bonferroni adjustment*. It adjusts the conventional type I error rate of 0.05 to $0.05/k$, where k is the total number of tests performed. The adjustment assumes that the hypothesis tests are independent, thereby making this approach quite conservative in the context of correlated tests.

To make the strategy more applicable to the scenario of SNPs in LD, one can estimate the effective number of independent SNPs in lieu of the total number of SNPs in the Bonferroni adjustment (Nyholt 2004). Because the number of effective independent SNPs will always be less than or equal to the total number of genotyped SNPs, this approach is less conservative than the standard Bonferroni correction. For GWAS, the generally accepted alpha-level for statistical significance based on the effective number of genome-wide tests is 5×10^{-8}. This concept of *genome-wide significance* should be used only when hypotheses are tested on the genomic scale. It is not appropriate for candidate gene studies or replication studies, for which the number of effective independent tests is substantially lower.

5.5.8.2 Permutation Testing

Permutation testing is a more computationally intensive method with which to adjust for multiple testing. It is less conservative than the Bonferroni adjustment because it incorporates the correlation between genotypes and/or phenotypes. It does so by randomly shuffling phenotypes in the dataset, effectively removing any association between phenotypes and genotypes, while maintaining the correlation among genotypes resulting from LD within an individual, and then testing for association again. Random reassignment of the data and association testing is repeated some prespecified number of times (generally in the thousands), and all of the test statistics for the associations of interest are computed for each permuted dataset. A permuted P value can then be obtained by comparing the original test statistic to the distribution of test statistics from the permuted datasets. Several statistical packages implement permutation testing, though the most commonly used is PLINK (Purcell et al. 2007).

5.5.8.3 False Discovery Rate

Another method to account for multiple testing is the *false discovery rate (*Benjamini and Hochberg 1995*;* Brzyski et al. 2017*)*. Rather than control the family-wise error rate as does the Bonferroni adjustment, the false discovery rate controls the expected proportion of false discoveries among significant results.

Under a null hypothesis of no true associations in a GWAS dataset, *P* values would conform to a uniform distribution between zero and one. The false discovery rate essentially corrects for the expected number of false discoveries under this null distribution. While it typically allows researchers to reject more null hypotheses than would a Bonferroni adjustment, it may still be overly rigorous in the context of GWAS or large-scale candidate gene association studies. In such scenarios, one may implement weighted or stratified false discovery rates to achieve greater power to detect true associations in subsets of SNPs with a higher proportion of true positives than in the full set of SNPs (Genovese et al. 2006; Greenwood et al. 2007; Roeder et al. 2006).

5.5.8.4 Bayesian Approach

Rather than correct for multiple comparisons via traditional frequentist methods, one might choose to implement a Bayesian approach to the false discovery rate. Under a Bayesian framework, the Bayes factor quantifies the strength of evidence for an association between a SNP and phenotype. It is weighed against the prior probability of an association to arrive at the posterior probability (Stephens and Balding 2009). The calculation does not reference the number of SNPs tested. While the expected number of false-positive associations will increase as more tests are performed, so too will the number of true positive associations under a reasonable set of assumptions. As such, the ratio of true to false positives will remain roughly constant. Several software packages (e.g., SNPTEST (Marchini et al. 2007) and BIMBAM (Servin and Stephens 2007)) accommodate genome-wide Bayesian analyses.

5.6 Concluding Remarks

Well-executed genetic association studies can contribute immensely to our understanding of the underpinnings of disease. For meaningful conclusions to be drawn, it is critical that they be designed with an appropriate population consisting of a sufficient number of subjects. This number will depend upon the design selected—candidate gene, GWAS, or otherwise—and should take into account the methodological nuances thereof. Accurate measurement of genetic information is also integral to the success of an association study, as are the quality control checks that validate it. Then, statistical analyses must consider the specific research question at hand, so as to make decisions that will best answer it.

There remain many gaps in our understanding and many association studies that have the potential to fill them in moving forward. Since the proposal of an exposome in 2005 (Wild 2005), investigators have striven to conceive of methods that incorporate all of the exposures that individuals experience in a lifetime into the study of their genetics. They have also been busy considering the question of pleiotropy so as to identify genes that affect multiple, sometimes seemingly unrelated, phenotypes. Many are developing methods that will better address rare variants and interactions. The collection of these efforts will further improve our

understanding of the disease process, risks, and response to therapy in this era of genomic discovery.

Acknowledgments This work was supported by National Institutes of Health grants R25CA112355, R01CA088164, and R01CA201358.

References

Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322:881–888

Aragaki CC, Greenland S, Probst-Hensch N, Haile RW (1997) Hierarchical modeling of gene-environment interactions: estimating NAT2 genotype-specific dietary effects on adenomatous polyps. Cancer Epidemiol Biomark Prev 6:307–314

Asimit J, Zeggini E (2010) Rare variant association analysis methods for complex traits. Annu Rev Genet 44:293–308

Asimit JL, Day-Williams AG, Morris AP, Zeggini E (2012) ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. Hum Hered 73:84–94

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007) GenABEL: an R library for genome-wide association analysis. Bioinformatics 23:1294–1296

Barbeira AN, Dickinson SP, Bonazzola R et al (2018) Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun 9:1825

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B 57:289–300

Bhattacharjee S, Rajaraman P, Jacobs KB et al (2012) A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. Am J Hum Genet 90:821–835

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331

Bowden J, Del Greco MF, Minelli C, Davey Smith G, Sheehan N, Thompson J (2017) A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. Stat Med 36:1783–1802

Brzyski D, Peterson CB, Sobczyk P, Candes EJ, Bogdan M, Sabatti C (2017) Controlling the rate of GWAS false discoveries. Genetics 205:61–75

Bulik-Sullivan B, Loh P-R, Finucane H et al (2015) LD score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 47:291–295

Buniello A, MacArthur JAL, Cerezo M et al (2019) The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 47:D1005–D1012

Burgess S, Butterworth A, Thompson SG (2013) Mendelian randomization analysis with multiple genetic variants using summarized data. Genet Epidemiol 37:658–665

Burton PR, Clayton DG, Cardon LR et al (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678

Cantor RM, Lange K, Sinsheimer JS (2010) Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Am J Hum Genet 86:6–22

Cardin NJ, Mefford JA, Witte JS (2012) Joint association testing of common and rare genetic variants using hierarchical modeling. Genet Epidemiol 36:642–651

Carlson CS, Matise TC, North KE et al (2013) Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. PLoS Biol 11:e1001661

Chanock SJ, Manolio T, Boehnke M et al (2007) Replicating genotype-phenotype associations. Nature 447:655–660

Chen GK, Witte JS (2007) Enriching the analysis of genomewide association studies with hierarchical modeling. Am J Hum Genet 81:397–404
Chen H, Lumley T, Brody J et al (2014) Sequence kernel association test for survival traits. Genet Epidemiol 38:191–197
Chen H, Huffman JE, Brody JA 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
Choi M, Scholl UI, Ji W et al (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci USA 106:19096–19101
Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT (2011) Basic statistical analysis in genetic case-control studies. Nat Protoc 6:121–133
Claussnitzer M, Cho JH, Collins R et al (2020) A brief history of human disease genetics. Nature 577:179–189
Clayton DG (2009) Sex chromosomes and genetic association studies. Genome Med 1:110
Cordell HJ, Clayton DG (2005) Genetic association studies. Lancet 366:1121–1131
Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229–232
Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004
Dick DM, Agrawal A, Keller MC et al (2015) Candidate gene-environment interaction research: reflections and recommendations. Perspect Psychol Sci 10:37–59
Dinu I, Potter JD, Mueller T et al (2009) Gene-set analysis and reduction. Brief Bioinform 10:24–34
Dutta D, Scott L, Boehnke M, Lee S (2019) Multi-SKAT: general framework to test for rare-variant association with multiple phenotypes. Genet Epidemiol 43:4–23
Elbers CC, van der Schouw YT, Wijmenga C, Onland-Moret NC (2009a) Comment on: Perry et al. (2009) interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach. Diabetes 58:1463–1467. e9; author reply e10
Elbers CC, van Eijk KR, Franke L et al (2009b) Using genome-wide pathway analysis to unravel the etiology of complex diseases. Genet Epidemiol 33:419–431
Evangelou E, Ioannidis JPA (2013) Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet 14:379–389
Evangelou M, Dudbridge F, Wernisch L (2014) Two novel pathway analysis methods based on a hierarchical model. Bioinformatics 30:690–697
Evangelou E, Warren HR, Mosen-Ansorena D et al (2018) Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. Nat Genet 50:1412–1425
Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587
Fehringer G, Kraft P, Pharoah PD et al (2016) Cross-cancer genome-wide analysis of lung, ovary, breast, prostate, and colorectal cancer reveals novel pleiotropic associations. Cancer Res 76:5103–5114
Frazer KA, Ballinger DG, Cox DR et al (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861
Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10:241–251
Gabriel SB, Schaffner SF, Nguyen H et al (2002) The structure of haplotype blocks in the human genome. Science 296:2225–2229
Gamazon ER, Wheeler HE, Shah KP et al (2015) A gene-based association method for mapping traits using reference transcriptome data. Nat Genet 47:1091–1098
Genovese CR, Roeder K, Wasserman L (2006) False discovery control with p-value weighting. Biometrika 93:509–524
Gilbert W, Maxam A (1973) The nucleotide sequence of the lac operator. Proc Natl Acad Sci USA 70:3581–3584
Gnirke A, Melnikov A, Maguire J et al (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 27:182–189

Goring HH, Terwilliger JD, Blangero J (2001) Large upward bias in estimation of locus-specific effects from genomewide scans. Am J Hum Genet 69:1357–1369
Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI (2008) Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet 82:100–112
Gray R, Wheatley K (1991) How to avoid bias when comparing bone marrow transplantation with chemotherapy. Bone Marrow Transplant 7(Suppl 3):9–12
Greenwood CM, Rangrej J, Sun L (2007) Optimal selection of markers for validation or replication from genome-wide association studies. Genet Epidemiol 31:396–407
Guey LT, Kravic J, Melander O et al (2011) Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. Genet Epidemiol 35:236–246
Gusev A, Ko A, Shi H et al (2016) Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet 48:245–252
Hahn LW, Ritchie MD, Moore JH (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19:376–382
Han B, Eskin E (2011) Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. Am J Hum Genet 88:586–598
Haycock PC, Burgess S, Wade KH, Bowden J, Relton C, Davey SG (2016) Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. Am J Clin Nutr 103:965–978
Hindorff LA, Sethupathy P, Junkins HA et al (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362–9367
Hirschhorn JN, Altshuler D (2002) Once and again-issues surrounding replication in genetic association studies. J Clin Endocrinol Metab 87:4438–4441
Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genet Med 4:45–61
Ho LA, Lange EM (2010) Using public control genotype data to increase power and decrease cost of case–control genetic association studies. Hum Genet 128:597–608
Hodges E, Xuan Z, Balija V et al (2007) Genome-wide in situ exon capture for selective resequencing. Nat Genet 39:1522–1527
Hoffmann TJ, Kvale MN, Hesselson SE et al (2011a) Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. Genomics 98:79–89
Hoffmann TJ, Zhan Y, Kvale MN et al (2011b) 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
Hoffmann TJ, Van Den Eeden SK, Sakoda LC et al (2015) A large multiethnic genome-wide association study of prostate cancer identifies novel risk variants and substantial ethnic differences. Cancer Discov 5:878–891
Hoffmann TJ, Passarelli MN, Graff RE et al (2017) Genome-wide association study of prostate-specific antigen levels identifies novel loci independent of prostate cancer. Nat Commun 8:14248
Hong MG, Pawitan Y, Magnusson PK, Prince JA (2009) Strategies and issues in the detection of pathway enrichment in genome-wide association studies. Hum Genet 126:289–301
Hong J, Lunetta KL, Cupples LA, Dupuis J, Liu CT (2016) Evaluation of a two-stage approach in trans-ethnic meta-analysis in genome-wide association studies. Genet Epidemiol 40:284–292
Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 44:955–959
Huang BE, Lin DY (2007) Efficient association mapping of quantitative trait loci with selective genotyping. Am J Hum Genet 80:567–576
Huang L, Li Y, Singleton AB et al (2009) Genotype-imputation accuracy across worldwide human populations. Am J Hum Genet 84:235–250
Huang QQ, Ritchie SC, Brozynska M, Inouye M (2018) Power, false discovery rate and Winner's curse in eQTL studies. Nucleic Acids Res 46:e133

Huedo-Medina TB, Sanchez-Meca J, Marin-Martinez F, Botella J (2006) Assessing heterogeneity in meta-analysis: Q statistic or I2 index? Psychol Methods 11:193–206
International HapMap Consortium (2003) The International HapMap Project. Nature 426:789–796
International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320
Ioannidis JP (2005) Why most published research findings are false. PLoS Med 2:e124
Ioannidis JP (2006) Common genetic variants for breast cancer: 32 largely refuted candidates and larger prospects. J Natl Cancer Inst 98:1350–1353
Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. Nat Genet 29:306–309
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 92:841–853
Jorgenson E, Witte JS (2006) Coverage and power in genomewide association studies. Am J Hum Genet 78:884–888
Karlsson Linner R, Biroli P, Kong E et al (2019) Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. Nat Genet 51:245–257
Katan MB (1986) Apolipoprotein E isoforms, serum cholesterol, and cancer. Lancet 1:507–508
Kraft P, Wacholder S, Cornelis MC et al (2009) Beyond odds ratios—communicating disease risk based on genetic profiles. Nat Rev Genet 10:264–269
Lander ES, Linton LM, Birren B et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921
Larson NB, McDonnell S, Cannon Albright L et al (2017) gsSKAT: rapid gene set analysis and multiple testing correction for rare-variant association studies using weighted linear kernels. Genet Epidemiol 41:297–308
Lee S, Emond MJ, Bamshad MJ et al (2012) Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet 91:224–237
Lettre G, Lange C, Hirschhorn JN (2007) Genetic model testing and statistical power in population-based association studies of quantitative traits. Genet Epidemiol 31:358–362
Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83:311–321
Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype imputation. Annu Rev Genomics Hum Genet 10:387–406
Li R, Conti DV, Diaz-Sanchez D, Gilliland F, Thomas DC (2012) Joint analysis for integrating two related studies of different data types and different study designs using hierarchical modeling approaches. Hum Hered 74:83–96
Li Z, Li X, Liu Y et al (2019) Dynamic scan procedure for detecting rare-variant association regions in whole-genome sequencing studies. Am J Hum Genet 104:802–814
Lindquist KJ, Jorgenson E, Hoffmann TJ, Witte JS (2013) The impact of improved microarray coverage and larger sample sizes on future genome-wide association studies. Genet Epidemiol 37:383–392
Liu M, Jiang Y, Wedow R et al (2019) Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat Genet 51:237–244
Lloyd-Jones LR, Robinson MR, Yang J, Visscher PM (2018) Transformation of summary statistics from linear mixed model association on all-or-none traits to odds ratio. Genetics 208:1397–1408
Loh PR, Tucker G, Bulik-Sullivan BK et al (2015) Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet 47:284–290
Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33:177–182
Luca D, Ringquist S, Klei L et al (2008) On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. Am J Hum Genet 82:453–463

Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5:e1000384
Magi R, Morris AP (2010) GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics 11:288
Maher B (2008) Personal genomes: the case of the missing heritability. Nature 456:18–21
Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM (2010) Robust relationship inference in genome-wide association studies. Bioinformatics 26:2867–2873
Manolio TA, Brooks LD, Collins FS (2008) A HapMap harvest of insights into the genetics of common disease. J Clin Invest 118:1590–1605
Marchini J, Howie B (2008) Comparing algorithms for genotype imputation. Am J Hum Genet 83:535–539. author reply 539-540
Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906–913
Marquez A, Kerick M, Zhernakova A et al (2018) Meta-analysis of immunochip data of four autoimmune diseases reveals novel single-disease and cross-phenotype associations. Genome Med 10:97
Mavaddat N, Michailidou K, Dennis J et al (2019) Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. Am J Hum Genet 104:21–34
McAllister K, Mechanic LE, Amos C et al (2017) Current challenges and new opportunities for gene-environment interaction studies of complex diseases. Am J Epidemiol 186:753–761
Mefford J, Witte JS (2012) The covariate's dilemma. PLoS Genet 8:e1003096
Mitchell BD, Fornage M, McArdle PF et al (2014) Using previously genotyped controls in genome-wide association studies (GWAS): application to the stroke genetics network (SiGN). Front Genet 5
Morgenthaler S, Thilly WG (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res 615:28–56
Morris AP (2011) Transethnic meta-analysis of genomewide association studies. Genet Epidemiol 35:809–822
Morris AP, Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34:188–193
Motsinger AA, Ritchie MD (2006) Multifactor dimensionality reduction: an analysis strategy for modelling and detecting gene-gene interactions in human genetics and pharmacogenomics studies. Hum Genomics 2:318–328
Mutsuddi M, Morris DW, Waggoner SG, Daly MJ, Scolnick EM, Sklar P (2006) Analysis of high-resolution HapMap of DTNBP1 (Dysbindin) suggests no consistency between reported common variant associations and schizophrenia. Am J Hum Genet 79:903–909
Nam D, Kim SY (2008) Gene-set approach for expression pattern analysis. Brief Bioinform 9:189–197
Neale BM, Rivas MA, Voight BF et al (2011) Testing for an unusual distribution of rare variants. PLoS Genet 7:e1001322
Nelson SC, Doheny KF, Pugh EW et al (2013) Imputation-based genomic coverage assessments of current human genotyping arrays. G3 (Bethesda) 3:1795–1807
Ng SB, Turner EH, Robertson PD et al (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461:272–276
Nolte IM, van der Most PJ, Alizadeh BZ et al (2017) Missing heritability: is the gap closing? An analysis of 32 complex traits in the lifelines cohort study. Eur J Hum Genet 25:877–885
Nyholt DR (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet 74:765–769
Paltoo DN, Rodriguez LL, Feolo M et al (2014) Data use under the NIH GWAS data sharing policy and future directions. Nat Genet 46:934–938
Pan W (2009) Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet Epidemiol 33:497–507
Panagiotou OA, Willer CJ, Hirschhorn JN, Ioannidis JPA (2013) The power of meta-analysis in genome-wide association studies. Annu Rev Genomics Hum Genet 14:441–465

Pierce BL, Burgess S (2013) Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. Am J Epidemiol 178:1177–1184
Pirinen M, Donnelly P, Spencer CC (2012) Including known covariates can reduce power to detect genetic effects in case-control studies. Nat Genet 44:848–851
Piro RM, Di Cunto F (2012) Computational approaches to disease-gene prediction: rationale, classification and successes: computational disease-gene prediction. FEBS J 279:678–696
Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909
Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11:459–463
Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–228
Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575
Rava M, Ahmed I, Demenais F, Sanchez M, Tubert-Bitter P, Nadif R (2013) Selection of genes for gene-environment interaction studies: a candidate pathway-based strategy using asthma as an example. Environ Health 12:56
Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517
Ritchie MD, Hahn LW, Roodi N et al (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69:138–147
Ritchie MD, Hahn LW, Moore JH (2003) Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol 24:150–157
Roeder K, Wasserman L (2009) Genome-wide significance levels and weighted hypothesis testing. Stat Sci 24:398–413
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 78:243–252
Roshan U, Chikkagoudar S, Wei Z, Wang K, Hakonarson H (2011) Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest. Nucleic Acids Res 39:e62–e62
Santorico SA, Hendricks AE (2016) Progress in methods for rare variant association. BMC Genet 17(Suppl 2):6
Servin B, Stephens M (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 3:e114
Sinnott JA, Kraft P (2012) Artifact due to differential error when cases and controls are imputed from different platforms. Hum Genet 131:111–119
Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38:209–213
Skol AD, Scott LJ, Abecasis GR, Boehnke M (2007) Optimal designs for two-stage genome-wide association studies. Genet Epidemiol 31:776–788
Smith GD, Ebrahim S (2003) 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol 32:1–22
Stephens M, Balding DJ (2009) Bayesian statistical methods for genetic association studies. Nat Rev Genet 10:681–690
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
Teer JK, Mullikin JC (2010) Exome sequencing: the sweet spot before whole genomes. Hum Mol Genet 19:R145–R151
Thomas D (2010) Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. Annu Rev Public Health 31:21–36

Thomas DC, Haile RW, Duggan D (2005) Recent developments in genomewide association scans: a workshop summary and review. Am J Hum Genet 77:337–345
Thomas DC, Casey G, Conti DV, Haile RW, Lewinger JP, Stram DO (2009) Methodological issues in multistage genome-wide association studies. Stat Sci 24:414
Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ (2005) Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci USA 102:13544–13549
Toland AE (2019) Polygenic risk scores for prostate cancer: testing considerations. Can J Urol 26:17–18
Torkamani A, Wineinger NE, Topol EJ (2018) The personal and clinical utility of polygenic risk scores. Nat Rev Genet 19:581–590
Venter JC, Adams MD, Myers EW et al (2001) The sequence of the human genome. Science 291:1304–1351
Wacholder S, McLaughlin JK, Silverman DT, Mandel JS (1992) Selection of controls in case-control studies. I Principles. Am J Epidemiol 135:1019–1028
Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N (2004) Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Natl Cancer Inst 96:434–442
Wang DG, Fan JB, Siao CJ et al (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077–1082
Wang K, Li M, Hakonarson H (2010) Analysing biological pathways in genome-wide association studies. Nat Rev Genet 11:843–854
Wang L, Jia P, Wolfinger RD, Chen X, Zhao Z (2011) Gene set analysis of genome-wide association studies: methodological issues and perspectives. Genomics 98:1–8
Ware JH (2006) The limitations of risk factors as prognostic tools. N Engl J Med 355:2615–2617
Wild CP (2005) Complementing the genome with an “exposome”: the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomark Prev 14:1847–1850
Willer CJ, Li Y, Abecasis GR (2010) METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26:2190–2191
Witte JS (1997) Genetic analysis with hierarchical models. Genet Epidemiol 14:1137–1142
Witte JS, Greenland S (1996) Simulation study of hierarchical regression. Stat Med 15:1161–1170
Witte JS, Elston RC, Schork NJ (1996) Genetic dissection of complex traits. Nat Genet 12:355–356. author reply 357–358
Witte JS, Elston RC, Cardon LR (2000) On the relative sample size required for multiple comparisons. Stat Med 19:369–372
Wojcik GL, Fuchsberger C, Taliun D et al (2018) Imputation-aware tag SNP selection to improve power for large-scale, multi-ethnic association studies. G3 (Bethesda) 8:3255–3267
Wu R, Kaiser AD (1968) Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. J Mol Biol 35:523–537
Wu Michael C, 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
Wu R, Taylor E (1971) Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage lambda DNA. J Mol Biol 57:491–511
Wu MC, Kraft P, Epstein MP et al (2010) Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet 86:929–942
Wu MC, Maity A, Lee S et al (2013) Kernel machine SNP-set testing under multiple candidate kernels. Genet Epidemiol 37:267–275
Xing C, Huang J, Hsu YH et al (2016) Evaluation of power of the Illumina HumanOmni5M-4v1 BeadChip to detect risk variants for human complex diseases. Eur J Hum Genet 24:1029–1034
Xiong M, Guo SW (1997) Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. Am J Hum Genet 60:1513–1531
Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL (2014) Advantages and pitfalls in the application of mixed-model association methods. Nat Genet 46:100–106

Zaitlen N, Lindstrom S, Pasaniuc B et al (2012) Informed conditioning on clinical covariates increases power in case-control association studies. PLoS Genet 8:e1003032

Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zollner S (2010) Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. Am J Hum Genet 87:604–617

Zeggini E, Ioannidis JP (2009) Meta-analysis in genome-wide association studies. Pharmacogenomics 10:191–201

Zhang Z, Ersoz E, Lai CQ et al (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42:355–360

Zhou X, Stephens M (2012) Genome-wide efficient mixed-model analysis for association studies. Nat Genet 44:821–824

Zhu Z, Lee PH, Chaffin MD et al (2018) A genome-wide cross-trait analysis from UK biobank highlights the shared genetic architecture of asthma and allergic diseases. Nat Genet 50:857–864