20

Association Mapping

Whereas the association signals detected can help to define regions of interest, they cannot provide unambiguous identification of the causal genes. WTCCC (2007)

The initial focus of a GWAS should be on the study design, determining the specific trait for investigation, the ascertainment of samples, and the choice of SNPs to be interrogated for association. Without thoughtful consideration of these design issues ... the results of any downstream association will be meaningless. Morris and Cardon (2019)

Version 12 Dec. 2022

The original idea behind an association analysis (Chapter 17) was to test a population sample for correlations between different genotypes at a candidate gene and trait values (or disease risk). While the scored genotypes in early studies were often presumed to be causal sites (such as SNPs coding for different amino acids), this was not always the case. More generally, the scored genotypes were simply markers in, or near, a candidate gene, assumed to in high linkage disequilibrium (LD; Chapter 5) with causal sites (QTLs). With the rise of genomic technologies that score hundreds of thousands of markers, this notion of a hypothesis-driven test of the impact of a *specific locus* expanded to an exploratory *scan of the genome* (with no specific candidates in mind), searching for associations over a dense set of markers covering the genome. This strategy is commonly referred to as a **genome wide association study**, or **GWAS**. In the era of electronic health records, human geneticists often perform a series of GWAS over thousands of recorded traits, which Zhou et al. (2018) referred to as a **phenome-wide association study** (**PheWAS**).

GWAS is now the dominant means of searching for QTLs. While the focus of GWAS studies is typically on humans, equally impressive work has been done in other species, such as with maize using the NAM lines (Chapter 18). As detailed below, a limiting factor to performing a GWAS is the ability to accurately, and cheaply, score vast numbers of markers. Genomics has made this economically feasible in humans, model organisms (e.g., mice, rats, *Drosophila*, *Arabidopsis*), high-value agricultural species (e.g., maize, rice, dairy cattle, chickens), and is becoming increasingly feasible for an ever-growing list of species. While GWAS usually involves either a random sample of individuals from a defined population or a specialized sample of cases with a matching set of controls, plant breeders and investigators using model organisms often use a collection of inbred lines as their GWAS population. Such a mapping population is referred to as an **association panel**.

Much of the basics of a GWAS follow from the marker-trait approach used in linkage-based mapping (Chapters 18 and 19). This chapter largely focuses on the design and analysis aspects of a GWAS, while Chapter 21 examines the data generated by GWAS (as well as by linkage-based studies) to provide an overview of our current picture of the genetic architecture of traits and diseases.

MOTIVATION AND HISTORY

LD Assocation Between a Marker and a QTL

The marker signal from a nearby QTL in both linkage-based (Chapters 18 and 19) and association mapping (Chapter 17) is a *marker-trait association*: the different genotypes at the marker have different mean trait values (or differental risks for binary traits). Such

signals arise when the structure of the mapping population generates correlations between markers and nearby QTLs. This naturally occurs in families (due to the excess of parental over recombinant gametes for linked loci; Example 5.6). Such associations also occur in a population due to linkage disequilibrium (**LD**; Chapters 5 and 17), although spanning a much smaller genomic scale than that generated by linkage following a cross.

To see how naturally occurring LD arises, suppose a QTL allele, q, undergoes a mutation to an new functional allele, Q. If this happens close to a SNP, there is an initial association with the SNP allele on the chromosome (or **background haplotype**) on which the mutation arose. Suppose that an Mq chromosome mutates to MQ. Initially Q is *only* found with marker allele M. At the time of the mutation, $|D'| = 1$ (Example 5.5), namely **complete disequilibrium** (no recombination between the alleles is seen in the sample), as Q-bearing chromosomes *always* contain M. However, it is r^2, the correlation between alleles, that measures how much of the signal at a causative site is imparted onto a nearby marker, with $r^2 = 1$ called **perfect disequilibrium**. As Example 5.5 showed, here $r^2 < 1$, as M-bearing chromosomes can be found *without* Q, so that the $M - Q$ correspondence is not perfect.

Equation 5.13d showed that the intial value of r^2 decays at a rate of $(1-c)^2$ per generation. Thus, if c is very small, the bulk of this initial associations can persist for thousands of generations. Assuming a one percent recombination rate per megabase, then for two loci separated by 10 kilobases, $c = 0.01 \cdot (10/1000) = 0.0001$. The expected time ($t_{1/2}$) for 50% of the initial r^2 value to decay in this setting satisfies $(1 - 0.0001)^{2t_{1/2}} = 0.5$. Solving gives $t_{1/2} = \ln(0.5)/[2\ln(0.999)] = 346$ generations. Assuming a 25 year human generation span, this is roughly 8,600 years. After 1000 generations, the r^2 value is still 14% of its intial value. If the markers are one kilobase apart, then 98% of the intial r^2 value is present after 1000 generations (again assuming a c of 0.01 per megabase).

The extend of linkage disequilibrium within a population is strongly influenced by its past evolutionary history. The time to the most recent common ancestor (TMRCA) largely sets the size of the **LD block**, as the longer the TMRCA, the more generations a region has to experience recombination. Because TMRCA scales with effective population size N_e (WL Chapters 2, 3, and 8), populations with historically large effective population sizes should show smaller average LD-block sizes than historically smaller populations, or than populations that passed thorough a bottleneck (as is often the case with domesticated lines).

This point is well made by the data on crop LD reviewed by Buckler and Gore (2007). In maize, LD spans less than a kilobase (kb) for wild populations and landraces, yet is greater than 100 kb for elite inbreds. In barley, LD spans less than a kb in wild populations, between 80 and 100 kb in landraces, and greater than 200 kb in elite lines, while soybeans have LD of around 40-80 kb in wild populations, but greater than 300 kb in elite lines. Similarly, in humans, African populations show much shorter blocks of LD than do European or Asian populations. Africa is the ancestral (and, hence, older) population, while the latter two populations are the result of migrations from this base population that passed through bottlenecks.

Because of population-specific variation in LD block size (which also varies widely *within* the genome of a given population), before performing a GWAS, one should employ a large set of randomly spaced SNPs to examine the extend of LD in the target population or association panel. Such a pilot study informs the investigator of the average marker density needed, and hence the limit of mapping resolution set by the sample LD.

Example 20.1 A common misconception in GWAS studies is the assumption that *markers closer to a causal site will have larger LD values*, and hence larger marker effects (via larger r^2 values). ***Such need not be the case***. Consider a new causal allele, *Q*, that arose on an *NM* marker background (haplotype), where the QTL locus is much closer to *M* than *N*. Suppose that recombination is sufficiently rare such that no *Q* alleles are found on any other haplotypes in the GWAS sample. $D'_{MQ} = D'_{NQ}$ = 1 in the case (as *Q* is *only* found on an *N* or an *M* background), but their r^2 values (which determine how much of the actual variance is

accounted for by the marker variance) are a function of the marker allele frequencies.

Equation 5.15c gives the correlation between *M* and *Q* as

$$r^2_{MQ} = \alpha_M(1 - p_M)/(1 - p_Q) \leq \alpha_M$$

where p_M is the frequency of the marker allele (*M*) and α_M the frequency of *M* alleles associated with *Q* (so that $\alpha_M p_M = p_Q$ is the frequency of *Q*). Suppose that the frequency of *Q* is 3% (making it a "common allele"), while the frequencies of *M* and *N* are, respectively, 60% and 5%. Here p_M = 0.6 and α_M = 3/60 = 0.05, giving r^2_{MQ} = 0.05·0.4/(1 − 0.03) = 0.0206, so that the marker variance accounts for only 2.06% of the causal (actual) variance. Conversely, for the more distant marker, *N*, p_N = 0.05 and α_N = 3/5 = 0.6, giving r^2_{NQ} = 0.6·0.95/(1-0.03) = 0.588. Thus, the marker variance for the more distant site captures almost 60% of the actual (causal) variance, a thirty-fold increase over the marker variance for the closer site, *M*. This is an illustration of the concept from Chapter 5 that r^2 LD values are ***largest when the causal and marker allele frequencies are similar***, and fall off as their absolute frequency difference increases. As we will see, this impacts the power of a GWAS to detect common versus rare alleles.

Advantages of Association Over Linkage Mapping

The limitations of linkage-based QTL mapping (Chapters 18 and 19) revolve around their need to be *based entirely on sets of relatives* (pairs, trios, families, or more extended pedigrees). This restriction has two consequences. First, it severely constrains the *obtainable sample size* of an experiment, as gathering large sets of relatives is nontrivial. Smaller samples constrain the SNP effect sizes that can be detected, with sample sizes in the thousands needed to reliably detect QTLs of modest effect (less than five percent; Table 19.2). A related issue with linkage-based studies, especially for rare diseases, is that families often show **singletons**, with only one family member being affected, providing no linkage-based information. Second, using relatives restricts the average number of recombination events that separates two sampled individuals. Any two random gametes from relatives in a linkage-based design likely have experienced only a very small number of recombination events since their shared ancestor. As a result, LD spans a significant fraction of a chromosome (at least tens of megabases), restricting the level of resolution for fine mapping.

In contrast, it is straightforward (at least conceptually) to obtain a large or very large population sample. Human GWAS studies now routinely use tens-of-thousands of individuals, offering enormous power to *detect* small marker effects (differences in the mean values of genotypes at a given marker). Meta-analysis (Appendix 6) can be used to combine multiple GWAS, generating total sample sizes in the hundreds-of-thousands to millions. Further, the association signal is generated by very fine-scale linkage disequilibrium, often on the order of tens of kilobases (or less). Two individuals in a linkage study have experienced only a few rounds of meioses, while two unrelated individuals from an association sample may be separated by hundreds or thousands of meioses (Chapter 8). As a result of this disparity in recombination history, linkage generates very long-range associations, while LD associations decay very quickly and only persist over very small scales (Chapters 5, 8, and 17). Linkage studies, however, can have an advantage when causal alleles are very rare, as a population-level rare allele can be common within carefully selected pedigrees (such as those showing an unusual number of cases, early-onset cases, or extreme phenotypes).

The History of GWAS

The idea of exploiting population-level LD with nearby markers to fine map a gene beyond what is possible with linkage-based approaches was suggested by Bodmer (1986), and also by Lande and Thompson (1990), the latter with respect to finding molecular markers for the construction of marker-selection indices. Hästbacka et al. (1992) and Lehesjoki et al. (1993) noted that LD mapping (Chapter 17) allows for fine mapping (resolution of less than one cM; $c \leq 0.01$), even for rare diseases where there are simply not a sufficient

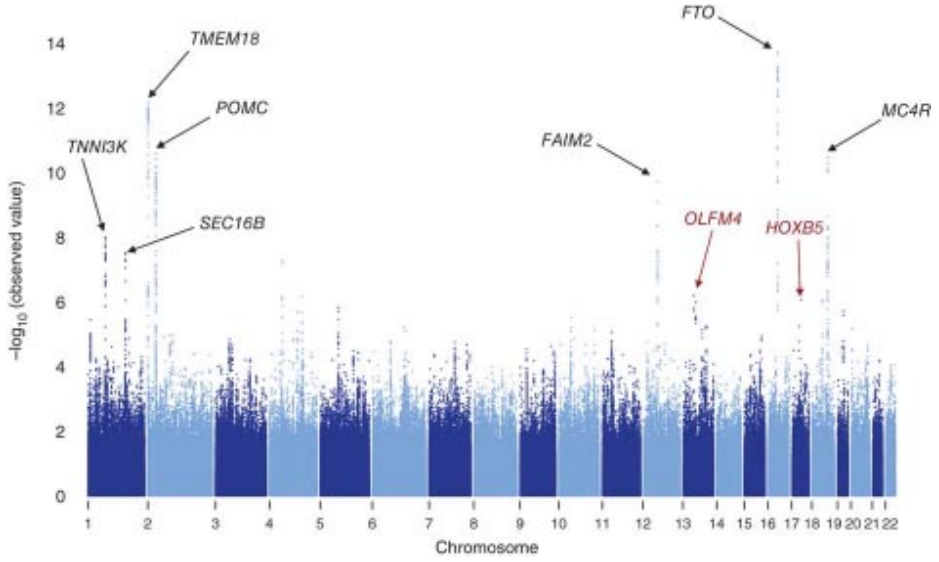

Figure 20.1 A **Manhattan plot**, where each pixel represents the position of a marker in the genome (horizontal axis, here arranged from chromosomes 1 to 22) and its significance, the value of $-\log_{10}(p)$, on the vertical axis. Sufficiently high peaks represent highly significant marker-trait associations, which are usually clustered, as nearby SNPs are typically in LD with each other. For most of the genome, the markers are so dense, and with only modest p values, that the pixels merge into a solid color. As one zooms in on a genomic location, these separate into individual pixels (Figure 20.2). In many plots, clear regions (a horizontal span with no vertical pixels) correspond to the centromeric regions of chromosomes, as these are often marker-poor (have a very low SNP density).

number of families to provide the number of meioses required to separate closely linked sites (Example 17.15). This approach starts with a known major gene that has been localized to a relatively small interval via linkage, which is then fine mapped using LD. Lyman et al. (1999) were among the first to exploit this idea in nonhuman populations, showing that certain *Drosophila* polymorphisms in the *scabrous* locus had an impact on bristle number. The notion that this same logic could apply in a genomic scan for QTLs was first emphasized in the very influential paper by Risch and Merikangas (1996), who pointed out that association studies have more power than sib-pair (allele-sharing; Chapter 19) approaches for detecting genes of modest to small effects. In contrast to the candidate-gene approach, this is a *hypothesis-free approach* (sites are simply chosen because they collectively span the genome).

Given the costs and logistical efforts involved, it is not surprising that the first large-scale GWAS were for human diseases. One of the first was that of Ozaki et al. (2002), who examined myocardial infraction risk in a Japanese population, using a sample of 1133 cases and 1006 controls. Of these, an initial sample of 94 cases and 658 controls were typed for 66,000 SNPs, searching for SNP alleles that were over-represented in the cases. One strong signal was a five SNP haplotype that spanned 50 kb, which included two SNPs in the lymphotoxon-α, LTA, gene. A follow-up analysis within this region with a set of 1133 cases and 1006 controls showed that individuals homozygous for the two LTA SNPs had a significant inflated risk (odds ratio of 1.78) of myocardial infraction.

This pioneering work was followed by two studies on the eye disease age-related macular degeneration (AMD) by Klein et al. (2005) and Dewan et al. (2006). Klein et al. examined 96 cases and 50 controls in Caucasians, scored for 166,000 SNPs. Individuals homozygous for a detected risk marker had a 7-fold increase in AMD. This SNP was in LD with an allele of complement factor H with a tyrosine to histidine change at residue 402. Dewan et al. examined 96 cases and 130 age-matched controls in a Chinese population, scored for 98,000 SNPs, finding a strong signal for AMD risk in the promoter region of $HTRA1$, a serine protease gene.

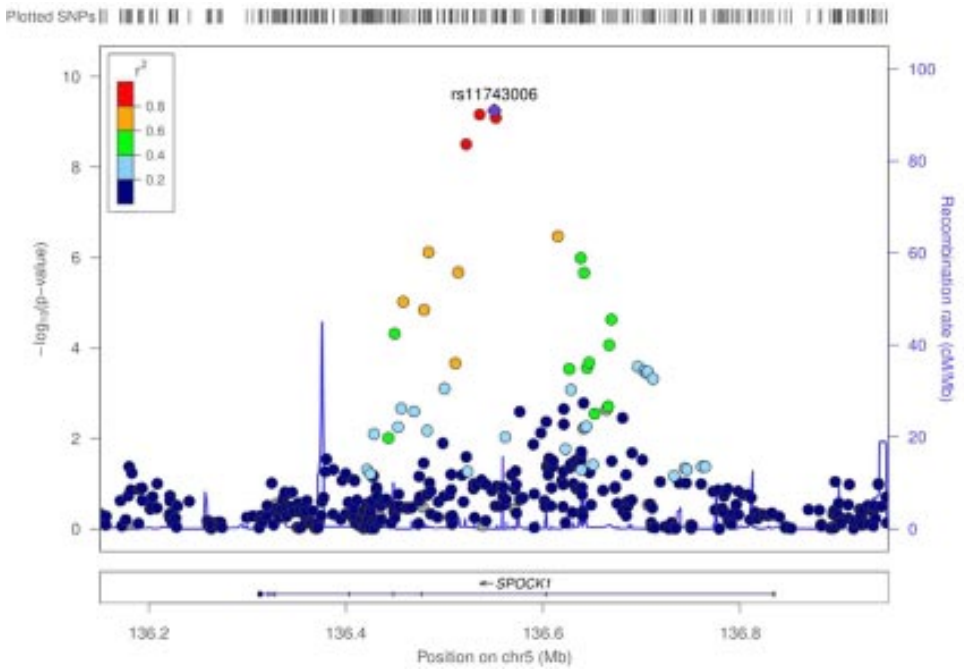

Figure 20.2 A **signal plot** for a cluster of significant SNPs, displaying the fine-scale view of a candidate region, often with functional genomic information. Here, the focus is the region around the *SPOCK1* gene on human chromosome 5, which Chen et al. (2017) found was associated with mathematical ability. The top of the plot shows the location of tested SNPs (tick marks), while the bottom of the plot shows the location on chromosome 5 (136.2 to 136.8 Mb), along with the *SPOCK1* gene and its direction of transcription. The left vertical axis is the $-\log_{10}(p)$ value for a SNP (displayed as a filled circle), while the right vertical axis is the value of localized recombination rate (displayed as a continuous curve; note the recombination hot spot just below 136.4 Mb). Finally, r^2 values for each SNP with respect to the lead SNP (rs11743006) are indicated by the shade of their filled circle. (After Chen et al. 2017.)

While technologically impressive for their time, all three of these studies were woefully underpowered, given the small number of genotyped cases. Hence, they could only detect major genes, and runned the risk of seriously overestimating SNP effects (Beavis effects; Chapter 18). The first truly modern GWAS was the **Wellcome Trust Case Control Consortium**, **WTCCC**, (2007), which scored 14,000 cases of seven common diseases (roughly 2000 cases per disease) with 3000 shared controls, using 500K SNPs genotyped using a Affymetric genechip array. The WTCCC study established many of the GWAS protocols currently in use today, e.g., stringent quality control (QC) on SNP calling and marker choice, the use of Manhattan and signal plots, adjusting for population structure, use of *q*-*q* plots to examine model fit, dealing with multiple comparisons, and stressing that more QTLs of smaller effect would likely be detected in a larger sample size.

Support for a QTL at a particular genomic location is usually displayed by a likelihood profile plot under linkage mapping (e.g., Figures 18.1, 19.3, and 19.5) because the effects of linkage typically span large regions of a chromosome. Conversely, under a GWAS, an LD block typically covers only a very tiny fraction of a chromosome. As a result, each tested marker is associated with a pixel that represents the p value for a test of marker-trait association, such as a Cochran-Armitage trend test (Example 18.5), a chi-square test, or a marker-dosage regression (Equation 18.4a). Note that these are the same single-marker association tests used with linkage studies (Chapters 18 and 19). The resulting plot has marker location on the horizontal axis and (usually) $-\log_{10}(p)$ on the vertical axis, so that higher values indicate greater significance. The result is a smear of pixels (usually so dense that they appear solid), out of which occasionally trickles a few pixels arising

above some threshold (Figure 20.1). Such displays are typically called **Manhattan plots** (or, more whimsically, **Dubai plots**), for the skyline of Manhattan with its sky scrapers. An unsuccessful GWAS often generates what some have called a **Manhattan, Kansas plot**, which, like the afore-mentioned town, is noted for its flatness and lack of any significant peaks. While a Manhattan plot gives a quick snapshot of the entire genome, **signal plots** (also called **regional association plots**) around significant peaks provide a more focused picture (WTCCC 2007). These depict the fine structure of the SNPs within a small region, and often incorporate additional genomic information, such as known open reading frames and localized recombination rates (e.g., Figure 20.2).

BASICS OF GWAS DESIGNS

The success, or failure, of a planned GWAS often hinges on the amount of care put into the initial study design. The enormous power of a modern GWAS (generated by their massive sample sizes) has a downside in that very small amounts of bias can propagate into false positives with very significant p values. Issues such as careful quality control (QC) when choosing SNPs, care in matching cases and controls, appropriate phenotyping, and population sampling strategies all need meticulous consideration. Even though it was the first truly large-scale GWAS, the WTCCC (2007) study remains the gold standard in dealing with these issues, and anyone contemplating a GWAS should careful review their paper before proceeding. Members of the WTCCC team have produced a series of protocols and short reviews on basic GWAS design (Zondervan and Cardon 2007), marker selection (Pettersson et al. 2009), quality control (Anderson et al. 2010), and basic analysis (Clarke et al. 2011; Morris and Cardon 2019; Wang et al. 2019; Uffelmann et al. 2021).

Gathering the Data

Three basic sampling designs are used in epidemiological studies. A **prospective cohort** study identifies a population (cohort) of individuals and follows then through some defined interval. When done well, these are non-trivial to perform in humans, as they often involve following individuals (and their descendants) over decades. The classic example is the **Framingham Heart study**, started in 1948 in Framingham, Massachusetts with 5200 adults, and is now on its third generations of participants. Another is the **1958 British birth cohort** (Power and Elliott 2006). A limiting factor with cohort designs for GWAS is they are often relatively small, so that the number of actual cases of a specific disease they contain may be modest, resulting in low power. A **cross-sectional** study gathers a single sample over some narrow time window. These are common in natural populations and agricultural settings. As with a cohort study, their limitation for disease mapping is that the resulting number of sampled cases may be small. In contrast to these two population sampling schemes, a **case-control** design specifically samples individuals with known cases and tries to match them with appropriate controls (e.g., WTCCC 2007). This is the typical design used for most diseases, and is an example of selective genotyping (Chapters 18 and 19), choosing individuals on the basis of their extreme phenotypes (here, disease presence). This design can be quite powerful, but it is also critically dependent on obtaining a control sample that is well-matched with the cases. For example, the control (ideally) should be roughly the same constitution as the case group in terms of (among other factors) sex, age, population of origin, and known exposures to environmental risk factors.

When the mapping population consists of an association panel of inbred lines, good experimental design can be used to improve the power of a GWAS (Myles et al. 2009). In humans, observations are made on single individuals, each (outside of monozygotic twins) representing a unique genotype, typically with their phenotype scored by a single observation. Conversely, when a panel of inbred lines is used, *replication* of each line can significantly reduce the residual variance of line values, significantly increasing power when the individual heritability (h^2) is low (Chapter 17). In these settings, even a modestly sized panel may have power in a GWAS. For example, using only 95 accessions of *Arabidopsis*

thaliana, Atwell et al. (2010) were able to locate a number of significant marker effects, in part because they could replicate each line, and thus measure each line value with very high precision and relatively little residual error. Replicative designs for estimating line means are examined in Chapters 24, 27, and Appendix 9. One issue with these designs is whether one should do a one-stage or a two-stage analysis. In a two-stage analysis, one first obtains the adjusted entry means (the estimates of the line means following correction for the design, e.g., block, plot, and environmental effects) and then uses these values for the GWAS. In a one-stage design (Cullis et al. 1998), these two operations are done simultaneously. Stich et al. (2008) showed that the two-stage design performed almost as well as the more exact (and much more computationally demanding) one-stage design for a GWAS with replicate lines from an association panel. Experimental design principles can also be used in a collection of outbred individuals. When phenotypes have high measurement error, the use of **repeated measures** (Chapter 31) can result in increased GWAS power (Kang et al. 2008).

Defining the Phenotype

As with linkage-based studies, poor phenotyping can doom a GWAS. At best, inconsistent phenotyping reduces power. This is especially a concern with somewhat ambiguously defined diseases, given that most human disease GWAS are now performed using a consortium of multiple institutions. Hence, very precise criteria, globally reproducible across clinical groups, are essential, such as the McDonald criteria, and its subsequent tweaks, for diagnosing multiple sclerosis (McDonald et al. 2001; Polman et al. 2005, 2011; Thompson et al. 2018). Similarly, careful definition and scoring of diseases in agricultural settings is critical, especially if the study involves a collaboration of multiple groups.

Likewise, for quantitative traits, careful, consistent, and accurate measures that can be applied across multiple investigative groups are essential. While there is considerable focus on quality control (QC) for *marker data*, a proper GWAS should have *at least* the same level of concern about the *quality of phenotyping*. An important class of quantitative traits in disease studies are **endophenotypes**, intermediate traits presumably upstream of a disease phenotype (Chapter 21). For example, transcript or protein levels at potential candidate genes, amounts of specific metabolites, or other biomarkers that may indicate disease risk or severity. Because of the impressive buffering of most biological systems, even large perturbations in these upstream actors can map into only small changes in the final phenotype. Hence, in some cases, a GWAS on well-chosen endophenotypes may be more powerful, and potentially more informative, than a GWAS directly on the disease itself, as these upstream elements may have larger (relative) effects. Again, the same concept holds in agricultural settings, such as measuring some quantitative feature of a disease (e.g., number of necrotic spots per leaf, parasite load) rather than simply scoring the disease as present or absence. Balancing the potential increased effect of an upstream endophenotypes is that such traits are often more challenging to measure than the target phenotype. This can result in a much smaller sample size that a GWAS based on the target phenotype, and hence lower power, even when an upstream effect size is larger.

SNP Selection and QC/QA

We assume here that a DNA chip, or related, technology is used to score a preselected set of markers, with whole-genome sequencing examined later in the chapter (and in Chapter 21 as well). When a chip (or some other mass-scoring of a preset collection of markers) is used, five issues can impact a GWAS: (i) coverage, (ii) quality of the marker data, (iii) systematic genotyping bias (such as batch effects), (iv) the appropriateness of the marker set for the population(s) of interest, and (v) the appropriateness of the marker set for the trait(s) of interest. We examine these issues in order.

The **coverage** of a chip (or a set of markers) is the fraction of sites not directly genotyped that are in high LD (and hence somewhat predictable) with genotyped markers. Coverage is usually not an issue with modern chips for humans and high-value agricultural species, but may be a concern in nonmodel species where resources have been more scarce. It is important

to stress that coverage is a *joint property of a chip and a target population*. Because the average length of an LD block changes with population history, a chip can have excellent coverage in one population but modest to poor coverage in another. A final issue with coverage is a point raised in the last two chapters: everything else being equal, power is usually increased more by *scoring more individuals, rather than more markers*. Spencer et al. (2009), focusing on human chips, reiterated this point, noting that lower coverage chips may result in more power for a GWAS if their use allows more individuals to be scored.

Quality control (**QC**), and the related issue of **quality assurance** (**QA**), involve, respectively, procedures to ensure quality as the genomic data is generated, and post-production review of the genotype quality (Laurie et al. 2010). We offer only a few brief comments here, as many of the QC/QA issues are platform specific. Standard measures of quality are the rate of missing calls (markers that fail to be scored by the genotyping process), departures from Hardy-Weinberg, and mismatches between the stated gender of a sample and the sex chromosome signal. Additional, and more sophisticated, QC metrics are reviewed by Anderson et al. (2010), Laurie et al. (2010), and Morris and Cardon (2019). A critical point, especially with case-control data, is randomizing individuals across genotyping batches. Systematic bias in a batch can impart false signals if the cases are run as one set of batches and controls as another. A similar issue can arise if all controls were genotyped at one center and all cases at another. Even when apparently extracted and scored under identicial conditions, Clayton et al. (2005) noted scoring biases in some markers between cases and controls, generating spurious associations.

The match of a set of markers to the study population is also critical. Historically, most of the initial human chips were very European-centric and hence could miss important polymorphisms in non-European populations. One might imagine that this issue is equally problematic in many agricultural settings when outlier populations are of interest (such exotic breeds or landraces). Likewise, as mentioned, a chip with good LD coverage in one population may have poor coverage in another.

The final issue with SNP choice is a bit more technical, but its roots follow from Example 20.1. If the allele frequencies of markers do not roughly match the allele frequencies of causal sites, the result is low r^2 values, and hence poor power. The distribution of **minor allele frequencies** (**MAFs**) over a large set of sites is called the **allele frequency spectrum**, and for neutral markers is generally strongly L-shaped, with most markers having rare alleles (the Watterson distribution; WL Equation 2.34). This distribution can be shifted by population history (e.g., bottlenecks or expansions) and past selection (WL Chapters 2, 8, and 9). Conversely, markers chosen for inclusion on a chip are generally biased towards more polymorphic markers (minor allele frequencies above one to five percent), namely, a bias toward **common SNPs**. To the extend that the marker frequency spectrum is discordant with the causal frequency spectrum (usually because a shift towards higher frequency values in markers), many causal sites will be missed. We return to the point in some detail below, and examine how whole-genome sequencing impacts GWAS. Lastly, while our focus here is on SNP-based GWAS, other classes of markers (Chapter 8) can also be used. One example are **copy number variants** (**CNVs**), which include duplicated or deleted regions (e.g., Ionita-Laza et al. 2009). Issues in detecting CNVs under different sequencing strategies are reviewed by Teo et al (2012).

Imputation

Even a very dense SNP chip only directly scores a tiny fraction of the existing genetic variation. Fortunately, the presence of strong LD among scored markers allows one to predict (or **impute**) the genotypes at unscored markers, *provided* that one has a scored reference set of haplotypes for these markers. Hence, given a sufficient reference set (representative of the population being sampled), only a modest fraction of these SNPs need actually be scored, allowing one to greatly extend the number of tested SNPs. The idea is similar to the multipoint mapping used for QTL mapping (Chapters 18 and 19), where known recombination rates allowed us to test for a QTL at arbitrary positions between adjacent markers. Impu-

tation (Li et al. 2009), and the closely related issue of haplotype phasing (Li and Stephens 2003; Scheet and Stephens 2006), is a fully mature field, with a highly technical literature, dissecting the virtues of various algorithms to accomplish these goals. Despite this richness, our discussion here will be brief. Applications of imputation in association mapping are reviewed by Servin and Stephens (2007), de Bakker et al. (2008), Guan and Stephens (2008), Marchini and Howie (2010), and Pei et al. (2010).

The basic idea is that one has a set of **genotyped markers** for individual i, a set of **reference markers** (known haplotypes with very dense scored markers), and a set of **target markers** that are present in the reference set, but were not scored in i. More formally, suppose our reference set is N haplotypes consisting of L scored SNPs over a particular genomic region. While the number of haplotypes could be as many of $N = 2^L$, in reality, SNPs are correlated because of LD, and the number of haplotypes is a much more modest number (**strong LD structure**; WL Chapter 9). Let $\mathbf{g}_i^T = (g_{i,1}, g_{i,2}, \cdots, g_{i,L})$ be the vector of SNP genotypes for individual i, where $g_{i,k}$ is the number of reference alleles at marker k. For scored markers, these values are 0, 1, or 2. However, some of these values are *missing* (unscored) and the idea of imputation is to use the reference collection of haplotypes to infer (impute) these missing values. Provided that the level of LD is high, the genotypes at target markers can generally be imputed with fairly high accuracy (e.g., Li and Stephens 2003; Marchini et al. 2007). The idea is that a haplotype with missing marker information is a *mosaic of the reference haplotypes*. Imputation involves estimating the break points between mosaic segments and inferring segment identity for intact regions. This is done using **Hidden Markov models**, which start at the first marker and then progressively moves along a chromosome through the L markers. These models use the local recombination rates, along with the observed marker data, to estimate the probability of a break after a marker (change of reference haplotypes) as one moves along the haplotypes for i. If a break occurs, the scored marker information and frequencies of reference haplotypes are used to compute the probability that a specific reference haplotype contributes the next segment. The resulting full set of markers (scored plus imputed) is often referred to as an ***in silico* genotype**. Note that imputation can only make inferences about SNPs that are present in the reference sample.

The accuracy of imputation can be checked by treating a scored marker as unknown and examining how accurately its genotype was predicted. Performing this operation over a large number of markers returns an estimate of the **imputation error rate**. Lin et al. (2010) proposed an **imputation quality score** (**IQS**), which adjusts the observed match frequency (P_0) for a called SNP by the probability (P_c) that it matches by chance,

$$IQS = \frac{P_0 - P_c}{1 - P_c}$$

A value of one indicates a perfect performance, while negative values indicate performance worse than expected by chance. More granularity can be obtained by considering the error over different local amounts of LD or over markers with different minor allele frequencies. The imputation error increases as LD and/or MAF decrease. Indeed, rare alleles are difficult to impute with accuracy. Part of this is due to the size of the reference sample, as imputation error decreases as the reference population size increases (Marchini and Howie 2010). By simulating very large reference sizes, Browning and Browning (2016) and Browning et al. (2018) showed that even rare alleles can be imputed with accuracy if the reference size is sufficiently large. They noted in a simulated reference panel of 200,000 Europeans, that markers with at least nine copies of the minor allele (MAF $\geq 4.5 \times 10^{-5}$) could be imputed with high accuracy ($r^2 \geq 0.8$). Huang et al. (2014) found that a variant with a MAF of 0.1% could be imputed with an $r^2 = 0.5$ using a sample of roughly 5,000 UK reference samples.

Most imputation methods return a posterior probability distribution of the number of reference alleles at an imputed marker, e.g., $\widehat{g}_{i,k} = (\widehat{g}_{i,k,0}, \widehat{g}_{i,k,1}, \widehat{g}_{i,k,2})$, where $\widehat{g}_{i,k,j}$ is the probability that the imputed copy number for marker k in individual i is j. One can extract

a summary statistic from these, such as the **best guess genotype**,

$$\widehat{g}_{i,k,max} = \max_j (\widehat{g}_{i,k,j})$$

namely, the genotype with the highest posterior probability, or the mean copy number (**posterior mean** or **allelic dosage**), $\widehat{g}_{i,k,ave} = \widehat{g}_{i,k,1} + 2\,\widehat{g}_{i,k,2}$, and then use these values in an association analysis. Alternatively, one could use the full posterior distribution to compute a weighted association statistic (e.g., Guan and Stephens 2008).

The use of imputation usually results in greater power for a GWAS, even when imputation accuracy is modest (Marchini et al. 2007; Guan and Stephens 2008; Marchini and Howie 2010; Pei et al. 2010). As might be expected from the behavior of the imputation error, power improves with higher LD and with higher MAFs for causal SNPs. A caution, however, is that high *average* imputation accuracy does not improve power in regions of low LD (Pei et al. 2010). The other important use of imputation is in meta-analysis (Appendix 6), wherein the results for a given trait from a number of studies are consolidated into a single analysis. If these studies used different SNPs, imputation can convert them all to a common set (e.g., de Bakker et al. 2008), provided that an appropriate reference sample exists.

BASIC STATISTICAL ANALYSIS OF GWAS DATA

As with single-marker linkage analysis (Chapters 18 and 19), each SNP in a GWAS is tested for a marker-trait association (significant differences in trait means over marker genotypes). We start with metrics for continuous traits and then examine discrete traits. These basic metrics are easily extended by adding cofactors, such as sex or age effects, and effects associated with the design, such as batch and specific-lab effects. A more delicate, but no less important, class of confounding effects are shared ancestry and population structure, and we treat these concerns in the next section.

Continuous Traits

Initially, one usually screens each SNP using the additive (or **gene-dosage**) model (Equation 18.14a; also called a **trend test**). Consider individual i and its genotype at a biallelic SNP marker locus k. We designate one of the SNP alleles as the **reference** (often chosen to be the minor allele) and let $N_{i,k}(= 0, 1, 2)$ be the number of reference alleles at marker k in individual i. These correspond to the nonreference homozygote, the heterozygote, and the reference homozygote, respectively (this could equivalently be coded as $-1, 0, 1$). From Equation 18.14a, the linear model for marker locus k becomes

$$z_i = \mu + b_k N_{i,k} + e_i \tag{20.1a}$$

where z_i denotes the trait value for individual i. For each reference allele added at marker k, the trait mean is changed by b_k. The test statistic for b_k can be written as $T^2 = (\widehat{b_k})^2/\sigma^2(\widehat{b_k})$ which follows an F distribution with 1 and $n-2$ degrees of freedom, which approaches a χ_1^2 distribution for large n.

If the SNP is declared significant, further insight might be gleamed by fitting the general genotype model, with (from Equation 18.29c)

$$z_i = \mu + b_k N_{i,k} + d_k H_{i,k} + e_i \tag{20.1b}$$

where $H_{i,k}$ is an indicator variable for being a heterozygote, so that $H_{i,k} = 1$ if i is a heterozygote at marker k, otherwise it has value zero. This is a two degrees of freedom test, and the significance of dominance ($d_k \neq 0$) can also be tested (one degree of freedom). While the trend test (with its fewer degrees of freedom) is usually more powerful than the general genotype test, the former has poor power for detecting fully recessive alleles, unless that allele is relatively frequent (Lettre et al. 2007).

Equations 20.1a and 20.1b form the basic framework upon which more complex models are built, usually by adding additional cofactors and random effects. For example, one could incorporate both a general sex effect (a sex-specific mean for the trait), and a sex-specific effect for an allele, using the model

$$z_i = \mu + s_i\mu_s + b_k N_{i,k} + c_k s_i N_{i,k} \tag{20.1c}$$

where s_i is an indicator variable for sex (0 for male, 1 for female). Under this model, μ and $\mu + \mu_s$ are the male and female means for our focal trait. Likewise the impact of each copy of the reference allele on the mean is b_k in males and $b_k + c_k$ in females. Using standard linear model machinery (Chapter 10), the effects of sex on the overall mean ($\mu_s \neq 0$) and/or sex-specific allelic effects ($c_k \neq 0$) can be tested. The general genotype model (Equation 20.1b) can be extended in the same fashion. The interaction with other cofactors is similarly modeled (Morris and Cardon 2019), such as epistasis (e.g., Equation 18.14d).

There are two rather different, but not necessarily exclusive, reasons for including additional factors in a linear model: ***reducing the residual variance*** and ***protecting against false-positives***. Including factors that influence the variability of a trait (such as age or sex differences) reduces the residual variance, potentially resulting in increased power to detect SNP effects. We say *potentially* because while their inclusion reduces the residual variance, it does so at the cost of degrees of freedom. Second, as detailed below, variables that introduce trait-SNP correlations in the absence of LD (such as shared relationships and/or population structure) can generate false-positives. The inclusion of factors accounting for such **confounding variables** can reduce the false-positive rate (see below).

An important caveat for the choice of covariates was noted by Aschard et al. (2015). Some covariates (e.g., height, body mass) are themselves heritable traits, and their inclusion can bias estimates of SNP effects when there are correlations between the covariate and the focal trait. Aschard et al. recommended either to avoid the inclusion of such covariates, or if one wishes to model then, to do so in a more formal multiple-trait GWAS framework (see below).

Adding SNPs as Cofactors

Segura et al. (2012) suggested that the power of a GWAS can be improved by incorporating SNPs with significant effects as cofactors. This is an extension of the composite interval mapping (CIM) method used in Chapter 18 for linkage mapping in inbred-line crosses. Recall that one feature of CIM was to include markers with significant effects that reside outside of the focal chromosome being testing. Their inclusion reduces the residual variance, increasing power. Segura et al suggested that the same is true for a GWAS, and proposed a stepwise approach for SNP inclusion. Suppose that SNP ℓ is the included cofactor, then when testing SNP k, the model becomes

$$z_i = \mu_i + b_\ell N_{i,\ell} + b_k N_{i,k} + e_i \tag{20.1d}$$

If the SNP cofactor shows dominance, then a $d_\ell H_{i,\ell}$ term can be added to the model. Once the most significant SNP is included in the model as a cofactor, a stepwise approach is used to scan for the next most significant SNP, with this process continuing until some stopping criteria is satisfied. Segura et al. applied their approach to both human and *Arabidopsis* data sets, and were able to identity new associations missed by previous analyses. This method is expected to be most powerful when there are causal sites with (at least) modest effects underlying the trait (as opposed to sites having nearly identical infinitesimal effects).

Discrete Traits: Contingency Table Analysis

Much of the motivation for human GWAS studies is the search for **disease susceptibility** (**DS**) **genes**. In this setting, the case-control design is typically used, where a series of cases are chosen by some criteria, and then matching controls are selected (e.g., WTCCC 2007). The resulting data for a given SNP are in the form of a contingency table, which can be expressed

either in terms of genotypes or alleles. For a diallelic SNP, the **genotypic contingency table** is given by

	Marker Genotype			
	mm	Mm	MM	Totals
Present	n_{P0}	n_{P1}	n_{P2}	n_P
Absent	n_{A0}	n_{A1}	n_{A2}	n_A
Totals	n_0	n_1	n_2	n

In the contingency table setting, a marker-trait association is indicated by a lack of independence between genotype and trait states (genotype frequencies differ between cases and controls). Thus, one could use a standard chi-square test for independence. This is a two degree of freedom tests and is the analog of the general genotype model for continuous traits. If one or more of the cell numbers in the table are small, the more precise Fisher's exact test can be used (Chapter 2). Alternatively, one could apply the Cochran-Armitage Trend Test (Equation 18.15), which uses one degree of freedom, and corresponds to the additive model (on the scale of measurement).

Instead of focusing on genotypes, one could simply count the number of reference alleles in each class (two for a homozygote, one for a heterozygote), yielding an **allelic contingency table**

	Marker Allele		
	m	M	Totals
Present	$2n_{P0} + n_{P1}$	$2n_{P2} + n_{P1}$	$2n_P$
Absent	$2n_{A0} + n_{A1}$	$2n_{A2} + n_{A1}$	$2n_A$
Totals	n_m	n_M	$2n$

Sasieni (1997) noted that the chi-square test is only appropriate for allelic data when Hardy-Weinberg holds. Under these conditions, the allele test asymptotically approaches the trend test on the genotypic data. There are a number of variants on these basic tests (such as combining results from different tests), which are reviewed by Kuo and Feingold (2010), who concluded that the trend test tends to be the most robust.

Example 20.2 The effects of genotypes on binary traits is usually expressed in terms of odds ratios (Example 19.10). Consider the data (from Example 18.5) of Zhang et al. (2005) on the association between genotypes at the DNA repair gene *ADPRT* and lung cancer:

	Genotype		
	mm	*Mm*	*MM*
Present	307	509	184
Absent	359	504	137
Odds	0.855	1.010	1.343
OR	1.000	1.181	1.571

To see these calcuations, consider the odds for genotype *mm*. From Equation 19.52a,

$$\frac{\Pr(\text{Disease Present} \mid mm)}{\Pr(\text{Disease Absent} \mid mm)} = \frac{307/(307+359)}{359/(307+359)} = \frac{307}{359} = 0.855$$

Setting *mm* as the standard, the odds ratio (OR) with respect to the other two genotypes are given in the table. Hence, relative to *mm*, the odds of the disease in *Mm* are increased by 18%, and by 57% for *MM*. Parameterizations for the effects of loci underlying threshold (discrete) traits are examined more detail in Chapter 30.

Example 20.3 The allelic contingency table for the data of Zhang et al. (2005) on the association between *ADPRT* genotypes and lung cancer becomes

	Marker Allele		
	m	M	Totals
Present	2·307 + 509 = 1123	2·184 + 509 = 877	2000
Absent	2·359 + 504 = 1222	2·137 + 504 = 778	2000
Totals	2345	1655	4000

The resulting χ^2_1 test statistic for independence is 9.8985, with an associated p value of 0.0017, whereas Example 18.5 computed the Cochran-Armitage statistic as 10.5733 (p = 0.0011).

Discrete Traits: Logistic Regression

An alternative strategy for discrete traits, which allows for cofactors and other confounders such as relatedness and population structure, is **logistic regression** (Truett et al. 1967; Chapter 14). The idea is that a linear model is constructed on some underlying (or *liability*) scale y (Chapter 30), which can potentially take on negative values and/or values greater than one. This liability value is then mapped via the logistic function (Equation 14.14a) into a range of (0,1), corresponding to the expected chance, $p(y_i)$, that an individual with a liability score of y_i displays the disease. Hence, we have an **observed value** (the disease state z_i that is either 0 or 1), an **underlying scale** (or **latent value**) y_i, and finally a mapping from the underlying scale y into an *expected value* on the z scale. The actual *observed value* of z_i follows a Bernoulli distribution (a binomial with a single draw; $n = 1$ in Equation 2.19a) with success parameter given $p(y_i)$. From Equation 14.15b, the logistic function mapping from y_i into the expected value of z_i is

$$E[z_i \mid y_i] = p(y_i) = \frac{1}{1 + \exp\left[-\left(y_i\right)\right]} \tag{20.2a}$$

Recalling Equation 14.14b, y is the predicted value of the **logit score**,

$$\text{logit}(p|y_i) = \ln\left(\frac{p|y_i}{1 - p|y_i}\right) = y_i \tag{20.2b}$$

Namely, the liability value y_i corresponds to the log of the **odds** (Examples 19.10 and 20.3), so that e^y corresponds to the expected disease odds for an individual with liability y. Note that Equation 20.2a maps y into p, while Equation 20.2b maps p into y.

Suppose on this underlying y scale, we assume an additive model for the effect of the kth SNP (Equation 20.1a) plus m added cofactors,

$$y_i = \mu + b_k N_{i,k} + \sum_{j=1}^{m} \beta_j x_{i,j} \tag{20.3a}$$

then

$$E[z_i \mid y_i] = p(y_i) = \frac{1}{1 + \exp\left[-\left(\mu + b_k N_{i,k} + \sum_j \beta_j x_{i,j}\right)\right]} \tag{20.3b}$$

and

$$\text{logit}(p|y_i) = \ln\left(\frac{p|y_i}{1 - p|y_i}\right) = \mu + b_k N_{i,k} + \sum_{j=1}^{m} \beta_j x_{i,j} \tag{20.3c}$$

where Equation 20.3c is the predicted value for the log of the odds of the trait being present in individual i.

On the liability scale, the interpretation of the b_k value for a SNP is clear. Under the additive model, each copy of the reference allele changes the log of the odds by b_k, so

that e^{b_k} is the change in the odds. Further note that the effects on the frequency scale are multiplicative, as the odds (OD) for y_i can be expressed as

$$\text{OD}(y_i) = \frac{p|y_i}{1-p|y_i} = \exp(y_i) = \exp\left(\mu + b_k N_{i,k} + \sum_{j=1}^{m} \beta_j x_{i,j}\right) \tag{20.3d}$$

$$= \exp(\mu) \cdot \exp(b_k N_{i,k}) \cdot \exp\left(\sum_{j=1}^{m} \beta_j x_{i,j}\right) \tag{20.3e}$$

$$= \text{OD(average)*OD(marker effect)*OD(Cofactor effects)} \tag{20.3f}$$

This decomposition shows that the odds for the trait being present in a given individual can be written as the product of the odds for a random individual, the odds for their marker genotype, and the odds associated with any of their additional (known) risk cofactors incorporated into the model.

Logistic regression is often used under a case-control design, and this can have subtle implications for the inclusion of cofactors into the model. In a linear regression setting, or in a logistic regression setting under a random population sample, the inclusion of covariates (such as age or sex) usually results in an increase in power. ***Such not need be the case in under a case-control design***. Under logistic regression, inclusion of a covariate increases both the estimated marker effect size *and* its sample variance (Robinson and Jewell 1991; Kuo and Feingold 2010; Clayton 2012; Pirinen et al. 2012). This implies that power can actually be *reduced* in some settings by inclusion of a covariate if the decrease in precision is larger than the increase in effect size. Pirinen et al. (2012) examined a simple logistic with and without a single covariate for three low-frequency human diseases, and found that the incursion of the covariate resulted in decreased power. For the case of Ankylosing spondylitis, a disease with a prevalence of 0.0025, the inclusion of a specific HLA risk factor resulted in a logistic regression that needed more than double the case sample size to have the same power as a logistic regression where this factor was excluded. Pirinen et al. found that when the disease is common (prevalence over 20%), covariate inclusion typically increased power, but decreases power when the disease is rare.

What might generate such an effect? Consider a case-control design, with a rare disease impacted by both genetic factors and, independently, by the covariant (e.g., smoking status). In the case population, both the rare genetic factor *and* the disease covariate are oversampled relative to their frequencies in the general population, which can create an association between the two in the overall GWAS, impacting power (Mefford and Witte 2012). Zaitlen et al. (2012a, 2012b) developed strategies for the appropriate conditioning on covariates in ascertained case samples that result in improved power.

Finally, one issue that can arise with a logistic regression is the problem of **separation**, in which the likelihood converges, but one of the estimates is infinite. This situation arises when, by chance, one of the covariates perfectly predicts the outcome, and can occur for rare cofactors (such as a rare allele) and/or in small samples or with sparse data. Firth (1993) propose a penalized likelihood correction for removing ML bias in certain settings, and Heinze and Schemper (2002) noted that the **Firth biased-corrected test** deals with separation (or near separation, where estimates are very unstable). Related approaches are reviewed by Mansournia et al. (2018).

Joint Testing of Multiple Markers: Gene-based GWAS

While above methods test markers (SNPs) one at a time, there are setting where it is advantageous to *simultaneously test a set of markers*. For example, there might be a set of weak signals over a region, that, when combined, result in a much stronger signal. When all of the tested markers reside within a given gene, this is called a **gene-based test for association** (Neale and Sham 2004), and is our focus here. In Chapter 21, we examine more general

gene set analysis (**GSA**) approaches, where the signals from all of the known genes in a given set or pathway are combined (e.g., Yu et al. 2009; Biernacka et al. 2012; Chen et al. 2019). Changing the unit of analysis to the gene or to a set of genes (or a pathway) can be considered a **secondary analysis** of the original GWAS, and can be accomplished either using the original data, or summary statistics (such as p values for each SNP).

There are several potential benefits of a gene-based, over a single-SNP based, analysis, but the relative advantage depends on the underlying trait architecture. If a causal gene typically has only a single causal site, then single-SNP analysis is favored. Conversely, if a gene has multiple causal sites, potentially of more modest effect, then a gene-based approach can be more powerful. A gene-based approach can also be ***more robust to genetic heterogeneity***. If there are multiple potential causal sites in the gene whose frequencies vary dramatically over samples, a single-SNP approach can miss these (even in a meta-analysis framework; see below), but a gene-based approach can capture them. This was seen by Peng et al. (2010), who examined type II diabetes in two different GWAS. They found no replication of *SNPs* (using genome-wide significance), but found seven *genes* that were successfully replicated over the two studies. Finally, the multiple comparisons burden can be somewhat alleviated with a gene-based analysis, as one is testing thousands of genes versus millions of SNPs.

We defer the important, and subtle, discussion of what group of SNPs constitutes a gene until Chapter 21. Gene-based GWAS approaches can be broken into two major categories. The first are approaches combining the p values for the scored SNPs in a gene. One advantage of these **combining methods** is that they only need summary-level statistics (and information on the correlation among SNPs for their extensions that do not assume independence). The second approach is more model-based, built around multiple regressions and their extensions (such as using PC scores as the predictor variables, penalized approaches such as ridge regression or the LASSO, or variance-components). These approaches typically require genotype data, not simply summary statistics, which can limit their applicability. Finally, there is considerable overlap between the multiple-SNP approaches considered here and rare-allele methods discussed later in the chapter.

As reviewed in Appendix 6, there are a number of methods for combining the individual p values for each tested marker into a single global p value for their gene. The foundational approach for such tests is that the distribution of p values under the null follows a uniform distribution. Suppose that one observes ten p values and the lowest 6 are all 0.10 (and hence not individually significant). This is a huge enrichment of small p values (under a unifrom, we expect only one in ten to have value ≤ 0.1), suggesting a signal for the *collection* of tests, but one that is too weak to be detected by any *single* test. More complicated distributions—such as using the first k of n ordered p values, or the distribution of p values below some threshold—can be easily computed by simulating draws of a vector of n independent uniform variables (or n correlated uniform variables in more general settings).

One of the most common p-value combining approaches is **Fisher's method** (Equation A6.1a): for k independent tests, where p_i denotes the p value for test i, the sum

$$X^2 = -2\sum_{i=1}^{k} \ln(p_i) \tag{20.4a}$$

approximately follows a χ^2_{2k} distribution. Loughin (2004) showed that while Fisher's method works well when the signal against the alternative is strong and concentrated in a small fraction of the tests, when a modest signal is spread over a number of tests, **Stouffer's Z score** (Equation A6.2) is more powerful. This method first translates the individual p_i values into unit normal scores, Z_i, and then computes an overall Z score,

$$Z_s = \sum_{i=1}^{k} Z_i/\sqrt{k} \tag{20.4b}$$

Further, as shown by Whitlock (2005), a weighted Z method (Equation A6.2.c) outperforms both Fisher and the unweighted Z when sample size varies (as would occur when allele frequencies vary over the tested markers). Bhattacharjee et al. (2012) discussed SNP weighting for a Z-based gene GWAS. The limitation with these approaches is the strong assumption of independence of tests, which fails for SNPs within a gene. Fisher's method has been extended to correlated tests by Brown (1975) and Makamb (2003), which requires estimates of the correlation among SNPs (e.g., Moskvina et al. 2011). More generally, one of the above statistics is computed, and its significance assessed via permutation (see below), but this usually requires access to the original data.

A number of modifications of Fisher's approach have been proposed that may improve its power and performance. Their starting point is to express Fisher's statistic as a product

$$X^2 = -2\ln\left(\prod_{i=1}^{k} p_i\right) \tag{20.4c}$$

Instead of considering the full product, various *truncated approaches* have been proposed that consider only a subset of the p values, namely those that are sufficiently small. The **truncated product method** (**TPM**) of Zaykin et al. (2002) considers only those p values below some threshold level (τ),

$$X^2_{trun} = -2\ln\left(\prod_{i=1}^{k} p_i\, I[\,p_i \le \tau]\right) \tag{20.4d}$$

where the indicator function $I[\,p_i \le \tau]$ has value one if $p_i \le \tau$, otherwise is zero. The logic is that by discarding modest to large p values, power may be improved, as most of the signal for region-wide significance resides in the smaller p values. A modification is the **rank truncated product method** (**rank-TPM**) of Dudbridge and Koeleman (2003). Here, one ranks the p_i values from smallest $p_{[1]}$ to largest $p_{[k]}$ and the product is taken over the $K < k$ smallest of these. We can extend truncated products to correlated tests, once again using permutation to obtain critical values. Further, these methods can be extended into **adaptive** versions (e.g., Yu et al. 2009; Chen et al. 2013; Yan et al. 2014; Pan et al. 2015), by varying the thresholds (τ or K), and then choosing the threshold yielding the smallest value (with significance again assessed via permutation). Even more generally, Fisher's method is a special case of the **Gamma method** (Zaykin et al. 2007; Biernacka et al. 2012), where the inverses of gamma functions (Table A7.1) are used to place differential weights on p values below some threshold. Zaykin et al. (2007) provides a nice overview these, and other, combining p-methods.

As mentioned, the significant of these approaches, even when markers are highly correlated, can be assessed using permutations, where the genotypes of an individual are kept intact (preserving their SNP correlation structure), but the phenotypes are randomized (shuffling the phenotypic labels over the genotype vectors). However, this can computationally intense, and requires access to the original data. Seaman and Müller-Myhsok (2005) developed **direct simulation approaches** (**DSAs**) to sample values from the distribution of truncated products under the null (the p_i follow a uniform distribution, but may be correlated), which is computationally much more efficient, and less restrictive, than a permutation approach.

A related approach to combination methods follows from corrections for multiple comparisons (Appendix 6). Recall that if we wish to have a FWER of γ, the Bonferroni correction over n independent tests uses a significance level for each test of $\alpha = \gamma/n$. One can essentially invert this as follows. Letting $p_{[i]}$ be the ith smallest p value, then the significance of the collection (the probability that none of the SNPs are significant) is approximately $p = np_{[1]}$ (p and $p_{[1]}$ taking the roles of γ and α, respectively). One could replace n with some measure of the effective number, n_e, of tests (Appendix 6), which can be estimated from the eigenvalues of the correlation matrix of the SNPs (Cheverud 2001; Nyholt 2004;

Li and Ji 2005; Patterson et al. 2006; Li et al. 2011). More powerful version of Bonferroni exists that discount for each significant test (Appendix 6). One such is **Simes-Hochberg correction** (Simes 1986; Hochberg 1988), which (following the logic above) leads to

$$p_{Simes} = \min_j \left(\frac{n\, p_{[j]}}{j} \right) \tag{20.4e}$$

which assumes uncorrelated tests. Li et al. (2011) proposed using an extended version of the Simes test, based on the effective number, n_e, of tests over the entire gene and the effective number, $n_{e[j]}$, based on the first j smallest p values. Their **GATES** (***g*ene-based *a*ssociation *t*est using *e*xtended *S*imes procedure**) test uses

$$p_{GATES} = \min_j \left(\frac{n_e\, p_{[j]}}{n_{e[j]}} \right) \tag{20.4f}$$

The second category of gene-based GWAS approaches are based on regressions and their extensions, which generally requires access to the raw genotype data (as opposed to summary statistics). One approach would be to use **Hotelling's** (1931) $\boldsymbol{T^2}$ **statistic** to jointly test the impact of n markers (Xiong et al. 2002; Chapman et al. 2003; Fan and Knapp 2003). If one considers a multiple regression using all markers in a defined set, $z_i = \mu + \sum \beta_j N_{i,j} + e_i$ (where $N_{i,j}$ is the number of copies of the reference allele at SNP j in individual i), then T^2 is a test that all of the β_k are zero. Recall that T^2 is the generalization of the Student's t test to multiple variables (for a single variable, T^2 reduces to t^2, the square of the t statistic). Again, the idea is that by jointly considering a well-chosen marker set, cumulative small signals might result in a larger significance for the collection as a whole. The tradeoff with a Hotelling-based approach is that inclusion of noncausal markers increases the degrees of freedom, without increasing the noncentrality parameter (the model signal from causal variants), lowering power (Appendix 5). An alternative approach is to use penalized regressions (Example 20.4) that downweight (or remove) model parameters with very low impact.

While Hotelling's test is generally performed using continuous trait data, it can also be modified for binary data. Xiong's T^2-based test for case-control samples is formatulated as follows: Let $g_{0,i,k}$ denote the genotype for the kth control individual at marker i (using the coding $-1, 0, 1$). For n markers of interest, let $\mathbf{g}_{0,k} = (g_{0,1,k}, g_{0,2,k}, \cdots, g_{0,n,k})^T$ be the vector of marker scores for control individual k, and $\mathbf{g}_{1,j} = (g_{1,1,j}, g_{1,2,j}, \cdots, g_{1,n,j})^T$ be the corresponding genotype vector for case indvidual j. For n_0 controls and n_1 cases, the Hotelling's T^2 test statistic becomes

$$T^2 = \frac{n_0\, n_1}{n_0 + n_1} \left[(\overline{\mathbf{g}}_0 - \overline{\mathbf{g}}_1)^T \mathbf{S}^{-1} (\overline{\mathbf{g}}_0 - \overline{\mathbf{g}}_1) \right] \tag{20.5a}$$

where $\overline{\mathbf{g}}_i = n_i^{-1} \sum_{j=1}^{n_i} \mathbf{g}_{i,k}$ (for $i = 0, 1$) are the mean vectors for cases and controls and

$$\mathbf{S} = \frac{1}{n_0 + n_1 - 2} \left[\sum_{k=1}^{n_0} (\mathbf{g}_{0,k} - \overline{\mathbf{g}}_0)(\mathbf{g}_{0,k} - \overline{\mathbf{g}}_0)^T + \sum_{j=1}^{n_1} (\mathbf{g}_{1,j} - \overline{\mathbf{g}}_1)(\mathbf{g}_{1,j} - \overline{\mathbf{g}}_1)^T \right] \tag{20.5b}$$

is the pooled sampling covariance matrix. One can think of T^2 as a generalized distance measure between the mean vectors, $\overline{\mathbf{g}}_0$ and $\overline{\mathbf{g}}_1$, of marker frequencies in cases versus controls. Asymptotically,

$$\left(\frac{n_0 + n_1 - n - 1}{n(n_0 + n_1 - 2)} \right) T^2 \sim F_{n, n_0 + n_1 - n - 1} \to \chi^2_n \tag{20.5c}$$

The last step follows because as the denominator degrees of freedom for an F distribution becomes large, it approaches a χ^2 with the numerator degrees of freedom (Appendix 5).

This notion of contrasting the differences in multilocus genotype similarity between some defined set of markers in cases and control has been extended by other authors (e.g., Schaid et al. 2005; Wessel and Schork 2006; Kwee et al. 2008). This approach is often framed using **kernel methods** (Schaid 2010; Larson et al. 2018)—which can be loosely thought of as generalized distance measures—such as the **kernel-based adaptive cluster**, **KBAC**, method of Liu and Leal (2010) and the **kernel-based association test**, **KBAT**, of Mukhopadhyay et al. (2010). Many of these approaches can also be framed as random-effects models, where the test for a gene effect can be based in terms of variance components (Chapter 32).

Treating SNPs as single markers and then accounting for their correlation structure is one approach for modeling under high levels of LD. The other is to base the unit of analysis on **haplotypes**, the specific configuration of SNP alleles over a small section of a chromosome. Coding the minor allele as one and the major as zero, the haplotype can be encoded as a string of zeros and ones, so that $h_1 = (00001)$ would be a haplotype with major alleles at the first four SNPs and the minor allele at the fifth. By focusing on haplotypes, the unit of analysis moves from a set of diallelic SNPs to a single marker with h alleles (one for each of h defined haplotypes). While conceptually straightforward, this approach raises several issues. The first two are operational: how to define an optimal haplotype length (a sliding window coupled with a cross-validation approach might offer a solution) and how to obtain their phase (extracting the two haplotypes given a diploid multilocus genotype; Chapter 5). The final issue is statistical: haplotype models quickly soak up degrees of freedom. With h haplotypes, the degrees of freedom range from $h - 1$ on the low end (analyzing haplotypes as additive alleles: extending Equation 20.1a and allelic contingency tables to h alleles) to $h(h - 1)/2$ on the high end (the full genotype model: extending Equation 20.1b and the genotype contingency tables to all diploid genotypes $h_i h_j$). Further, one would expect a small number of common haplotypes and a larger number of rarer ones. As a result, a haplotype-based analysis often focuses on the relatively common haplotypes, and might lump the rest into a single group (similar to rare alleles approaches considered below). As noted above, one approach to deal with such high-dimensional models is to use penalized regressions (Example 20.4)

The basic structure of a haplotype-based analysis treats each haplotype as an allele in an otherwise standard marker analysis, e.g., using the regression $z_i = \mu + \sum_j \beta_j H_{i,j} + e_i$. Here $H_{i,j}$ is either the number of copies of haplotype j in individual i (when they can be scored directly), or can be replaced an expected value, $H_{i,j} = \Pr(\text{haplotype } j \mid \mathbf{g}_i)$, where $\mathbf{g}_i$ is the vector of SNP values for individual i over the region of interest (e.g., Schaid 2002; Schaid et al. 2002).

An intermediate approach that nicely accommodations the challenging issues of defining haplotypes and controlling their dimensionality is to use **principal components** (**PCs**; Chapter 9) of the covariance matrix $\mathbf{C}$ for the markers of interest. Let $\mathbf{e}_i$ be the eigenvector associated with eigenvalue λ_i (ranked from largest to smallest). Recall that the sum of the first k eigenvalues divided by the total sum of eigenvalues gives the amount of variation explained by the first k PCs (Chapter 9). Gauderman et al. (2007) and Wang and Abbott (2008) suggested that PCs can be used as the predictor variables in a regression, with

$$z_i = \mu + \sum_{j=1}^{k} \beta_j \gamma_{i,j} + e_i, \quad \text{where} \quad \gamma_{i,j} = \mathbf{e}_j^T \mathbf{g}_i$$

Thus, $\gamma_{i,j}$ is the projection of the SNP data from the focal gene ($\mathbf{g}_i$) from individual i onto the jth PC. Gauderman et al. suggested that k should correspond the the number of eigenvalues that accounts for between 80 to 90 percent of the total scored SNPs variation in the target gene. Gauderman et al. and Ballard et al. (2010) found that PC-based approaches usually out perform haplotype or multiple SNP regressions. One standard concern with any PC method is how to interpret the PC axes (Chapter 9). However, when a causal SNP is found on several different haplotypes (diffusing its haplotype-specific signal), the SNP weighting on the first few PCs could pick out such a signal.

As with the various p-value combining methods discussed above, given the strong dependency (LD) between SNPs within a genic region, the gold standard for regression-based significance testing is permutation. A general summary statistic is obtained for the observed SNP set, and then trait values are shuffled over individuals, while keeping their marker genotypes intact, generating a test statistic under the null (e.g., the **set-based test** in PLINK; Purcell et al. 2007). When a large number of genes are tested, permutation becomes computationally intense due the small critical values required to correct for multiple comparisons. To obtain a reasonably stable estimate of the test value corresponding to probability α under the null requires roughly (at least) $10/\alpha$ permutation samples, which can be the order of 10^7 or greater when a large number of genes are tested. One interesting solution to this problem was suggested by Knijnenburg, et al. (2009), drawing on results from extreme value theory (WL Example 27.2). Under rather general assumptions, the extreme tails of a probability distribution converge to a generalized Pareto distribution (WL Equation 27.7), whose parameters can be estimated using ML. This allows a modest number of simulations (say 5000 to 10000) to be used to estimate its shape and then extreme values (say the threshold value corresponding to $p = 10^{-8}$) are computed analytically from the resulting distribution.

As a result, a number of authors have proposed simulation-based approaches (drawing test statistic values from the expected null distribution), with the notion that the computational burden for a single simulation draw is much less than for a single permutation value. Lin (2005) proposed a general scheme for many score-based tests using draws of independent normal random variables which are then adjusted by the correlation structure of the data. A similar approach (**VEGAS**, for **versatile gene-based association study**) was proposed by Liu et al. (2010). Even for these simulation approaches, the computation burden of obtaining $10/\alpha$ values can be daunting. One potential solution is **fastBAT** approach of Bakshi et al. (2016). They noted that, under the null, the test statistics for many of the above tests can be written as the quadratic products of a MVN (Chapter 9, Appendix 3), and one can use standard large-sample MVN quadratic product approximations (e.g., Davies 1973; Kuonen 1999) to obtain critical values, as opposed to using simulations.

Example 20.4 A common situation that arises in modern quantitative genetics are regressions whose number of parameters p exceeds, often greatly, the sample size n. In a standard least-square regression framework, estimation proceeds using generalized inverses (Appendix 3), resulting in a set of solutions. A more powerful approach is to use **penalized** (or **regularized**) **regressions** (Chapter 31). Consider the regression model

$$y_i = \mu + \sum_{j=1}^{p} \beta_j X_{i,j} + e_i$$

In the standard OLS framework, one solves for the β_j that minimizes the sum of squared residuals,

$$\text{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \mu - \sum_{j=1}^{p} \beta_j X_{i,j} \right)^2$$

Penalized regressions start with this framework, and then place constraints on the β_j. Under **ridge regression** (**RR**), one instead minimizes RSS $+\lambda \sum \beta_j^2$ (Hoerl and Kennard 1970), for some **shinkage parameter** $\lambda > 0$ (Hoerl et al. 1975; Lawless and Wang 1976, and Cule at al. 2011 discuss the choice of λ). An alternative approach is the **least absolute shrinkage and selection operator**, or **LASSO**, which minimizes RSS $+\lambda \sum |\beta_j|$ (Tibshirani 1996). Note that for values of β near zero, β^2 is a much less harsh constraint than $|\beta|$ (as $|\beta| \gg \beta^2$ for $|\beta| \ll 1$), so that the LASSO shrinks most of the β_j to exactly zero (yielding a **sparse estimate**), and hence is often used in **model selection** (choosing the parameters in the final model as those with nonzero βs). Finally, the two approaches are combined in the **elastic net** (Zou and Hastie 2005), which seeks to minimize RSS $+\lambda_1 \sum |\beta_j| + \lambda_2 \sum \beta_j^2$. These approaches can also be

extended to generalized linear models, e.g., Le Cressie and van Houwelingen (1992) proposed a logistic ridge regression method.

The use of penalized regressions for marker selection, treating epistatic interactions, and dealing with a set of very highly correlated markers in a GWAS setting has been examined by a number of authors (e.g., Li et al. 2007; Malo et al. 2008; Wu et al. 2009, 2011; Chen et al. 2010; Han and Pan 2010; Zhou et al. 2010; Li et al. 2011; Xu et al. 2012; Gao et al. 2013; Yi et al. 2015; Schaid et al. 2020).

Detecting Gene-gene (Epistatic) Interactions

As was the case of linkage-based mapping (Chapter 18), it is trivial to extend GWAS to search for epistasis by looking for interactions between marker genotypes. For example, the presence of epistatic interactions between the genotypes at two different SNP markers can be tested as follows. Let $\overline{z}_{ij}$ be the mean value of individuals with two-locus genotype ij, with i indexing the SNP genotypes at locus i (values of 0, 1, 2), and let $\overline{z}_i$ be the mean trait value for single-locus genotype i. A simple test for epistasis given by

$$T = \sum_{i=1}^{3}\sum_{j=1}^{3} \left(\overline{z}_{ij} - \overline{z}_i - \overline{z}_j + \mu\right)^2$$

Under the null (additive interaction) $E[T] = 0$ and T follows an F distribution with $n - 9$ degrees of freedom (n being the sample size), while $E[T] > 0$ under the alternative (epistasis). In a GWAS setting, this, and other, simple tests can be greatly complicated by three factors: multiple comparisons, power, and the effects of imperfect tagging of causal sites by markers. We consider these in turn. Complications induced by imperfect tagging turn out to have especially subtle effects and can generate false positives in non-obvious settings. There is a growing, and increasingly technical, literature on the search for GWAS epistasis, see Cordell (2009) and Wei et al (2014) for more detailed reviews.

The major complication in the search for epistasis under linkage mapping was the massive explosion in the number of comparison as the number of markers increased, basically scaling as m^2 (more exactly, $m[m-1]/2$), a concern that is greatly magnified under a GWAS. As discussed in Chapter 18, this increase in model space can be handled by either using a hypothesis-free approach (an **exhaustive search** over all such combinations) or by a hypothesis-driven approach (such as testing specific loci, or, more generally, loci first detected by having marginal effects). The latter approach, called a **two-step scan**, can either be based on one, or both, of the markers showing a significant marginal effect (Chapter 18). Marchini et al. (2005) and Evans et al. (2006) examined these different strategies, and found that there were situations where an exhaustive search (despite its far greater multiple-testing burden) had higher power than a two-step approach. This required that the interaction variance is greater, often far greater, than either of the marginal additive variances for the two candidate loci, which in turn, requires that both loci have minor allele frequencies near 50% (Chapter 5). In a GWAS setting, satisfying this condition is generally unlikely, even when the pair of loci show strong functional epistasis, as this still generally results in weak statistical epistasis, the signal detected by a GWAS.

Generally speaking, the GWAS signatures for (statistical) epistasis (significant epistatic variances) are weak. As noted by Wei et al. (2014), "*searching for epistasis has contributed rather little to the understanding of complex traits, except for the important observation that large interaction terms are very unlikely to exist between pairwise SNPs*". However, just as the finding of generally small additive variances *does not* imply the absence of major-effect alleles (they are just at low frequencies), the observation of (at best) weak epistatic variances does not imply that functional epistasis is generally absent, but rather that even if it occurs, the two-locus genotype frequencies are such that most effects are loaded onto the additive variance (Chapter 5; Purcell and Sham 2004; Hill et al. 2008; Mäki-Tanila and Hill 2014).

Even under a very dense GWAS, causal QTLs can be imperfectly tagged by marker SNPs. For a single causal locus, only r^2 of its additive variance is captured by a marker

in LD, while only r^4 of its dominance variance is captured by the marker (dominance) variance. With epistasis, even less is captured. If $\sigma^2(A^kD^\ell)$ denotes the true causal epistatic variance, the resulting marker epistatic variance (assuming each causal site is correlated by r to a SNP) becomes $r^{2k}r^{4\ell}\sigma^2(A^kD^\ell)$, so that only fractions r^4, r^6, and r^8, respectively, of the $\sigma^2(AA)$, $\sigma^2(AD)$, and $\sigma^2(DD)$ causal variances are capture by the marker pair (Wei et al. 2014). The impact of imperfect LD on estimating epistatic variances is thus far greater than its impact on detecting single-marker effects (additive variances). Coupled with the least-squares decomposition tending to place most of the variance into additive terms (Chapter 5), it is thus not surprising that estimates of marker epistatic variances, when even detected, tend to be very small. This, coupled with the much greater multiple comparison burden, means that there is generally very low power to detect most epistatic interactions.

A final complication is that even when an epistatic signature is detected, the subtle effects of imperfect tagging of a causal SNP by a marker can result in false positives, resulting in **phantom epistasis**. This can occur over two scales, between linked loci, and, surprisingly, between unlinked loci. As noted by Wood et al. (2014), Hemani et al. (2014), and de Los Campos et al. (2019), **haplotype effects** can occur when two marker loci are both in imperfect LD with a causal SNP, say $M1 - Q - M2$. When the LD is imperfect, the causal additive variance is not fully accounted for by either marker alone, resulting in their interaction term, capturing part of this residual, being significant. Hence, under a purely additive model, imperfect LD can generate non-additive signals between markers in LD. Finally, even when marker loci are unlinked *and* in linkage equilibrium with each other, imperfect tagging of causal sites can still generate phantom epistasis. This arises by skewing the distribution of T away from a normal (the effects of a genotypic class now follow a mixture, instead of a normal, distribution; Chapter 16). This can inflate the F statistic over its expected value under the null, generating false positives (see Hemani et al. 2021 for details).

Corrections for Multiple Testing

Given the massive numbers of tested markers (either directly or via imputation) in a GWAS—orders of magnitude higher than in a linkage analysis—adjustment for multiple testing is a serious concern. Recall that linkage-based mapping considered two multiple comparison scenarios: the **sparse-map approximation** with roughly independent markers (in which case Bonferroni corrections can provide reasonable control), and the **dense-map approximation** with correlated markers (essentially testing every position along a chromosome) wherein more elaborate modeling is required (Chapters 18 and 19). The situation under a GWAS falls somewhat in between these two settings. Under linkage mapping, the signal from a QTL can propagate over most of a chromosome (albeit with its impact diminishing over distance). In contrast, the LD structure for many populations is such that a chromosome can be considered as a series of very small LD blocks, within which LD is very high, but with near independence between blocks. This difference of long-range correlations under linkage mapping versus only very sort-range correlations in a GWAS is a function of the former experiencing only a few meioses per chromosome, whereas the later may involve thousands (or more) of recombinations since the common ancestor.

Because of LD block structure, the *effective* number of independent tests (n_e) is less than the actual number, with the Bonferroni correction testing each marker at $p = \gamma/n_e$ to provide a genome-wide Type I error control of γ (Equations 18.13a and 18.13b). A number of approaches using the eigenstructure of the correlation matrix of tests have been proposed to estimate n_e (e.g., Cheverud 2001; Nyholt 2004, 2005; Ji and Li 2005; Patterson et al. 2006; Appendix 6), most of which are very conservation under strong LD structure (Salyakina et al. 2006; Moskvina and Schmidt 2008). However, the **simpleM** method of Gao et al. (2008) seems to perform fairly well in such settings (Gao et al. 2010; Gao 2011; Hendricks et al 2014; Davis et al. 2016). Ranking the eigenvalues of the test correlation matrix from largest (λ_1) to smallest (λ_m), the Gao estimator of n_e (for n tests) is the value that satisfies

$$\sum_{i=1}^{n_e} \lambda_i / \sum_{i=1}^{n} \lambda_i = C$$

namely the smallest number of eigenvalues that account for a fraction C (typically taken as 0.995) of the total variation (sum of all eigenvalues). One can also use Bayesian approaches to estimate the effective number of tests (Q. Wang et al. 2015; S.-B. Wang et al. 2016).

An alternative approach is to estimate the number of LD blocks and take this as the value of n_e. In their landmark paper promoting association over linkage analysis. Risch and Merikangas (1996) suggested that a modern GWAS might involve testing up to 500,000 sites, for a $p = 1 \times 10^{-7}$ (choosing a γ of 5%). More recent suggestions in humans are around 10^6 LD blocks for European populations yielding $p = 0.05/10^6 = 5 \times 10^{-8}$ (Altshuler et al. 2008; Fadista et al. 2016), and double that number for African populations (Pe'er et al. 2008) with their deeper ancestry, and hence more generations of recombination in a population sample. Estimates of the rough number of independent LD blocks in other species can be much higher (as in outbred maize populations) or much lower (as in a panel of elite inbred lines). In theory, a large set of SNPs with good genomic coverage could be used to estimate the number of blocks (e.g., Example 20.5), or at a minimum, suggest the average spacing of SNPs for adequate power. With a value of n_e in hand, one can further improve on Bonferroni methods using sequential corrections (Manly et al. 2004; Appendix 6). In the simplest setting (Holm's method), the p values are ordered $(p_{[1]}, \cdots, p_{[n]})$, and if the smallest is rejected (e.g., $p_{[1]} \leq \gamma/n_e$), then the next smallest is tested at $p_{[2]} \leq \gamma/(n_e - 1)$, and so on until failure to reject. More powerful sequential approaches (e.g., Simes-Hochberg and Hommel) are reviewed in Appendix 6.

A second approach to control for multiple comparisons are permutation tests (reviewed in Chapter 18). Where feasible, these are generally regarded as the gold standard for multiple comparison corrections. The idea is simple: one keeps the marker data intact for each individual, but randomly shuffles their trait values. This generates a sample from the null (no marker-trait association), so that a GWAS performed on this data has all markers under the null. Repeating this procedure a few thousand of times generates a distribution of the smallest genome-wide p value, the smallest for any particular chromosome, etc., under the null. However, given the size and scale of a modern GWAS, resampling incurs a very heavy computational burden. A further complication is that a permutation requires that one has identified **exchangeable units**, which is not trivial in the presence of population structure.

A computationally more efficient approach is to approximate a permutation test by using large-sample theory. Under the null, a vector of test statistics is asymptotically multivariate normal (MVN) with a correlation structure than can be estimated from marker data. One can then either sample from this distribution or use numerical interaction to estimate critical values (Lin 2005; Conneely and Boehnke 2007; Han et al. 2009). Hendricks et al. (2014) noted that such applications of **extreme tail theory**, and well as the Gao n_e estimator, were efficient alternatives to permutation.

Besides the obvious issue of adjusting for multiple, correlated tests, a much more subtle issues arises in a modern GWAS. Subsequent to the original analysis, imputation and the combination of results from multiple studies are often done, resulting in ***the number of GWAS tests initially performed being rather different from the number of tests that ultimately derive from this data*** (Thomas and Clayton 2004). Consider the setting where one million SNPs are directly scored and tested. Using this group of markers, one could estimate n_e for a Bonferroni-style correction or use a permutation approach based on this set of markers. Now suppose that a reference set of sequences becomes available that allows for the imputation of an addition two million SNPs on the same dataset. This would involve no change in the number of scored individuals, but results in the original data now being used for three million tests. The genome-wide threshold p value obtained via the original permutation is no longer appropriate as additional tests are involved. The very real possibility exists that highly significant markers detected in the initial study would no longer achieve genome-wide significance as additional markers (and therefore tests) are added. Likewise, estimates of n_e based on correlations among the original tests are also no longer appropriate.

Such concerns led to the extension of the infinitely dense marker approximation for

linkage studies from Chapters 18 and 19 (where all tests on a chromosome are correlated because of insufficient recombination) to an analog of the infinitely dense setting in an LD framework. Here there are a large number of nearly independent islands of SNPs, within which marker correlations are extremely high. Example 20.5 discusses one approach for obtaining a limiting threshold value in such **fully saturated** marker settings, effectively adjusting for all current, and possible future, single-marker association tests based on the original data.

Example 20.5 Dudbridge and Gusnanto (2008) proposed an interesting approach for obtaining the fully-saturated threshold value via permutation testing using increasing marker density. Their idea was that the threshold value increases with the number of tests, but ultimately approaches some asymptotic value. They took a set of SNPs (from the target population of interest) and randomly assigned them into cases and controls, and then performed a permutation test to obtain the critical p value required for a FWER of γ. They then examined ever-smaller subsets of the SNPs, with $f(x)$ denoting the critical p value for x scored SNPs, which is expected to decrease as x increases. The resulting number of effective tests for a given value of x to obtain a global γ FWER is $n_e(x) = \gamma/f(x)$. The authors suggested fitting a least-squares regression on such data using the **Monod function**

$$n_e(x) = \frac{\mu x}{h + x} + e \tag{20.6a}$$

The estimate of h (the **half-saturation parameter**) is the number of SNPs to achieve half of the threshold value, $f(h) = \mu/2$, and $\widehat{\mu}$ is the limiting value, so that using

$$p = \gamma/\widehat{\mu} \tag{20.6b}$$

provides the fully-saturated Type I error control of γ. An alternative estimator for n_e, obtained by fitting a beta distribution (Equation A7.37) for the observed minimum p values over the permutation samples, was developed by Dudbridge and Koeleman (2004), Dudbridge and Gusnanto (2008), and Saffari et al. (2016). See their papers for details.

Dudbridge and Gusnanto applied Equation 20.6a using a set of roughly 360,000 SNPs from a UK Caucasian population. The limiting threshold was around $\widehat{\mu} \simeq 693{,}000$, with Equation 20.6b giving the fully-saturated threshold (for $\gamma = 0.05$) as $p = 7.2 \times 10^{-8}$ for this population. They also noted that an eigenvalue-based estimate of n_e (Patterson et al. 2006) based on the 360,000 SNPs returned a value of around 33,000, an order of magnitude too small, while the permutation test for the full set of SNPs returned a value of roughly 228,000 SNPs, also below the fully-saturated value.

Bonferroni, sequential Bonferroni, and resampling are all examples of **family wide error rate** (**FWER**) approaches (Manly et al. 2004; Rice et al. 2008), where the goal is controlling the experiment-wide (here **genome-wide**) error rate. An alternative approach is the **false discovery rate** (**FDR**) framework (Appendix 6). The motivation for Bonferroni is the belief that all of the tests are likely under the null, while FDR assumes that some small fraction are likely *not* from the null. In the FDR setting, control is not over *all tests* (as is done with FWER procedures), but rather over ***all tests declared to be significant***. An FDR value of 5% implies that, of the declared significant tests, at most 5% are false discoveries (i.e., the true marker effect is zero). As detailed in Appendix 6, the basic underpinning of FDR amounts to adjusting the p-value threshold, π_δ, for declaring a test to be significant until the desired FDR is achieved. Under the assumption that very, very few of the n tests are true positives, the expected number of false discoveries is $n\pi_\delta$, while $n_{S(\delta)}$ is the observed number of tests with $p \leq \pi_\delta$, and hence are declared to be significant. The FDR thus becomes $n\pi_\delta/n_{S(\delta)}$. Decreasing the value of π_δ decreases both $n\pi_\delta$ and $n_{S(\delta)}$, and one adjusts the threshold value until this ratio achieves the desired FDR value (e.g., 0.05); see Appendix 6 for details.

FDR represents an attempt to strike a balance between **sensitivity** (Type II error, the power to detect a difference, e.g., avoiding false negatives) and **specificity** (Type I error, avoiding false positives). The Bonferroni framework, with its extremely conservative p values, has high specificity but poor sensitivity as many markers with small effects do not generate enough of a signal to exceed the extreme p threshold required to be declared significant. The FDR approach lowers this threshold (improving sensitivity), while still providing good specificity (Manly et al. 2004). The goal of an initial GWAS is usually to distill out a small set of candidate regions for future study. In such a setting, an FDR framework is usually better than an FWER approach, as the latter may exclude important regions from future study. While false positives occur in the FDR candidate set, their number is controlled, and the cost of including such false candidates is often offset by much greater sensitivity, and hence the inclusion of true candidates that might otherwise be excluded. Brzyski et al. (2017) noted that an FDR analysis applied to GWAS is not truly on a SNP-by-SNP basis, as blocks of SNPs in LD will share the same signal from a single causal site. Rather, the FDR unit of analysis is a cluster of SNPs in LD. This is the same issue noted by Chen and Storey (2006) for linkage mapping (Chapter 18 and Appendix 6).

Finally, a rather different approach was suggested by Wacholder et al. (2004), WTCCC (2007), and subsequent authors (Thomas and Clayton 2004; Wakefield 2007, 2008, 2012), namely using a Bayesian framework. An excellent discussion of Bayesian approaches for multiple GWAS comparisons is given by Stephens and Balding (2009). As suggested by Thomas and Clayton (2004), the basic tenet of this framework is that

> "it is not the number of tests performed but rather the prior credibility of the hypotheses that is important in interpreting a set of observed associations. That is, when a hypothesis is unlikely to be true *a priori*, we should require strong evidence to be convinced of its truth"

In this framework, the strength of evidence can be expressed as the **posterior odds ratio** in favor of a true association. Letting T denote the value of the test statistic and τ the significance threshold, then the odds ratio in favor of a true association when the test is deemed significant can be written as

$$\begin{aligned}\frac{\Pr(H_1 \mid T > \tau)}{\Pr(H_0 \mid T > \tau)} &= \frac{\Pr(T > \tau \mid H_1)\Pr(H_1)/\Pr(T > \tau)}{\Pr(T > \tau \mid H_0)\Pr(H_0)/\Pr(T > \tau)} \\ &= \left[\frac{\Pr(T > \tau \mid H_1)}{\Pr(T > \tau \mid H_0)}\right]\left[\frac{\Pr(H_1)}{\Pr(H_0)}\right] = \frac{1-\beta}{\alpha}\,\frac{\Pr(H_1)}{\Pr(H_0)} \qquad (20.7a)\end{aligned}$$

where the first step follows from Bayes theorem (Equation 3.3b). Here α is the Type-I error rate, β the Type-II error rate (for a power of $1-\beta$), and $\Pr(H_1)$ and $\Pr(H_0)$ are the prior probabilities for, and against, an association. To apply Equation 20.7a, one must have some loose idea about the fraction of independent regions that generate associations with the trait, and some details about the effect size (and frequency) in order to specify β. WTCCC (2007) suggested values for $\Pr(H_1)$ in the range of 10^{-4} to 10^{-6}.

Equation 20.7a is very closely related to Morton's (1955b) **posterior error rate** (**PER**)—which Wacholder et al. (2004) refer to as the **false positive report probability** (**FPRP**)—see Appendix 6. Denoting the posterior odds ratio (Equation 20.7a) as PO, an alternative metric is the **posterior probability of association**, **PPA** (Stephens and Balding 2009), where

$$PPA = \frac{PO}{1+PO} \qquad (20.7b)$$

Finally, a more general approach is to replace

$$\frac{\Pr(T > \tau \mid H_1)}{\Pr(T > \tau \mid H_0)} \quad \text{with} \quad \frac{\Pr(T \mid H_1)}{\Pr(T \mid H_0)} \qquad (20.7c)$$

Namely, replacing a threshold being exceeded with the actual value, T, of the test statistic for that marker. This ratio of support for the data (T) under the alternative versus the null

hypothesis is called a **Bayes factor**, **BF** (Appendix 7). Computing the BF requires assumptions about the prior distribution of allele effects (and their associated allele frequencies), see Wakefield (2007, 2008, 2012) and Stephens and Balding (2009) for details. Stephens and Balding make the important point that as GWAS analysis moves beyond one-marker-at-a-time considerations to more complex units of interactions, Bayesian approaches can offer much more flexibility than frequentist methods.

Example 20.6 As an application of Equation 20.7a, consider the scenario where ten regions, out of one million LD blocks, influence the trait of interest, and suppose that there is 50% power to detect each effect. To obtain an odds-ratio of ten to one in favor of a true effect, rearranging Equation 20.7a gives a required α value of

$$\begin{aligned}\alpha &= \left(\frac{\Pr(H_0 \mid T > \tau)}{\Pr(H_1 \mid T > \tau)}\right)\left([1-\beta]\,\frac{\Pr(H_1)}{\Pr(H_0)}\right) \\ &= \frac{1}{10}\cdot 0.5 \cdot \frac{10/10^6}{1-10/10^6} = 5\times 10^{-7}\end{aligned}$$

Under these parameters, we expect an average of $0.5 \cdot 10 = 5$ true discoveries and $5 \times 10^{-7} \cdot (10^6 - 10) = 0.5$ false discoveries, for an expected FDR of 0.5/5.5, or around 9%. The PPA for an odds ratio of 10 (PO = 10) becomes PPA = 10/11 = 0.909. Similarly, for a posterior odds ratio of 20 (with a resulting PPA of 20/21 = 0.952), $\alpha = 2.5 \times 10^{-7}$, which yields an expected 0.25 false discoveries and an expected FDR of 0.25/5.25, or slightly under 5%.

Now suppose a highly polygenic trait, such as height, which may have a large number of regions (say 1000), but, given their smaller effect sizes, lower power of detection (say $\beta = 0.8$, or a power of 20%). The critical value for a odds ratio of ten becomes

$$\alpha = \frac{1}{10}\cdot 0.2 \cdot \frac{1000/10^6}{1-1000/10^6} = 2\times 10^{-5}$$

In this setting we expected an average of $0.2 \cdot 1000 = 200$ true discoveries and $2 \times 10^{-5} \cdot (10^6 - 1000) \simeq 20$ false discoveries, for an expected FDR of 20/220, or again around 9%. For a posterior odds ratio of 20:1, $\alpha = 1 \times 10^{-5}$, yielding an expected FDR of 10/210, again slightly under 5%.

Power, Replication, and the Winner's Curse

While it would seem that a modern GWAS, with its very large sample size, has ample power, this ignores two realities. First, most detected GWAS effects are small. This is perhaps not surprising, as the impact of a marker on a GWAS usually appears through the *variance* attributable to the linked causal site. For an additive QTL, this is $2a^2p(1-p)$, where a is the allelic effect, with this signal further attenuated by r^2 due to scoring the effect using a marker in LD with the causal site. In nature, large effect alleles tend to be at low, to very low, frequencies, most likely as a result of past selection against them (Chapter 21; WL Chapter 28). Further, because common SNPs are typically the markers of choice, the mismatch in frequencies between a marker allele and a rare causal allele results in a small r^2 value (Example 20.1), even under complete disequilibrium ($|D'| = 1$), further reducing the signal from the causal site. Hence, a small GWAS effect *does not* imply a small effect allele, but rather that the causal site only accounts for a small fraction of the total trait variance. Second, the blessing (and curse) of a modern GWAS is the vast number of markers that are considered, resulting in very small α values being used to test each marker in order to account for multiple comparisons (as just discussed). The cost of such increased specificity is a substantial loss of sensitivity (power).

Table 20.1 Critical values for χ^2 tests with one or two degrees of freedom (df) for a given α, and the associated NCP values (λ) required for a desired level of power. See Example 20.7.

			Required λ for a power of		
α	df	Critical	50%	80%	90%
10^{-6}	1	23.93	23.93	32.87	38.11
	2	27.63	26.62	36.11	41.62
$5 \cdot 10^{-7}$	1	25.26	25.26	34.43	39.79
	2	29.02	28.01	37.71	43.35
10^{-7}	1	28.37	28.37	38.05	43.67
	2	32.24	31.23	41.43	47.32

One consequence of low power, *lack of repeatability*, led to considerable initial skepticism about the validity of GWAS detected sites (often referred to as **GWAS hits**). As noted by Colhoun et al. (2003), poor replication could result from false positives, lack of power, or differences between initial and confirmatory studies (such as using populations segregating different causal alleles; or differences in how a trait, such as a complex disease, is defined). Much of this initial angst, which was especially common in the psychiatrical genetics community (Levinson et al. 2016), has faded as the sample sizes of GWAS increased, and more care was given to ensure that initial and confirmatory studies use consistent methodologies, resulting in a dramatic improvement in repeatability (Visscher et al. 2012, 2017; Marigorta et al. 2018).

Finally, as developed in Chapter 18, one consequence of low power is that the effects of markers declared to be significant are overestimated, with the degree of bias increasing as power decreases (Figure 18.8; Equation 18.43). This is called the Beavis effect in QTL linkage mapping, but in GWAS it is commonly referred to as the **winner's curse** (a term from the epidemiological literature, but whose roots trace back to the analysis of competitive bids of oil leases; Capen et al. 1971). The GWAS implications of the winner's curse, and potential corrections, have been extensively discussed in the literature (e.g., Siegmund 2002; Sun and Bull 2005; Wu et al. 2005; Garner 2007; Yu et al 2007; Zöllner and Pritchard 2007; Ghosh et al. 2008; Zhong and Prentice 2008, 2010; Bowden and Dudbridge 2009; Xiao and Boehnke 2009; Faye et al. 2011; Xu et al. 2011; Ferguson et al. 2013; Poirier et al. 2015; Wang et al. 2016; Palmer and Pe'er 2017; Dudbridge and Newcombe 2019; Wang et al. 2020; Xie et al. 2021).

Example 20.7 As we have seen, a number of classic association metrics in a GWAS are either chi-square tests, or can be reformulated as such. In these cases, the test statistic follows a (central) chi-square when there is no association, and a noncentral chi-square (Appendix 5) when there is a marker effect. The key quality for the later distribution is its **noncentrality parameter** (or **NCP**) which can be thought of as the inflation of the test statistic over its value under the null. Specifically, if a test follows a χ^2_k under the null, then its expected value is its degrees of freedom, k, while if the test is from the alternative, its expected value is $k + \lambda$, where λ is the NCP (Equation A5.14b), a function of the sample design and marker effects.

To see how power calculations are performed with noncentral chi-squares, consider the threshold for significance under a chi-square test (with one degree of freedom) when we take $\alpha = 5 \times 10^{-7}$. We can quickly compute the value of this threshold using the `R` command `qchisq(1-5*10^(-7),1)`, which returns a value of 25.26. Hence, any marker whose associated chi-square statistic is less than this value is declared to be nonsignificant. The corresponding critical values for $\alpha = 10^{-6}$ and 10^{-7}, are, respectively, 23.93 and 28.37. What NCP value is needed to give a high probability that a test of a true effect is declared significant (i.e., high power)? For a test to have power $1 - \beta$ (under $\alpha = 5 \times 10^{-7}$), we need to solve for the value of the NCP λ that satisfies $\Pr(\chi^2_{1,\lambda} \geq 25.26) = 1 - \beta$. Recall that (in `R`) $\Pr(\chi^2_{k,\lambda} \leq X)$ is obtained as `pchisq(X,k,`λ`)`. Suppose that the NCP for a marker under

the GWAS design is $\lambda = 15$. What is the power? This is $\Pr(\chi^2_{1,15} \geq 25.26)$, computed as `1- pchisq(25.26,1,15)`, which returns a value of 0.12. Hence, the power of this test is only 12%, meaning it would not be declared significant 88% of the time. What value of λ is required for 80% power? A grid search yields $\lambda = 34.43$. This can also be obtained by using `R` to plot $\Pr(\chi^2_{1,15} \geq 25.26)$ for different values of λ, e.g., `curve(1- pchisq(25.26,1,x), 10,35)` plots the power for λ over (10, 30). Similarly, for this critical value, the corresponding λ values for 50% and 90% power become 25.26 and 39.79. Table 20.1 gives the required NCP values for other parameter combinations.

A number of authors have developed expression for power, usually by obtaining expressions for the noncentrality parameter (λ) generated by a specific design (sample size) and marker effects (Schork 2002; Xiong et al. 2002; Chapman et al. 2003; Purcell et al. 2003; Edwards et al. 2005; Viscchер and Duffy 2006; Altshuler and Daly 2007; Klein 2007; Yu et al. 2008; Spencer et al. 2009; Yang et al. 2010a; Sham and Purcell 2014; Wang and Xu 2019). As developed by these authors, the basic structure of the NCP is of the form $\lambda = nr^2 g(p, a)$ where n is the sample size, r^2 is the correlation between marker and causal alleles, and $g(p, a)$ is the causal gene effect (a function of, at least, the allelic effect a and frequency p). We first consider power calculations for a continuous trait under a random sample, and then for a binary trait under a case-control design.

For a continuous trait,

$$\lambda = n\left(\frac{r^2_{MQ}\sigma^2_Q}{1 - r^2_{MQ}\sigma^2_Q}\right) = n\left(\frac{\sigma^2_M}{1 - \sigma^2_M}\right) \tag{20.8a}$$

where σ^2_Q is the causal variance and $\sigma^2_M = r^2_{MQ}\sigma^2_Q$ is the corresponding trait variance explained by the marker. We can write the causal variance as $\sigma^2_Q = 2p(1-p)\beta^2$, where β (the slope of a regression of trait value on SNP allele count; Equation 20.1a) is the effect size in phenotypic standard deviations. More generally, if additional cofactors are added to the model, then the standard deviation refers to the *residual* variance, e.g., Wang and Xu (2019). Because we typically expect $r^2_{MQ}\sigma^2_Q \ll 1$,

$$\lambda \simeq nr^2_{MQ}\sigma^2_Q \tag{20.8b}$$

The power for an imputed site follows by replacing r^2_{MQ} by r^2_{impQ}. The required sample size to achieve a desired amount of power follows directly using the critical values in Table 20.1. For example, to have 80% power using a significance level of $\alpha = 5 \times 10^{-7}$ requires a NCP value of 34.43. Solving for n

$$n \geq \frac{34.43}{r^2_{MQ}\sigma^2_Q}$$

Suppose the causal site accounts for 0.1% of the total variance and $r^2 = 0.4$, then

$$n = \frac{34.43}{0.4 \cdot 0.001} = 86,075$$

The fraction of this sample size required for 50% power is 25.26/34.43 (the ratio of the NCPs from Table 20.1), or 75% (63,150).

Now consider a case-control design, which is a form of selective genotyping (Chapter 18). The power of this design is a function of the difference in marker allele frequencies between cases and controls (Klein 2007; Yang et al. 2010a; Evans and Purcell 2012). To translate the effect of an underlying locus into the expected case-control allele frequency difference, suppose the three genotypes have (multiplicative) risks of $f : f\gamma : f\gamma^2$, and let p be the frequency of the risk-increasing allele ($\gamma > 1$), and let K denote the disease incidence (prevalence) in the population. For a design of ηn cases and $(1-\eta)n$ controls,

$$\lambda = nr^2_{MQ}\left(\frac{2p(1-p)(\gamma-1)^2\eta(1-\eta)}{(1-K)^2[1+p(\gamma-1)]}\right) \tag{20.9a}$$

In order to compare the case-control power with that of a continuous trait, Yang et al. (2010a) considered the **threshold-liability model** (Chapter 30). Under this model, there is an underlying liability value y (assumed to be normally distributed with variance one), with the disease expressed when the liability value exceeds some threshold (T), so that $\Pr(y \geq T) = K$. On this scale, the fraction of the liability variance explained by the causal locus is $\sigma^2_Q \simeq 2p(1-p)(\gamma-1)^2/\iota^2$, where ι is the mean value for the liability above the truncation value, and is given by $\varphi(T)/K$, where $\varphi(x)$ is the unit normal function evaluated at x (Chapter 30). The resulting NCP becomes

$$\lambda \simeq n r^2_{MQ} \sigma^2_Q \left(\frac{\iota^2 \eta(1-\eta)}{(1-K)^2} \right) \tag{20.9b}$$

Suppose that a causal locus for a continuous trait (CT) and a causal locus under a case-control (CC) design both have the same variance (σ^2_Q on the observed, and liability, scales, respectively). From Equations 20.8b and 20.9b, the ratio of the two resulting NCP values becomes

$$\frac{\lambda_{CC}}{\lambda_{QT}} = \left(\frac{n_{CC}}{n_{QT}} \right) \left(\frac{\iota^2 \eta(1-\eta)}{(1-K)^2} \right) \tag{20.9c}$$

$$\simeq \frac{\iota^2}{4(1-K)^2} \tag{20.9d}$$

with Equation 20.9d applying for a design with equal number of cases and controls ($\eta = 1/2$) and the same total design size ($n_{CC} = n_{QT}$). Consider the relative power for a continuous trait (say height) with that of schizophrenia. For this disease, $K = 0.01$, so that $T = 2.33$ (as for a unit normal U, $\Pr[U \geq 2.33] = 0.01$) and $\iota = \varphi(2.33)/0.01 = 2.64$, giving $\lambda_{CC}/\lambda_{QT} = 2.64^2/[4 \cdot 0.99^2] = 1.78$. Hence, the same power for the case-control design occurs at 1/1.78 = 56% of the sample size of the continuous trait. For example, the power for a design with 120,000 individuals for the continuous traits is that same as for a design with around 33,000 cases and 33,000 controls. When the disease is rare, the power gain from in the case-control design more than compensates for the reduction of signal when translating from the liability to the observed (binary) scale.

Multitrait GWAS

As was the case for linkage mapping, marker-trait association machinery can be extended from a single trait to a *vector* of traits. Multivariate traits arise naturally in many settings, such as eQTL mapping (Chapter 21), analyzing correlated traits (Chapter 26), and exploring genotype-by-environment (G x E) interactions (Chapter 27). As such, much of our discussion on multiple trait GWAS is spread over these chapters, with just a few overview remarks here.

As might be expected given the complexity of the task, a variety of different strategies have been suggested for implementing a multitrait GWAS. These include fully multivariate **direct approaches**, such as estimating the vector $\boldsymbol{\beta}$ of regression coefficients of a focal SNP on each of the traits (e.g., Wisser et al. 2011; Korte et al. 2012; Zhou and Stephens 2014; Turley et al. 2018; Carlson et al. 2019) or multivariate path analysis (e.g., Grotzinger et al. 2019; Igolkina et al. 2020; Pritikin et al. 2021). For estimation reasons (Chapter 26), direct approaches are generally restricted to just a few traits (five or fewer). With a very large number of traits, **indirect approaches** are often used. These extract some lower-dimensional feature(s) from the trait vector, such as using the first PC of the phenotypic covariance matrix, and then map this composite trait in a univariate GWAS (e.g., Aschard et al. 2014; Zhang et al. 2018; Carlson et al. 2019). Approaches have also been suggested for combining information from univariate analyses, such as incorporating the correlation structure between univariate estimates (e.g., Hunag et al. 2011; van der Sluis et al. 2013). Finally, while often framed in terms of testing each variant (SNP) separately, one can also use a gene-based (or more generally,

set-based) approach as the predictor variables (e.g., Kim et al. 2016). For example, estimating the vector of trait regression coefficient not for a SNP but rather for some composite SNP score over a region/set of interest.

Galesloot et al. (2014), Porter and O'Reilly (2017), and Chung et al. (2019) reviewed a number of multitrait GWAS approaches and found that there is no one method which does best under all settings. However, with correlated traits, all were generally more powerful that performing a series of univariate analyses. A multiple trait GWAS borrows information from correlated traits to improve the estimate of a focal trait, information that is not exploited by a univariate analysis. Even when traits are genetically uncorrelated, they can still be environmentally correlated (Chapter 27)—such as having correlated measurement errors—and incorporating this residual covariance structure improves estimates.

Many of the same general issues about testing and combining information from multiple sources that arose in gene-based GWAS also apply to a multitrait GWAS. For example, in a gene-based setting, one might first test for the significance of the overall gene and then perhaps attempt to detect signals associated with particular variants within such selected genes. At a multitrait level, one might first test whether the $\boldsymbol{\beta}$ vector associated with a SNP (or a more general set, such as a gene or pathway) is significantly different from zero. If so, then one might test particular traits within this vector. Peterson et al. (2016a) proposed a two-step FDR approach in such testing, first controlling the FDR on which variants have a significant nonzero $\boldsymbol{\beta}$ vector and the controlling the FDR for the nonzero elements of $\boldsymbol{\beta}$ within those selected SNPs.

CORRECTING FOR CRYPTIC RELATEDNESS AND POPULATION STRUCTURE

A marker-trait *association* is just that, a *correlation* between the value of a genotype and the value of a trait. While our above discussion has assumed that this correlation arises because a marker locus is in LD with a causal locus (QTL), it can also arise from other factors, such as the presence of undetected relatives in the sample and/or undetected population structure. Given the enormous sensitivity of a modern GWAS to very small influences, these **spurious associations** are potentially very problematic. Chapter 17 discussed the Transmission Disequilibrium Test (TDT), wherein one examines whether a parent heterozygous at a marker passes (transmits), or fails to pass, a specific allele to its offspring. For a random allele, the probability of a transmitted or a nontransmitted event should be equal. However, if we partition individuals by trait value (such as disease presence/absence), this ratio can be skewed for alleles that are linked to causal loci (Example 17.20). By focusing on within-family transmission, population structure is fully controlled when using the TDT, and, given that it specifically uses known relatives, undetected relatives are not a concern. Under a linkage analysis, relatedness is *known, recent, and useful*, while under an association analysis, relatedness is typically *unknown, distant, and a nuisance* (Astle and Balding 2009). Hence, during the early 2000s, the TDT was taken as the gold standard for replication of a proposed candidate gene or association signal. However, this strategy is in conflict with the motivation for transitioning to population-level association analysis, which was to *avoid* the difficulty and resources required to collect a sufficiently large collection of families to have reasonable power. Furthermore, the TDT can be challenging with late-onset diseases, where it may be difficult to sample relatives (parents or sibs) of a case.

As detailed below, in an attempt to move away from family-based studies, a number of approaches have been proposed using marker data to detect relatives and population structure, and to correct for their effects. Examples 9.13 and 9.14 foreshadowed the nature of these corrections, which are typically performed using mixed models (Chapter 10). We will first examine corrections for the presence of relatives before turning to adjustments for population structure. Essentially just one approach has been proposed to adjust for relatives, while multiple approaches have been suggested for population structure. Reviews on a number of the issues discussed below are given by Pritchard et al. (2000b), Pritchard and Donnelly (2001), Hoggart et al. (2003), Voight and Pritchard (2005), Patterson et al. (2006),

Tiwari et al. (2008), Astle and Balding (2009), Choi et al. (2009), Price et al. (2010a), Zhang and Deng (2010), Sillanpää (2011), Sul et al. (2018), and Toosi et al. (2018).

Correcting for Relatives in the Sample

Early GWAS studies typically went out of their way to exclude known relatives (e.g., WTCCC 2007). The reason was simple: relatives share alleles (IBD) and have correlated values for heritable traits (Chapter 7). This induces a correlation between shared alleles and trait values, generating marker-trait associations even at markers unlinked to QTLs. In humans, identification of close relatives upon sampling is often (but not always!) straightforward, while detecting more distant relatives can be more problematic. The situation is much more challenging in natural populations, where pedigree relationships are typically not observed, and hence must be inferred from marker data (Chapter 8). The same concern arises when constructing association panels for mapping QTLs using collections of distinct lines (e.g., Atwell et al. 2010). Some of these lines may have recent ancestors, and hence are relatives, which may be unknown to the investigator. The presence of undetected relatives in a sample is referred to as **cryptic relatedness**.

The remedy for these concerns was introduced in Chapter 19 in the form of mixed models (Chapter 10), with a random effect included for relatedness (Equation 19.29; Example 19.5). For example, the additive model for SNP k (Equation 20.1a) now becomes

$$z_i = \mu + b_k N_{i,k} + A_i + e_i \tag{20.10a}$$

where b_k is a fixed effect for the SNP, while A_i is a random effect for the background (additive) polygenic value (the effect of QTLs not in LD with the focal SNP). From Equations 7.11a and 19.30a, these polygenic values are correlated among relatives,

$$\sigma(A_i, A_j) = 2\Theta_{ij}\,\sigma_A^2 \tag{20.10b}$$

where σ_A^2 is the background additive variance for the trait, and the Θ_{ij} can be estimated from a known pedigree (Chapter 7) or from marker data (Chapter 8). Note that Equation 7.12 generalizes this approach to higher-order (nonadditive) shared polygenic effects.

The mixed model given by Equation 20.10a requires specification of the covariance matrix of the A_i random effects. This is given by the $n \times n$ matrix for all of the pairwise Θ_{ij} values among the n individuals in the population sample. The resulting matrix in GWAS studies is often called the **K** (for **kinship**) **matrix** by plant breeders or the **numerator relationship matrix** (**A** = 2**K**) by animal breeders, with the human literature using both. When estimated solely from marker information, this is often called the **genomic relationship matrix** (**GRM**; also the **realized relationship matrix**, **RRM**). Example 9.14 gives several different marker-based estimates of the GRM. To avoid double-dipping of SNP information (**proximal contamination** and a loss of power; Lippert et al. 2011; Sawcer et al. 2011; Listgarten et al. 2012; Yang et al. 2014), usually the **LOCO** (**leave one chromosome out**) approach is used (Yang et al. 2014). Here, the GRM is computed using marker information for all chromosomes *except* the one on which the tested SNP resides. Hence, the GRM varies slightly over chromosomes. Further, often a subset of the markers is used, trimming out those that are in high LD with each other. Zaitlen et al. (2013) and Jiang et al. (2019) showed that using a **sparse GRM**, truncating small values (e.g., $2\widehat{\Theta}_{ij} < 0.05$ or 0.01) to zero captures essentially as much variation as the full GRM.

Example 20.8 The resulting GWAS mixed model (testing marker k) that accounts for relatives in the sample can be written as

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta}_k + \mathbf{Z}\mathbf{u} + \mathbf{e}, \qquad \begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix} \sim \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_A^2\mathbf{K} & \mathbf{0} \\ \mathbf{0} & \sigma_e^2\,\mathbf{I} \end{pmatrix} \tag{20.10c}$$

where the elements of GRM **K** are marker-based estimates of $2\Theta_{ij}$ (Example 9.14). The covariance matrix for the residuals (**e**) assumes an OLS structure (σ_e^2 **I**), but can be replaced by σ_e^2 **R** to allow for more general GLS error structures (Chapter 10). Here **u** is the random vector of background polygenic effects (the A_i), the incidence matrix **Z** = **I** when all observations are singletons, and (assuming the only fixed effect beyond an additive SNP effect is the mean, μ) the structure of the vector $\mathbf{X}\boldsymbol{\beta}_k$ is

$$\mathbf{X}\boldsymbol{\beta}_k = \begin{pmatrix} 1 & 2 \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & N_{k,n} \end{pmatrix} \begin{pmatrix} \mu \\ b_k \end{pmatrix} = \begin{pmatrix} \mu + 2\,b_k \\ \mu + 0 \\ \mu + b_k \\ \vdots \\ \mu + N_{k,n}\,b_k \end{pmatrix} \tag{20.10d}$$

In this example, individual one (first row of **X**) is a reference homozygote ($N = 2$) at the focal SNP (marker k), two is a nonreference homozygote ($N = 0$), three is a heterozygote, and so on. Note that under fixed-effect modeling of SNP effects, this model is *rerun for each tested SNP*. Because σ_A^2 is the polygenic variance accounted for by QTLs *not* in LD with the tested marker, it ***potentially changes over markers***, and hence ***needs to be reestimated for each marker***. The reader might note that the expected change in σ_A^2 over markers should be small, but recall that a large modern GWAS can be influenced by *very* small effects. For all SNPs on the same chromosome, the GRM is unchanged, but the elements in the second column of **X** (corresponding to the genotypes at the focal SNP) vary with over markers.

If additional fixed effects (cofactors) are included, these add extra columns to **X** with the corresponding fixed effect added to $\boldsymbol{\beta}_k$ (Chapter 10). For example, if the general genotype model (Equation 20.1b) is fit, then

$$\mathbf{X}\boldsymbol{\beta}_k = \begin{pmatrix} 1 & 2 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ \vdots & \vdots & \vdots \\ 1 & N_{k,n} & H_{k,n} \end{pmatrix} \begin{pmatrix} \mu \\ b_k \\ d_k \end{pmatrix} = \begin{pmatrix} \mu + 2\,b_k \\ \mu + 0 \\ \mu + b_k + d_k \\ \vdots \\ \mu + N_{k,n}\,b_k + H_{k,n}\,d_k \end{pmatrix} \tag{20.10e}$$

Correcting for Population Structure: Genomic Control

Another confounding factor is **population structure** (or **population stratification**), wherein one has unknowingly sampled from several different subpopulations (or **strata**). More generally, there can also be individuals in the sample with different amounts of **admixture** (carrying genes from two, or more, of the subpopulations). Such individuals may not fall into discrete strata, but rather could form a potential continuum of admixture (**continuous admixture populations**). The concern is that if disease risks (or trait means) differ over these strata (or over amounts of admixture), and similarly some SNP frequencies also vary over strata, marker-trait associations can be created (Knowler et al. 1988; Spielman et al. 1993; Lander and Schork 1994; Risch 2000; Cardon and Palmer 2003; Freedman et al. 2004; Marchini et al. 2004; Clayton et al. 2005). For example, suppose allele *M* (at a particular SNP, unlinked to any QTL) is common in subpopulation 1, but rare in subpopulation 2. Simultaneously, subpopulation 1 has a higher disease risk (or trait mean) than subpopulation 2. The presence of *M* predicts membership in group 1, which, in turn, predicts a higher disease risk (or larger trait value), generating a marker-trait association. Note that population subdivision may not be obvious to even a careful investigator, leading to the concern of undetected **cryptic stratification** introducing false positives. The original large-scale GWAS (WTCCC 2007) did a careful analysis to ensure that there was little stratification, in both the cases and controls, as well as between them. The latter point is of special concern, in that cases and controls are assumed to be drawn from the same population. Failure to do so

creates stratification and the potential for numerous false positives. One potential generator of stratification at the case versus control level is **ascertainment bias**: for example, cases could easily be oversampled from one strata if its members are overrepresented in clinics (Astle and Balding 2009).

Example 17.19 showed a classic example of stratification with diabetes in members of the Pima nation in southern Arizona. Here the focal marker predicted whether an individual had some amount of admixture with Caucasians, a lower risk population. A more recent example is the work of Campbell et al. (2005), examining height in a collection of European Americans. No stratification signal was seen in a panel of roughly 200 SNPs, but a SNP in the *LCT* gene (associated with the persistence of lactase) showed a strong association with height ($p < 10^{-6}$). This allele showed a cline of increasing frequency from southern to northern Europe, roughly matching a similar trend seen in height. When individuals were corrected for ancestry, the signal was greatly reduced overall, and absent in some comparisons.

Increasingly sophisticated population genetics models have examined the impact of stratification and admixture on generating spurious associations (Devlin and Roeder 1999; Pritchard and Rosenberg 1999; Wacholder et al. 2000; Gorroochurn et al. 2004; Rosenberg and Nordborg 2006). A few generalizations emerge from this work. First, spurious associations usually tend to be less severe in admixed populations relative to the corresponding mixture of discrete subpopulations. Second, the largest impact occurs when lower-risk subpopulations make up a larger fraction of the sample. Finally, the severity of spurious associations usually decreases as the number of underlying subpopulations increases.

The first general approach to adjust for population structure was the method of **genomic control**, **GC** (Devlin and Roeder 1999; Bacanu et al. 2000, 2002; Devlin et al. 2001a, 2001b, 2004), using a single correction over all tests (a closely related method was independently proposed by Reich and Goldstein 2001). Suppose a disease GWAS is conducted using a trend test (Equation 18.15). Under the null (no linkage to a QTL), the resulting test statistic at each marker follows a chi-square with one degree of freedom (χ^2_1). Devlin and Roeder (1999) noted that when population structure is present, all of the marker test statistics should be enhanced by a common **genomic inflation factor**, λ. If S_k denotes the test statistic value for marker k, then under the null in the presence of population structure, $S_k \sim \lambda\chi^2_1$, so that $S_k/\lambda \sim \chi^2_1$, suggesting a correction for structure applicable to all tests. Devlin and Roeder showed that the inflation factor (with case-control data) is given by

$$\lambda = 1 + n\left[\frac{F_{ST}\sum_j (D_j - C_j)^2 - 2F_{ST}}{1 + F_{ST}}\right] \tag{20.11a}$$

where each subpopulation has n sampled cases and n sampled controls, D_j and C_j are the fraction of cases (diseased) and controls, respectively, in the jth population, and F_{ST} is a measure of population structure (the fraction of marker variation due to among-population differences; WL Chapter 2). Observe that λ *increases* with the sample size (n), so that larger samples should show *more* bias. For large n and small F_{ST},

$$\lambda \simeq 1 + n\left[F_{ST}\sum_j (D_j - C_j)^2\right] \tag{20.11b}$$

As first noted by Devlin and Roeder (1999), cryptic relatedness also inflates λ (e.g., Sawcer et al. 2011), and Voight and Pritchard (2005) developed expressions for λ under a variety of relationship scenarios. The latter's conclusion was that deep relationships have a minimum impact of λ, so that a random sample from a large single outbred population should not be impacted. However, sampling bias that favors choosing relatives (such as cases being weakly related) can have a profound impact.

How does one go about estimating λ? Because the expected value of a χ^2_1 is 1 (Equation A5.14b), one might initially consider using the mean value ($\overline{S}$) of test statistics over all

markers as an estimate of λ, e.g., the **mean estimator**

$$\widehat{\lambda}_{mean} = \overline{S} \tag{20.12a}$$

The limitation with this approach is that we expect some small fraction of the markers are indeed linked to QTLs, and would generate large S values. Including these when computing the mean value under the null inflates the estimate of λ. Statisticians often deal with such outliers by using the ranks of the data in place of their actual values (e.g., Conover 1999; Sprent and Smeeton 2007), and Devlin and Roeder (1999) used this idea to propose a **median estimator** for λ. Note that $\Pr(\chi_1^2 \geq 0.4549) = 0.50$, so that the expected median value of test statistics under the null is 0.4549 (half of the tests are expected to be above, and half below, this value). By considering ranks, rather than values, extreme outliers do not unduly influence rank-based statistics. Ordering the values of the test statistics over the m markers from smallest ($S_{[1]}$) to largest ($S_{[m]}$), the **Devlin-Roeder estimator** is given by

$$\widehat{\lambda}_{med} = \frac{\text{median}(S_{[1]}, \cdots, S_{[m]})}{0.4549} \tag{20.12b}$$

where median$(\cdot)$ denotes the value separating the upper half of test scores from the lower half. Other robust estimators for λ were discussed by Wang (2009). Finally, Devlin et al. (2004) noted that both Equations 20.12a and 20.12b ignore variation in the S_k, and suggested their **GCF estimator**, where

$$\frac{S_k}{\overline{S}} \sim F_{1,m} \tag{20.12c}$$

so that the GCF-adjusted test statistic follows a scaled F, rather than a scaled χ^2, distribution (for very large m, $F_{1,m}$ approaches a χ_1^2). Equation 20.12c follows because

$$\overline{S} = \frac{1}{m}\sum_{k=1}^{m} S_k \sim \frac{1}{m}\sum^{m} \chi_1^2 \sim \frac{\chi_m^2}{m} \tag{20.12d}$$

with the last step following because the sum of m χ_1^2 is a χ_m^2 random variable (Equation A5.14c). Hence, from the definition of the F distribution (Appendix 5),

$$\frac{S_k}{\overline{S}} \sim \frac{\chi_1^2/1}{\chi_m^2/m} \sim F_{1,m}$$

Dadd et al. (2009) found GCF outperformed GC based on either λ_{mean} (GC$_{mean}$) or λ_{med} (GC$_{med}$). We have framed the above discussion of GC in terms of testing the additive (or trend) model. Zheng et al. (2005) examined the power of using a trend-test based GC when dominance is present, while Zheng et al. (2006) extended GC to the general genotype model.

While conceptually elegant and simple, one immediate flaw with GC is that *all markers are treated equally*. As a result, the *relative rankings of the markers do not change* under genomic control. All marker tests are adjusted by the same amount, when in fact, a more marker-specific approach is required. In Equation 20.11, F_{ST} is the mean value over *all* markers ($\overline{F}_{ST}$), while markers with excessive divergence relative to the average ($F_{ST} \gg \overline{F}_{ST}$) should receive greater adjustments (Price et al. 2006; Kang et al. 2010). Thus we would expect that a proper correction for structure is likely to change the rankings of some of the markers. As a result, GC is usually no longer used to correct for population structure. However, as with many ideas in science, GC has been recycled, and is now often regarded as a useful *metric of model fit*. The values of λ before and after a model-based correction are compared to assess how well the model reduced the impact from confounding effects. The rough rule of thumb historically used is that a value of $\lambda \leq 1.05$ implies that the model has largely corrected for any structure (e.g., Price et al. 2010a). However, as we now show, one can observe significant inflation in the *absence of any confounders* if a large number of small effect QTL underlie a trait.

Adjusting for Inflation from Polygenicity: LD Score Regression

GC estimates of λ assumes that only a very, very small fraction of tested markers are in LD with QTLs. This assumption fails if a large number of small effect QTLs underpin a trait (the **polygenic model**). Yang et al. (2011a) showed that in this setting, significant inflation can be generated ***even in the absence of any population structure***. In particular, they showed that the impact of polygenes on the genomic inflation factor is

$$\lambda \simeq 1 + n\left(\frac{h^2\,\overline{r^2}\,\overline{s}}{m}\right) \tag{20.13a}$$

Here n is the sample size, h^2 the trait heritability, m the number of tested SNPs, $\overline{s}$ is average number of SNPs in LD with causal variants, and $\overline{r^2}$ is the average r^2 LD between a causal variant and a marker. Notice that, as with inflation from population structure, λ increases with the number of individuals genotyped (n), reflecting the increased power to detect small effects.

Yang et al. (2011a) examined two GWAS studies for human height: the QIMR study of just under 4000 individuals (Yang et al. 2010b) and the GIANT composite GWAS with a combined sample size of around 184,000 (Lango Allen et al. 2010). Substituting GWAS estimates from the QIMR study into Equation 20.13a suggested a λ median estimate of around 1.03, consistent with the observed value. For the much larger GIANT data set, the observed λ value (based on the median estimator) was 1.55, again consistent with the value predicted by Equation 20.13a. Hence, for a highly polygenic trait, a significant fraction of the markers will show elevated values, resulting in a genomic inflation factor much larger than one, even in the absence of population structure. Figure 20.3 underscores this point. Further, ***this effect is expected to be more pronounced as the GWAS size increases***.

Bulik-Sullivan et al. (2015) proposed the method of **LD score regression** to adjust for the inflation due to polygenicity. Their logic was that the impact of population structure should (generally) not scale with the LD reach of a variant, while the impact from polygenicity should (as the marker is impacted by a greater number of causal sites). For a given marker j, they defined its **LD score** as

$$\ell_j = \sum_i r_{ji}^2 \tag{20.13b}$$

where the sum is over scored variants in LD with the focal variant, and r_{ji}^2 the correlation between markers j and i. Under a model where average effect size increases as the minor allele frequency decreases (large-effect alleles tend to be rarer; Chapter 21), they showed that the expected test statistic for a non-causal marker j is

$$\mathrm{E}\left[\chi^2 \,|\, \ell_j\right] = \left[\frac{nh^2}{m}\right]\ell_j + n\alpha + 1 \tag{20.13c}$$

where m is the number of SNPs, h^2 is the trait heritablity, and α denotes the inflation factor from confounders (such as the bracketed term in Equation 20.11b). One then regresses the observed test statistic for marker j on its LD score, ℓ_j. Bulik-Sullivan et al. noted that this is most optimally done as a weighted regression (Example 10.8), and discuss various weighting schemes (see their paper for details). Note that the intercept from the fitted regression (which removes the impact of polygenicity) has expected value $1 + n\alpha$, so that the intercept minus one estimates $n\alpha$, the inflation from confounders. Figure 20.3 shows that most of the inflation seen in a number of large GWAS experiments appears to be due to polygenicity.

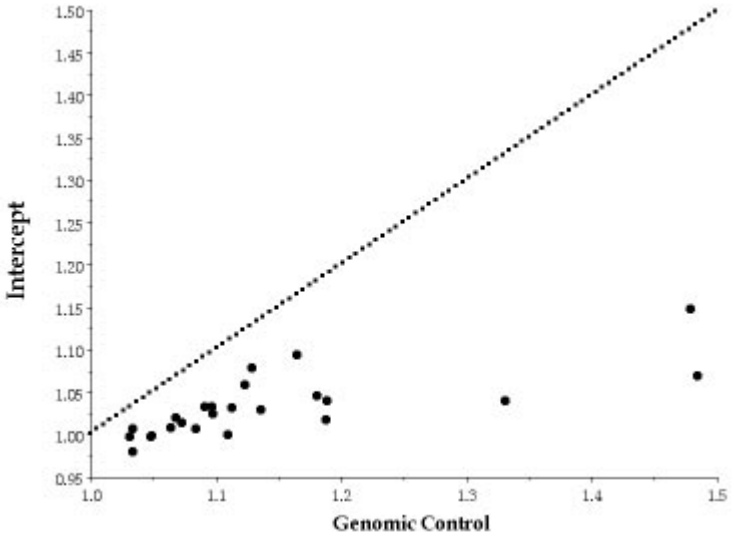

Figure 20.3 Plot of the LD score regression estimates of λ (the intercept of Equation 20.13c minus one) on the vertical axis versus the λ value estimated from GC on the horizontal axis. The data are from 20 different traits examined in large GWAS studies. Note that many of the GC values are rather large (the two largest, $\simeq 1.5$, are for height and schizophrenia), which might suggest major effects from confounders such as relatedness or population structure. However, the LD-score estimates are all much smaller, suggesting that the vast bulk of the inflation over all these traits is from polygenicity. (Data from Bulik-Sullivan et al. 2015).

Correcting for Population Structure: Structured Association Mapping

While GC attempts to *correct* for any population structure, next-generation methods further try to *quantify* the nature of this structure (i.e., how many strata, prediction of strata membership, etc.). These methods are based on the idea that marker data provide information about subpopulation membership. One such signal is that ***differences in allele frequencies between subpopulations generates LD between unlinked markers*** in the combined population sample. To see this, suppose all population 1 individuals are *AABB*, while all population 2 are *aabb*. When the populations are combined into a single sample, assuming equal contributions from both populations, $p_A = p_B = 0.5$, while the LD is

$$D_{AB} = \text{freq}(AB) - \text{freq}(A)\text{freq}(B) = 0.5 - 0.5^2 = 0.25$$

We present this extreme case (differential fixation) to show the basis for this phenomena. In most settings, very small amounts of LD are generated between pairs of unlinked loci that differ in frequency over subpopulations. Such subtle signals require a reasonable number of markers to detect.

This phenomena was exploited by Pritchard and Rosenberg (1999), Pritchard et al. (2000a, 2000b), and Falush et al. (2003, 2007), who assumed that a population sample of interest consisted of k subpopulations, each in Hardy-Weinberg proportions. They then used an MCMC sampler to develop a Bayesian classifier to assign individuals into clusters. Assignment was done in such a way as to minimize the LD between unlinked loci *within* each cluster. Their program, **STRUCTURE**, returns a vector of posterior probabilities of group membership for each sample member. For example, with $k = 3$, this vector might be (0.3, 0.1, 0.6), so that this individual is most likely from group three and least likely from group two. More generally, this could also be interpreted that this individual shows admixture, with 30% of its genes from population 1, 20% from 2, and 60% from three, although Zhu et al. (2002) and Zhang et al. (2003) suggest some caution with this interpretation. Approaches for estimating the number of subpopulations (S) were offered by Pritchard and Rosenberg

(1999), Zhu et al. (2002), Evanno et al. (2005), Price et al. (2006), Camus-Kulandaivelu et al. (2007), and Lawson and Falush (2012). The last authors noticed that the presence of highly related individuals in the sample can lead to unreliable group assignment. While SNPs are often the marker of choice, Thornsberry et al. (2001) showed that 141 highly polymorphic SSR markers (Chapter 8) were sufficient to allow STRUCTURE to account for population structure in a set of 92 maize lines.

Pritchard et al. (2000b) proposed using these probabilities of group membership to perform **structured association mapping** (**SAM**), essentially testing for association within each subpopulation. Using such latent (underlying, but unmeasured) variables to control for population structure was also suggested by Ripatti et al. (2001), Satten et al. (2001), and Sillanpää et al. (2001). Consider a quantitative (as opposed to a binary) trait. Let the vector $\mathbf{q}_i^T = (q_{i,1}, q_{i,1}, \cdots, q_{i,S-1})$ denote membership probabilities for individual i (because the q_{ij} sum to one, $q_{iS} = 1 - \sum_j^{S-1} q_{ij}$), and let $\mu + v_i$ denote the mean value of the trait from group i. In this setting, the adjusted group-weighted mean for individual i becomes

$$\mu_i = \mu + \sum_{j=1}^{S-1} q_{i,j}\, v_i \tag{20.14a}$$

Note under this coding the μ is the mean for group S. One could equivalently code this so that the v_j are deviations from the overall population mean (μ). The population-structure corrected additive model (Equation 20.1a) becomes

$$\begin{aligned} z_i &= \mu_i + b_k N_{i,k} + e_i \\ &= \left(\mu + \sum_{j=1}^{S-1} q_{i,j}\, v_i \right) + b_k N_{i,k} + e_i \end{aligned} \tag{20.14b}$$

The term in parentheses adjusts for mean trait values varying over groups, removing the effects of structure (to the extend that $\mathbf{q}_i$ captures this). In linear model form, for n individuals and S groups, we can add the vector of population structure corrections as

$$\mathbf{Qv} = \begin{pmatrix} \mathbf{q}_1^T \\ \mathbf{q}_2^T \\ \vdots \\ \mathbf{q}_n^T \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_{S-1} \end{pmatrix} = \begin{pmatrix} q_{1,1} & q_{1,2} & q_{1,3} & \cdots & q_{1,S-1} \\ q_{2,1} & q_{2,2} & q_{2,3} & \cdots & q_{2,S-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ q_{n,1} & q_{n,2} & q_{n,3} & \cdots & q_{n,S-1} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_{S-1} \end{pmatrix} \tag{20.14c}$$

where $\mathbf{Q}$ is a known matrix, and the vector $\mathbf{v}$ is estimated (as a fixed effect) from the data. The resulting linear model becomes

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Qv} + \mathbf{e} \tag{20.14d}$$

Example 20.8 gives the structure of $\mathbf{X}\boldsymbol{\beta}$ under both the additive and general genotype models.

Correcting for Population Structure: Principal Components

While the STRUCTURE approach is very elegant, it is computationally demanding and its performance in continuous admixture settings is somewhat unclear. An alternative strategy was proposed by Zhu et al. (2002), Zhang et al. (2003), and Price et al. (2006), using the **principal components** (**PCs**) of the covariance matrix $\mathbf{C}$ of marker information (a **PCA**, or **principal component analysis**, also called an **eigenanalysis**). PCs and eigenvectors were discussed in Chapter 9, and the reader might find it useful to review that material before proceeding. Recall that the **first principal component** (**PC1**) is the (unit-length) vector corresponding to the direction of the most variation in the distribution space described by $\mathbf{C}$. This is also the **leading eigenvector** ($\mathbf{e}_1$), namely that associated with the largest eigenvalue

of **C**. The idea is that instead of running a computationally demanding Bayesian classifier, one could take a small number of PCs and use these to capture structure in the marker data, presumably corresponding to distinct subpopulations and/or axes of admixture within the sample. Price et al. (2006) called this approach **EIGENSTRAT** (based on their program name).

The use of principal components to assign individuals into subpopulations traces back to Menozzi et al. (1978). There is sufficient theoretical work to justify this assumption. For example, Patterson et al. (2006) noted that "*natural models of population structure predict that most of the eigenvalues ... will be* '*small,*' *nearly equal, and arise from sampling noise, while just a few eigenvalues will be* '*large,*' *reflecting past demographic events*." Bryc et al. (2013) showed (for a sufficient number of individuals and markers) that there is a high probability of S large eigenvalues for a population consisting of S well-differentiated subpopulations. Novembre and Stephens (2008) showed that PCs can also pick up signals of spatial differentiation in an otherwise continuous population. One note of caution: while PCA of population structure is generally powerful, is not fool proof. Kimmel et al. (2007) and Zhao et al. (2007) noted that there are situations where PCs do not appropriately correct for population stratification. For example, if cases and controls come from different populations, then a PCA will detect this stratification, but the resulting correction will also remove the QTL association signals. A second complication is that common alleles can show different stratification signature relative to rare alleles in the same sample (Mathieson and McVean 2012; Zaidi and Mathieson 2020), and PC approaches typically use common alleles. We will return shortly to this point.

Patterson et al. (2006) developed a rather remarkable result, based in theoretical work on the eigenstructure of covariance matrices (Baik, Ben Arous, and Péché 2005). A covariance matrix with just a few expected large eigenvalues shows a **phase-change transition**, where if the amount of population signal is slightly above a threshold value there is a strong eigenvalue signal, while if the amount of signal is slightly below this threshold, no clear eigenvalue signal is present. Patterson called this value the **BBP threshold** (after the authors of the motivating theoretical paper), and noted that with m markers and n individuals, this threshold is

$$\tau_{BBP} = \frac{1}{\sqrt{nm}} \tag{20.15}$$

An F_{ST} value above τ_{BBP} generates a clear eigenvalue signal, while populations with F_{ST} below this level have little signal. For 100,000 markers scored in 1000 individuals, $\tau_{BBP} = 0.0001$, so that population structure such that $F_{ST} > 0.0001$ should have a strong PC signal.

The PC correction for stratification proceeds as follows. Consider a **marker matrix M** for the scored genotypes at m markers over n individuals, where

$$\mathbf{M}_{n\times m} = \begin{pmatrix} g_{1,1} & g_{1,2} & \cdots & g_{1,m} \\ g_{2,1} & g_{2,2} & \cdots & g_{2,m} \\ \vdots & \vdots & \cdots & \vdots \\ g_{n,1} & g_{n,2} & \cdots & g_{n,m} \end{pmatrix} \tag{20.16a}$$

Here $g_{i,j}$ is the number of copies of the reference (typically the minor) allele at marker j in individual i (taking on values of 0, 1, or 2). Hence, the ith row are the m marker scores for individual i, while the jth column are the values for marker j over the n individuals in the sample. **M** often represents a trimmed set of the SNPs used in the GWAS, but this need not be the case. Zou et al. (2010) showed that PCs can be influenced by sets of linked SNPs in high LD, falsely suggesting subpopulations. One approach to adjust for this is **thinning**, wherein SNPs are pruned such that adjacent markers in high LD are excluded (e.g., Fellay et al. 2007). Zou et al. noted that thinning discards potentially useful information, and instead suggested a weighting scheme to adjust individual SNPs based on their LD scores (Equation 20.13b). Note that the concern here is with *high levels* of LD between *tightly-linked* SNPs. The *low levels* of LD expected between *unlinked* SNPs due to the presence of structure

are not impacted by these corrections. Finally, the LOCO approach is used, so that each chromosome has a specific **M** matrix, based on all of the used markers *except* for those on the focal chromosome.

Typically, two adjustments are made to the elements of **M** before a PCA. First, the elements are **centered**

$$g_{c,i,k} = g_{i,k} - 2p_k, \qquad \text{where} \qquad p_k = \frac{1}{2n}\sum_{j=1}^{n} g_{j,k} \tag{20.16b}$$

so that the column average is zero. Second, the centered elements are then **normalized** (so that all have the same variance). Note that $g_{i,k}$ is just a binomal random variable (Equation 2.19a; with $n = 2$ and $p = p_k$), and hence (Equation 2.19c) has a variance of $2p_k(1 - p_k)$, leading to

$$g_{n,i,k} = \frac{g_{c,i,k}}{\sqrt{2p_k(1-p_k)}} \tag{20.16c}$$

Denote the centered, normalized marker matrix by $\mathbf{M}^*$. A PCA analysis then proceeds on the covariance between individuals in marker scores. The resulting $n \times n$ covariance matrix is given by

$$\mathbf{C}_{n\times n} = \frac{1}{m}\mathbf{M}^*_{n\times m}(\mathbf{M}^*)^T_{m\times n} \tag{20.16d}$$

where the ijth element of **C** is the inner product of the ith and jth rows of $\mathbf{M}^*$,

$$C_{ij} = \frac{1}{m}\sum_{k=1}^{m} M^*_{ik}M^*_{jk} = \frac{1}{m}\sum_{k=1}^{m} \frac{(g_{i,k} - 2p_k)(g_{j,k} - 2p_k)}{2p_k(1-p_k)} \tag{20.16e}$$

Note from Equation 8.15b is that this is just the correlation estimate of $2\Theta_{ij}$, which assigns more weight to loci with rare alleles. Hence, if the same set of markers is used to correct for close relatives and to correct for structure, **C** is just the GRM used for kinship.

Let $\mathbf{e}_j$ (length $n \times 1$) denote the jth eigenvector of **C**, namely that associated with the jth largest eigenvalue. The **projected ancestry**, a_{ij}, of individual i along the jth axis of variation in **C** is given by the ith entry in $\mathbf{e}_j$. The correction for differing means over the subgroups follows using the a_{ij}, with the adjusted mean value for individual i given by

$$\mu_i = \mu + \sum_{j=1}^{S} a_{ij}\, v_j \tag{20.17a}$$

where the v_j are fixed effects, estimated by the model. The full model correcting for population structure is given by Equation 20.14b, with Equation 20.14c now using

$$\mathbf{Qv} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1S} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2S} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nS} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_S \end{pmatrix} \tag{20.17b}$$

The number ($S < n$) of nontrivial PCs to include in **Q** is a classic PCA problem, with a number of proposed solutions (reviewed by Peres-Neto et al. 2005). One approach is a visual examination of the eigenvalue scree plot (Figure 9.1), given the expectation of a few large, and many small, eigenvalues. However, stopping based on some observed inflection point in the scree plot is ad-hoc and often chooses too many PCs. More formal tests have been proposed. For example, Patterson et al. (2006) used the **Tracy-Widom distribution of the largest expected eigenvalue** of a covariance matrix (Tracy and Widom 1994) to estimate the number of significant PCs. Simulations by Shriner (2011, 2012) showed that this approach also systematically overestimated the number of PCs, and that the bias is larger in admixed

populations (Patterson et al. also noted overestimate problems under admixture). Shriner suggested that part of this bias may occur because stratified and admixed populations display departures from Hardy-Weinberg (even if the underlying subpopulations are in HW). Shriner found that an alternative stopping rule, **Velicer's minimum average partial test** (Velicer 1976), correctly selected the number of PCs to include. Velicer's criteria stops including PCs when the average of squared corrections (following removal of the included PC effects) starts to increase. Alternatively, Tucker et al. (2014) suggested that as a general rule incorporating five PCs was generally sufficient to correct for stratification in many (real and simulated) settings. A caveat with any stopping rule is that the aim under GWAS is not to include *all* nontrivial PCs, but rather just those that *manage to capture the impact of subpopulation structure and/or admixture*.

Two final comments on the PCA approach. Lee et al. (2011) derived the relationship between PC corrections and genomic control, noting that the latter could be expressed as the sum of squared correlations between a PC and the trait value, each term weighted by the eigenvalue for the corresponding PC. Li and Yu (2008) extended the PC approach by jointly clustering individuals and then also correcting by PC score.

Using Admixed Populations Can Improve the Power of a GWAS

Finally, while structured populations are typically treated a complication that needs to be corrected, starting with Pasaniuc et al. (2011), it was realized that using admixed populations can often *improve* the power of a GWAS relative to a similarly sized unstructured population sample (Zhang and Stam 2014; Atkinson et at. 2021; Lin et al. 2021; Hou et al. 2021). The notion is that differential selection pressures and drift can result in differences in causal allele frequencies between the underlying subpopulations contributing (either directly or via admixture) to the mapping population sample. As a result, causal variants may be at more intermediate frequencies in an admixed population, and hence offer greater power. One measure of such power would be the F_{ST} values of the causal alleles of the focal trait. Lin et al. (2021) estimated that causal-site based F_{ST} values in excess of 0.5 were commonly found for traits in humans. With a strong reference sample (such as can be obtained for some human populations), one can use the haplotypes around a particular SNP allele to infer its ancestry, a proceed called **local ancestry inference** (**LAI**). The **Tractor** approach of Atkinson et al (2021) leverages such LAI estimates by modifying the basic gene-dosage regression (Equation 20.1a) to

$$z_i = \mu + \sum_{j=1}^{n_p} b_{k,j} N_{i,k,j} + e_i$$

where $N_{i,k,j} (= 0, 1,$or 2) is the number of reference alleles (from SNP k) in individual i that originated from ancestral population j. Hence, the modification is to partition the reference allele (at the focal SNP) further into the ancestral population that contributed it to individual i. The gain in power using this approach occurs when the effect size tagged by a marker effect varies substantially over the ancestral populations.

MIXED-MODEL GWAS ANALYSIS

The QK Model

Mixed linear models (Chapter 10) provide an ideal framework for the joint incorporation of corrections for recent (kinship) and deeper (population stratification) ancestry. While the above discussion showed the early roots of mixed-model based GWAS analysis (especially when correcting for cryptic ancestry), it was Yu et al. (2006) who jointly incorporated corrections for *both* kinship and stratification. The result was their **QK model**. Their motivation was in the analysis of complex association panels, such as those used by plant breeders. These panels are typically constructed from a very diverse set of inbred lines, leading to a collection of both related lines and sets of lines from very different subpopulations (often

representing a global sample of diversity). By their nature, such panels have both recent and deep shared ancestry.

The QK model proceeds by first partitioning fixed effects into three distinct sets of factors ($\boldsymbol{\beta}, \mathbf{v}, b_k,$). Two of these sets are common over all markers: the vector of trait covariates $\boldsymbol{\beta}$ (such as sex, age, or environmental effects) and the vector of population structure effects $\mathbf{v}$, while b_k is the specific SNP effect for the marker being tested. This is under an additive model, while under the general genotype model b_k is replaced by the vector (b_k, d_k) of additive and dominance effects (Equation 20.1b). Random effects enter as the vector $\mathbf{u}$ of polygenic values and the vector of residuals $(\mathbf{e})$. The resulting model for testing marker k becomes

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{N}_k b_k + \mathbf{Q}\mathbf{v} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{20.18a}$$

The vector $\mathbf{N}_k$ has $N_{i,k}$ as its ith component (the number of copies of the reference allele in individual i), while the vector $\mathbf{Q}\mathbf{v}$ adjusts for population structure (either by STRUCTURE or a PCA; Equations 20.14c or 20.17b, respectively).

Hence, we can write the vector of adjusted mean values as

$$\boldsymbol{\mu}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{Q}\mathbf{v} \tag{20.18b}$$

Combining the two random effects as $\mathbf{e}^* = \mathbf{Z}\mathbf{u} + \mathbf{e}$, Equation 20.18a can be written as

$$\mathbf{z} = \boldsymbol{\mu}^* + \mathbf{N}_k b_k + \mathbf{e}^* \tag{20.18c}$$

Assuming OLS for $\mathbf{e}$ (Equation 10.10), the variance-covariance matrix for $\mathbf{e}^*$ is

$$\mathbf{V} = \sigma_A^2 \mathbf{Z}\mathbf{K}\mathbf{Z}^T + \sigma_e^2 \mathbf{I} = \sigma_A^2 (\mathbf{Z}\mathbf{K}\mathbf{Z}^T + \delta^2 \mathbf{I}), \quad \text{where } \delta^2 = \frac{\sigma_e^2}{\sigma_A^2} \tag{20.18d}$$

Here, $\mathbf{K}$ is the genomic relationship (GRM), or kinship, matrix (whose elements are the marker-based estimates of $2\Theta_{ij}$), and these joint corrections for structure ($\mathbf{Q}$) and kinship ($\mathbf{K}$) lead to the term *QK* model. Some authors (e.g., Sun et al. 2010) use P when PCs are used and Q when STRUCTURE is used (e.g., PK vs. QK models). Adjusting the data to have a mean of zero (in the absence of a SNP effect) leads to

$$\mathbf{z}^* = \mathbf{z} - \boldsymbol{\mu}^* = \mathbf{N}_k b_k + \mathbf{e}^* \tag{20.18e}$$

With estimates of $\boldsymbol{\mu}^*, \sigma_A^2$, and σ_e^2, this is just a GLS regression (Equation 10.13a), with $\mathbf{V}$ given by Equation 20.18d. One can think of the adjustment for population structure as a fixed effect correction (a change in the mean), while cryptic relatedness enters not through the mean, but rather through correlations among the residuals ($\mathbf{e}^*$ not being of OLS form).

Example 20.9 Because both $\mathbf{Q}$ and $\mathbf{K}$ control for relatedness (distant and more recent, respectively), obvious questions for a QK model are whether both corrections are needed, and if only one correction is applied, which one is better. The answer depends on the (typically unknown) relationship structure of the sample. Yu et al. (2006) considered three traits (flowering time, ear height, ear diameter) in a diverse panel of 277 maize inbred lines. They compared the Q model (Equation 20.18a with no $\mathbf{u}$ term; a structured population analysis), the K model (Equation 20.18a with no $\mathbf{Q}$ term (a single-population model with polygenic control for shared recent ancestry), and the Q+K model (the full version of 20.18a). Over the three traits, all models showed better performance relative to the simple model where both $\mathbf{u}$ and $\mathbf{Q}$ were absent. Among the three corrections for confounding, the performance rankings were Q lowest, K better, and Q+K just slightly better than K alone. For example, for flowering time, roughly 38% of the examined 560 SNPs were significant ($p < 0.05$) under the simple model, while this fraction was reduced (presumably by reducing the fraction of false positives) to 14% for the Q model, 6.1% for the K model, and 6.0% for the Q+K model.

Yu et al. also examined power under different assumed effect sizes. The choice of trait impacted power calculations because the impact of both structure and close relatedness depend on the genetic structure of the particular trait. Both the Q and K models had greater power than the simple model. For ear height and flowering time, the Q+K model had the greatest power, while for ear diameter, K had slightly higher power than Q+K.

Zhao et al. (2007) also examined whether the K model alone was sufficient, or if **Q** must also be included in an association panel of 95 *Arabidopsis* accessions. As did Yu et al, they found that while the K model worked well, the Q+K model gave better results. However, when they computed kinship using percentage of shared haplotypes (which they called K^*), the K^* model alone was essentially as good as the K^*+Q model. Conversely, Bradbury et al. (2011) found that the K-only model was superior to other models over a wide range of conditions given the population structure of their collection of Barley lines.

Kinship and Structure: Recent Versus Deep Ancestry

The thoughtful reader might have wondered about our apparently sharp distinction between kinship and population structure (K versus Q), as both are metrics of shared ancestry, the former of recent ancestry, the latter of deeper ancestry. Further, the two corrections often use the same marker information, as the GRM for kinship correction is often used by the PCA to extract axes of population structure. However, this not need be the case, as the GRM for kinship correction and the marker matrix for the PCA correction for structure can contain different sets of markers. Indeed, there are suggestions in the literature that while the structure correction should use a very large set of markers (to capture small signals; e.g., Equation 20.15), kinship may be efficiently captured using a much smaller number of markers (Lippert et al. 2011; Listgarten et al. 2012, 2013; Tucker et al. 2014; Liu et al. 2016; Jiang et al. 2019).

Despite the fact that different set of markers can, in principal, be used for kinship and structure corrections, the issue remains as to why the Q+K model often outperforms either the K or the Q models (Example 20.9). The key is that the dimensionality of deep ancestry (i.e., population structure) is different from that for kinship. When structure is present, a random individual shares deep ancestry with a considerable fraction of the population, and this signal is often largely captured by the first few PCs of the marker covariance matrix. In contrast, a random individual (the absence of large extended pedigrees) likely shares recent kinship with only a very, very small fraction of the sampled population. Hence, while recent kinship itself might not be uncommon, it is caused by a large number of very small groups of close relatives. The small signal for each such cluster of close relatives is manifested by very low rank eigenvalues from the GRM. Hence, when the GWAS sample is a mixture of both deep and recent ancestry (such as often occurs with plant association mapping panels), the resulting eigenstructure has a few high-ranking eigenvalues from deep ancestry and a long tail of small eigenvalues from shared recent kinship. While PCs are a poor way to exploit information from these small clusters of recent relatedness, these correlations can be efficiently utilized by the GRM.

Another was to think about independent contributions from Q and K, even when both use the same GRM, is the REML variance estimation framework (Chapter 32). Under a REML analysis, one first removes the impacts from any fixed effects. The REML variance estimate is thus the residual variation that remains after the fixed effects are removed. Because PCAs are entered as fixed effects, the REML estimate for σ^2_A that forms the basis for the kinship correction extracts the remaining information not captured by the PCs.

Computational Improvements

There has been a considerable amount of theoretical work build around the basic QK structure for a GWAS (Table 20.2), most of it related to making this mixed model computationally

Table 20.2 Optimization and approximation methods for improving the efficiency of a mixed-model GWAS. Optimization methods attempt to find numerically more efficient methods for solving the exact mixed-model equations. Most approximation approaches start by assuming a constant value for polygeneic variance, rather than the exact model which allows this to vary over markers.

Continuous traits: Optimization of exact solutions		
EMMA	**Efficient mixed-model association**	Kang et al. (2008)
FaST-LMM	**Factored spectrally transformed LMM**	Lippert et al. (2011)
GEMMA	**Genome-wide efficient mixed-model association**	Zhou and Stephens (2012)
Continuous traits: Approximation of exact solutions		
EMMAX	**EMMA expedited**	Kang et al. (2010)
P3D	**Population parameters previously determined**	Zhang et al. (2010)
GRAMMAR	**Genomewide rapid association using mixed model and regression**	Aulchenko et al. (2007)
GRAMMAR-GC		Amin et al. (2007)
GRAMMAR-Gamma		Svishcheva et al. (2012)
BOLT-LMM		Loh et al. (2015a)
fastGWA		Jiang et al. (2019)
Binary traits: Linear model approximations		
LTMLM	**Liability-threshold mixed linear model**	Hayeck et al. (2015)
LEAP	**Liability estimator as a phenotype**	Weissbrod et al.(2015)
Binary traits: Logistic regression optimization/approximation		
GMMAT	**Generalized linear model association test**	Chen et al. (2016)
SAIGE	**Scalable and accurate Implementation of generalized mixed model**	Zhou et al. (2018)

more efficient so that it can scale to the size of a modern GWAS ($n > 10^4, m > 10^6$). Using standard methods for solving mixed models (Chapter 32), the run times scale as at least $O(mn^3)$. In the QK model, the vectors $\boldsymbol{\beta}$ and $\mathbf{v}$ of fixed effects, and the polygenic variance σ_A^2 (and its associated vector $\mathbf{u}$ of random polygenic effects) are all **nuisance parameters**, variables that must be included in the model, but that are generally not of much interest by themselves (although we will return to the polygenic variance in Chapter 32). In a GWAS, the sole interest is usually whether b_k is significantly different from zero.

The computationally demanding step in applying the full QK model to each of the m markers is the estimation of the background polygenic additive variance, σ_A^2. In an exact analysis, σ_A^2 must be reestimated for each different SNP as, in theory, it changes slightly depending in the tested SNP. Indeed, we could have written this as $\sigma_{A(-k)}^2$ to remind the reader of this fact, but choose not to for ease of presentation. Differences in σ_A^2 among markers are largest when comparing a marker with a large effect and one with a small effect. The latter has the large-effect site incorporated into σ_A^2, while the former has it removed. Hence the common assumption of an equal value of the polygenic variance σ_A^2 over all markers (Table 20.2) is not really problematic for markers of very small effects (their removal results in very small marker-specific changes in σ_A^2). With large-effect markers, however, this constant-variance assumption can introduce bias. One simple solution is to incorporate a few of the largest-effect SNPs as cofactors (Chapter 18), which should result in a more consistent value of σ_A^2 over markers.

Running a full mixed model analysis on the millions of markers (directly scored or imputed) in a modern GWAS is thus computational very demanding, as standard methods for the estimation of σ_A^2 typically scale as (at least) $O(n^3)$ (Chapter 32). Thus, a GWAS with tens of thousands of individuals and millions of markers is exceptionally challenging if an exact mixed-model analysis is performed on each marker. With σ_A^2 and σ_e^2 estimated, inversion of the $n \times n$ matrix $\mathbf{V}$ (Equation 20.18d)—itself not a trivial operation for a large

number of individuals—can be used to obtain exact estimates of the fixed effects (Equation 10.13a). However, as we detail in Chapter 32, estimation of variance is an *iterative procedure*, with each step usually involving the inversion of $\mathbf{V}_{[i]} = \sigma^2_{A[i]}(\mathbf{K} + \delta^2_{[i]}\mathbf{I})$, where i indexes the iteration. Because the updated variance components change the value of $\mathbf{V}_{[i]}$, a new inversion is required for each iterative step. Further, in the LOCO setting, the kinship matrix $\mathbf{K}$ is slightly different for each of the tested chromosomes. Given this background, much of the work on mixed-model GWAS analysis starts with some version of Equation 20.18a, and then explores either computational improvements or approximations to make such an analysis feasible for the scale of a modern GWAS. Summaries of the time complexity scaling for the different methods presented below can be found in Svischeva et al. (2012), Zhou and Stephens (2012), Zhou et al. (2018), and Jiang et al. (2019).

The first approach has been to develop computational improvements of the mixed-model estimation algorithms (typically by improving the estimation of σ^2_A). One of the first suggestions was the **efficient mixed-model association** (**EMMA**) method of Kang et al. (2008). They showed that using the eigenvalue decomposition (Equation A3.32a) of the GRM $\mathbf{K}$ allowed one to avoid having to invert $\mathbf{V}_{[i]}$ during each iteration of the variance estimation procedure. Computing the initial decomposition has complexity of order n^3, with subsequent iteration steps having complexity of order rn (for r iterations). Traditional iterative variance estimation schemes typically have a complexity of order rn^3 (Chapter 32). Under the LOCO setting with markers tested over ℓ chromosomes, there are ℓ separate $\mathbf{K}$ estimates, so that scaling becomes $O(\ell[n^3 + rn])$.

While EMMA reduces the computational complexity for each SNP, a new decomposition most be performed for each marker. Lippert et al. (2011) developed an improvement for setting where the number m_r of SNPs used to estimate the GRM $\mathbf{K}$ is less than n. Their **factored spectrally transformed LMM** (**FaST-LMM**) method requires just a single eigenvalue decomposition that can be used for all tested markers on a given chromosome. Zhou and Stephens (2012) introduced their **genome-wide efficient mixed-model association** (**GEMMA**) which, while also using an eigendecomposition of $\mathbf{K}$, maximizes the likelihood function in a more efficient manner than previous algorithms. As a benchmark, Zhou and Stephens noted that the computation times for EMMA, FaST-LMM, and GEMMA for two different human data sets were, respectively, $\sim$ 9 days, 6.8 hours, and 33 minutes for one set and $\sim$ 27 years, 6.2 hours, and 3.3 hours for the other.

This first generation of accelerated methods scaled between $O(mn^2)$ and $O(m^2n)$ (Loh et al. 2015a). While this is an improvement over the $O(mn^3)$ scaling of classic mixed model analysis, even with these gains, the number of computational steps for a full exact mixed model GWAS for a modern data set is *at least* 10^{14} to 10^{16}. Second generation methods, **BOLT-LMM** (Loh et al. 2015a) and **fastGWA** (Jiang et al. 2019), scale even faster, between $O(mn)$ and $O(mn^{3/2})$. Both methods estimate σ^2_A once (see below) and use clever computational tricks, such as sparse matrix inversion (Jiang et al. 2019).

In addition to more efficient methods for estimating the background polygenic variance associated with each marker, a complementary approach is to approximate the exact model. Several investigators suggested estimating σ^2_A and σ^2_e once under the null model of no SNP effect (Equation 20.18a with no $\mathbf{N}$ term), with these estimates subsequently used when testing each marker. This reduces Equation 20.18a to a GLS regression (Equation 10.13a), with $\mathbf{V} = \sigma^2_A\mathbf{K} + \sigma^2_e\mathbf{I}$. This constant polygenic variance approximation was first proposed (for family-based association mapping) by Chen and Abecasis (2007). Kang et al. (2010), using their EMMA variance estimator, called this approach **EMMA eXpedited** (**EMMAX**), while Zhang et al. (2010) called this **population parameters previously determined** (**P3D**). Zhang et al. also suggested that **compression** can be used, wherein individuals with very similar relationship values are clustered as a group, and their mean relatedness used. This has the effect of reducing the size of the kinship matrix, which can speed up calculations. Simulation results by these authors showed these approximations generally work well with low heritability traits, but become less precise as heritability increases. They also tended to be less powerful than the exact method. As mentioned above, incorporation of a few of the

largest-effect SNPs as cofactors should result in a more consistent value of σ_A^2 over markers.

A related approach was motivated by Equation 20.18e, with the idea that one could correct each individual once and then subsequently use the adjusted values in the marker regressions. Aulchenko et al. (2007) called this **genomewide rapid association using mixed model and regression** (**GRAMMAR**). Once again, Equation 20.18a is fit under the null model of no SNP effect (the $\mathbf{N}$ term is not included). With estimates of the vectors of fixed effect (BLUEs for $\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{v}}$) and predictions of the vector of random polygenic effects (BLUPs for $\widehat{\mathbf{u}}$) in hand, the trait value for each individual can be adjusted to remove these effects,

$$\mathbf{z}^* = \mathbf{z} - \widehat{\mathbf{z}}, \quad \text{where} \quad \widehat{\mathbf{z}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Q}\widehat{\mathbf{v}} + \mathbf{Z}\widehat{\mathbf{u}} \tag{20.19a}$$

The regression for any given SNP is then simply tested using

$$\mathbf{z}^* = \mathbf{N}_k b_k + \mathbf{e} \tag{20.19b}$$

This is just a GLS regression, and estimates for the marker effect, its sampling variance, and its significance are obtained as follows. Let $\mathbf{x}_k$ be the vector of mean-centered counts ($N_{i,k}$) for marker k, whose ith entry is $x_{i,k} = N_{i,k} - 2p_k$. With estimates of σ_A^2 and σ_e^2 in hand, the GLS estimates for the effect from SNP k, its variance, and its test statistic (distributed as χ_1^2), become

$$\widehat{b}_k = \frac{\mathbf{x}_k\mathbf{V}^{-1}\mathbf{z}^*}{\mathbf{x}_k\mathbf{V}^{-1}\mathbf{x}_k}, \qquad \sigma^2\left(\widehat{b}_k\right) = \frac{1}{\mathbf{x}_k\mathbf{V}^{-1}\mathbf{x}_k}, \qquad \chi^2_{[k]} = \frac{(\mathbf{x}_k\mathbf{V}^{-1}\mathbf{z}^*)^2}{\mathbf{x}_k\mathbf{V}^{-1}\mathbf{x}_k} \tag{20.19c}$$

Note the $\mathbf{V}$ is assumed to be the same for all markers (the polygenic variance σ_A^2 is constant over markers).

Aulchenko et al. originally corrected just for relatedness, but their approach easily extends to the full QK model. GRAMMAR turns out to give slightly biased estimates of SNP effects and p values that are conservative (the true p values are slightly less than the reported p values). To adjust for these effects, subsequent modifications were offered by Amin et al. (2007) (**GRAMMAR-GC**), and Svishcheva et al. (2012) (**GRAMMAR-Gamma**), where SNP effects and test statistics are divided by correction factors to reduce bias and make marker-effect tests less conservative.

One final comment on the GAMMA approach. At first sight, it appears that Equation 20.19a leads to the adjusted trait values (z_i^*) being free of correlations among relatives, and thus exchangeable. If true, this would allow for permutation tests to easily be applied (Chapter 18), randomizing the values of $\mathbf{z}^*$ over the vector of marker information, creating a random draw from the null. Unfortunately, this is only appropriately correct, as BLUPs (predict values) for the u_i have a correlated error structure (Chapter 31), which compromises the assumption of exchangeability. This concern does not apply to fixed-effects, so that the z_i^* are exchangeable with regard to any fixed-effect adjustments.

Mixed Models for Binary Traits

While the above mixed-model adjustments for cryptic relatedness and population structure were presented in the context of a continuous trait, many traits of interest are binary (e.g., disease presence/absence). One approach for such traits is to ignore their dichotomous nature ($z_i = 0, 1$) and simply use the continuous **linear mixed models** (**LMMs**) developed above (Equation 20.18a), e.g., Sawcer et al. (2011). This is an approximate approach, typically done to reduce the computational burden of the analysis. Alternatively, one could perform a more exact—and computationally more demanding— analysis using **generalized linear mixed models** (**GLMMs**), such as logistic regression. We consider these in turn.

The starting point for LMM and GLMM binary trait analysis is to consider a latent variable, which we will call the liability, y_i, associated with individual i. In QK model form, we can express this as

$$y_i = b_k N_{i,k} + \widehat{y}_i \tag{20.20a}$$

where

$$\widehat{y_i} = \mu + \sum_{j=1}^{m} \beta_j x_{i,j} + \sum_j q_{ji} v_j + u_i \tag{20.20b}$$

is the background liability for i, and, as before, our interest is in the significance of b_k, the effect of SNP k. The standard continuous LMM uses $z_i = y_i + e_i$, where the residuals are assumed to be homoscedastic, e.g., $\sigma^2(e_i) = \sigma^2(e)$, namely a constant over all i. This residual structure is invalid for a binary trait, as if $E[y_i] = p_i$, then $\sigma^2(e_i) = p_i(1 - p_i)$, which varies over i. If this variation is small, then the use of a LMM approximation may be reasonable. However, Chen et al. (2016) showed that residual heteroscedasticity was rather problematic in an asthma dataset of individuals of Hispanic/Latino heritage. Different subpopulations had different disease risk, and thus had stratification-specific variances. Using a LMM for this data resulted in incorrect p values. A second issue is that many binary traits use a case-control design. If the disease/trait is rare, then cases are a highly nonrandom sample from the population, and failing to accounting for this ascertainment results in a loss of power (Yang et al. 2014).

Hayeck et al. (2015) and Weissbrod et al. (2015) suggested improvements to account for ascertainment using LMM based on the threshold-liability model (Chapter 30). Under this model, if y_i exceeds some threshold value T, then the disease is present, else it is absent. Hence, $y_i > T$ for cases, while $y_i < T$ for controls. If the disease prevalence and the heritability on the underlying liability scale (Chapter 30) are known, then one can estimate the posterior mean liability for each individuals ($\widehat{y_i}$). This is accomplished using the GRM and the case/control status of all individuals. Then, in the same spirit as Equation 20.19, one performs a standard GWAS regression of $y_i^* = y_i - \widehat{y_i}$ on $N_{i,k}$ for each SNP. This approach was called the **liability-threshold mixed linear model** (**LTMLM**) by Hayeck et al. (2015) and **liability estimator as a phenotype** (**LEAP**) by Weissbrod et al. (2015). The improvement in power of these methods over a standard LMM analysis increases with sample size and with the rareness of the disease.

A more exact approach to acount for shared relatedness in a binary trait GWAS is to used a generalized linear mixed model, such as a logistic regression (Zhu et al. 2002; Setakis et al. 2006; Zheng et al. 2006; Chen et al. 2016; Banerjee et al. 2018; Shenstone et al. 2018). These formally model the correct (binomial) error structure, allowing for the residual variance to vary over individuals. As we saw above, under a logistic regression framework, the underlying latent value y_i is mapped into the expected value on the observed (binary) trait scale z via a logistic regression,

$$p(y_i) = E[z_i \,|\, y_i] = \Pr[z_i = 1 \,|\, y_i] = g(y_i) \tag{20.21a}$$

where the **link function** $g(y)$ is given by the logistic,

$$g(y_i) = \frac{1}{1 + \exp\left[-\left(y_i\right)\right]} \tag{20.21b}$$

where y_i is given by Equation 20.20. This expands on Equation 20.3a by adding fixed-effects terms for population structure (the v_j) and a random effect u_i for the background polygenic value.

Logistic regression introduces a few subtleties typically not present in a LMM analysis. First, the careful reader might recall our above discussion that adding covariates into a logistic regression can actually reduce the power of a GWAS, especially under a case-control design when the disease is fairly rare. This statement still holds for the β_j terms in Equation 20.20b. However, recall that fixed effects are also added to a model to account for confounding factors, reducing the number of false positives. This is the role played by the fixed effects v_j for population structure in Equation 20.20b. Second, the addition of the random effect u_i to control for close relatives makes this a **generalized linear *mixed* model** (**GLMM**), whereas our above discussion of logistic regressions (where the elements

of y were all fixed) are generalized linear models (Chapter 14). While exact solution to a mixed-model logistic can be obtained using either ML or Bayesian methods, the analysis of GLMM raises a number of computational issues (Breslow and Clayton 1993). As a result, they are even more computationally demanding than a LMM GWAS, facing serious scaling problems to accommodate both the number of individuals n and markers m in a modern GWAS.

One approximate approach, the **generalized linear model association test** (**GMMAT**), was suggested by Chen et al. (2016). Their logic was similar to that used for GRAMMAR (Equation 20.19), in that the mixed-model logistic is fit just once for y_i^* (i.e., the model with no SNP effect), using the mixed model parameters to predict this value, $\widehat{y}_i^*$, for individual i. One then does the GLS regression

$$z_i - \widehat{y}_i^* = b_k N_{i,k} + e_i \tag{20.22}$$

where the covariance matrix for the e_i can be expressed in terms of parameters estimated in the original mixed logistic regression (see Chen et al. for details). Zhou et al. (2018) developed an even more efficient computational approach for GLMM-based GWAS, **SAIGE** (**Scalable and Accurate Implementation of GEneralized mixed model**). They did so by exploiting optimization methods (such as saddlepoint approximations and preconditioned conjugate gradients; see their paper for details).

In the early period of GWAS, a case-control design was typically based on sampling a population for a specific disease, and then sampling an appropriate set of controls. However, with growth of large-scale **biobanks** (collections of electronic medical records and genome sequences), GWAS are now often conducted by sampling diseases from these datasets. The result is often very unbalanced numbers of cases versus controls (especially when a focal disease is rare). As noted by Zhou et al. (2018) such case-control imbalances can result in greatly inflated type I errors. Fortunately, the SAIGE approach handles even very unbalanced case/control designs while controlling the type I error.

In the broader context of GLMM, there are other candidates for the link function g (Equation 20.21b) that maps y into the expected value of z besides the logistic. For example, the threshold-liability model uses the probit function, $g(y) = \Pr(U \leq y)$ where U denotes a unit normal, as its link. The choice of the link function involves assumptions about how additivity on the underlying scale maps into interactions on the observed scale (Clayton 2012). Under a logistic link function, additive terms on the underlying scale become multiplicative odds terms on the observed value (Equation 20.3f).

Assessing Model Fit

In a GWAS, as in any model-fitting endeavor, the investigator wishes to assess model quality and determine whether improvements are required. Because the SNP effect is usually the only parameter not regarded as a nuisance variable, the goal is less about improving overall model fit, and more about improving power and controlling the type one error rate. WTCCC (2007) suggested the two approaches most widely used to both assess a current GWAS model and to compare it with others: **probability plots** and genomic control (λ) values.

The motivation for probability plots follows from the powerful observation that p values follow a uniform distribution (over $[0, 1]$) under the null hypothesis (Appendix 6). For a GWAS, the expectation is that *most* SNP effects are zero (i.e., from the null), but a tiny fraction should be true positives. Hence, there should be a slight excess of very small p values, while the rest of the values should be draws from a uniform. Probability plots are a useful device to visually inspect if this pattern holds. One of the most common such plots used in GWAS is constructed as follows. With m tested markers, compute $-\ln(p)$ for each and rank these from smallest to largest,

$$-\ln(p)_{[1]}, -\ln(p)_{[2]}, \cdots, -\ln(p)_{[m]}$$

where $-\ln(p)_{[k]}$ is the value of the kth smallest value of $-\ln(p)$. Note that this scaling awards the smallest p values the highest rankings, so that $-\ln(p)_{[1]}$ corresponds to the test with the

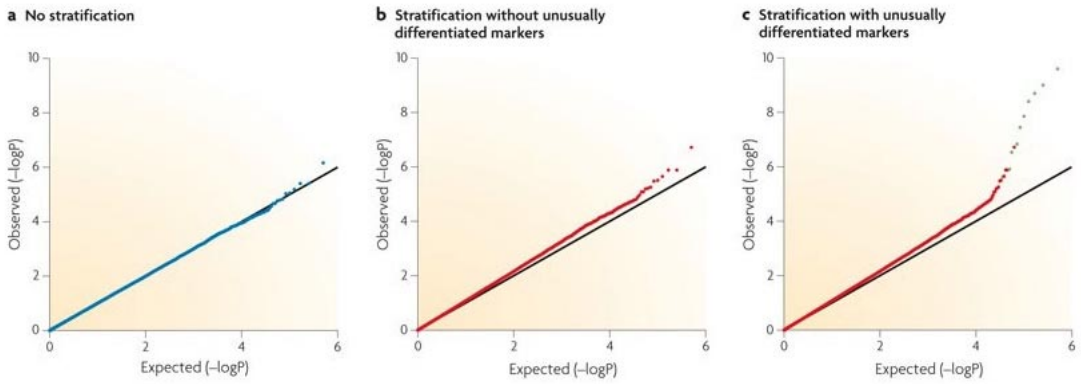

Figure 20.4 Accessing a GWAS model using probability plots. The straight line with a slope of one represents the expected realization between predicted and observed p values under the null (a uniform distribution), while pixels represent p values for individual tests. **Left:** For a population with no stratification, or one corrected for its effects, one might expect a few tests with very small p values (on the far right of the plot) representing true positives, while the rest of the p values follow the expectation under the null. **Center:** When stratification is present, p values are elevated over many of the tests, resulting in a large fraction of the p values exceeding their neutral expectations. **Right:** Stratification coupled with markers showing excessive divergence, resulting in not just departures above the neutral expectation, but in excessive departures. (After Price et al. 2010a.)

p value closest to one, while $-\ln(p)_{[m]}$ corresponds to the test with the smallest p value. This scaling also avoids the compression of very small p values near zero. One then plots $-\ln(p)_{[k]}$ against $-\ln(1-[k-1/2]/m)$, the latter representing the expected kth smallest value under a uniform (the use of $[k-1/2]$ instead of k is a continuity correction). The resulting plot of observed versus expected values should be a straight line with a slope of one, perhaps with a slight, to strong, upturn as k approaches m (representing the small p values of true positives). Deviations from this pattern suggest that the current model needs improvement (Figure 20.4).

There are a number of different approaches for constructing probability plots, which are typically used to test if an empirical distribution matches a theoretical candidate (Wilk and Gnanadesikan 1968; Gerson 1975). The two most common are **p-p** and **q-q** (or **quantile-quantile**) **plots**. If $p_{[k]}$ and $S_{[k]}$ denote the kth smallest p value and test statistic, respectively, then a standard p-p plot graphs $p_{[k]}$ against $(k-1/2)/m$. The common GWAS use of $-\ln(p)$ is a version of a p-p plot, emphasizing the smallest p values, as these same values would be visually indistinguishable from zero if plotted as $p_{[k]}$. Conversely, q-q plots compare the ranked values of the *test statistics* (as opposed to the associated p values of those statistics) with the corresponding expected ranked values from the null distribution. Suppose that the test statistic under the null is a χ^2_1. In a q-q plot, the value of the kth (of m) smallest test statistics is compared with value $f_{[k]}$ that satisfies $\Pr(\chi^2_1 \leq f_{[k]}) = (k-1/2)/m$, generating a point of $(S_{[k]}, f_{[k]})$ in the plot. Again, the expectation under the null is a straight line with a slope of one. WTCCC (2007) used q-q plots, while many other GWAS use the $-\ln(p)$ plots, but both essentially convey the same information. Stirling (1982) discusses the construction of confidence bounds for probability plots.

While an aberrant probability plot is cause for concern, an important cautionary tale was offered by Chen et al. (2016). They observed an apparently well-behaved probability plot for one of their simpler GWAS models. However, a more careful analysis showed this plot resulted from conservative p values in some SNPs balancing out anti-conservative p values from other SNPs. The result was a visually well-behaved probability plot that masked substantial underlying issues.

Typically, a probability plot is presented along with the inflation factor λ for a given

model, with the idea that a well-behaved model should also have a small inflation (typically $\lambda < 1.05$). Recall that λ is a general measure of the inflation of the test statistic for a random marker. Models with different complexities (such as Q, K, or Q+K) are often ranked by their λ values (smaller is better). If a model has a substantial inflation ($\lambda > 1.05$), historically the view was that additional corrections were required. However, as we have seen, this is not strictly correct. If a large number of small effect QTLs underlie the trait, this can inflate λ even after accounting for any confounding effects from structure or cryptic relatedness. Further, this polygenic inflation becomes more pronounced as sample size increases (Equation 20.13a). A more correct genomic-control metric for model comparison is the λ value computed using LD-score regression (Equation 20.13c), which adjusts for polygenicity (Bulik-Sullivan et al. 2015).

GWAS WITH RARE ALLELES

As we discuss in detail in Chapter 21, an ongoing debate is whether a common disease is due to common alleles of small effect or rare alleles of large effect. This is somewhat of a false dichotomy, as numerous studies show that both classes of alleles impact many diseases, and the more appropriate question is their relative contributions. Across a variety of diseases and traits, large-effect alleles generally tend to occur at very low frequencies (Bodmer and Bonillna 2008; WL Chapter 28). One explanation for this observation is that such alleles are likely under negative selection, either by their direct impact on a focal trait, or via pleiotropic effects on other fitness traits (WL Chapter 25). From the theory of mutation-selection balance (WL Chapter 7), we expect such alleles to have frequencies that are inversely proportional to the strength of selection against them. Hence, one view of disease variation is that major-effect alleles (such as amino-acid replacements or knockouts) are under strong selection, and thus tend to be rare, whereas alleles with smaller effects (such a minor regulatory changes) experience weaker selection and thus tend to be at higher frequency. In the extreme where an allele has no impact of fitness, neutral theory predicts (in an idealized equilibrium population) that frequencies follow the Waterson distribution (WL Equation 2.34a), where $\phi(x) = C/x$, with x the (derived) allele frequency and ϕ its density function (pdf). Hence, most neutral alleles are rare. Selection against alleles further shifts this distribution towards even smaller values (as do many demographic features, such as recent population expansions; Kryukov et al. 2009; WL Chapter 9).

A fair assessment of the relative contribution of common versus rare alleles requires unbiased methods that offer equal power of detection of alleles from both classes. The standard one variant at a time GWAS framework discussed above is rather powerful at detecting common alleles (minor allele frequency, MAF, greater than one to five percent), but is very underpowered for detecting rarer alleles (Li and Leal 2008; Asimit and Zeggini 2010; Bansal et al. 2010). There are two reasons for this. First, for an additive locus, the relative contribution is $\sigma^2_Q = 2a^2p(1-p)$ where a is the allelic effect. For two alleles with the same effect, σ^2_Q increases with the value of the minor allele frequency (p). Hence, there is a strong bias in favor of detecting common alleles when the common and rare allele have the same effect size. Given our discussion above, this may be less of a concern, as we might expect a to *increase* as p decreases, so that the average value for σ^2_Q between common and rare alleles could be more similar than we initially expected. Even in this case, a second factor, the correlation (r^2) between the marker and causal alleles, results in greatly reduced power for rare alleles. Example 20.1 shows the reason for this: r^2 decreases as the mismatch between the marker and causal allele frequencies increases. Hence, by mainly using common SNP alleles as markers, we are biasing detection toward common *causal* alleles, even if common and rare alleles have the same value of σ^2_Q. Because different methodologies from traditional GWAS are often used to search for rare variants, the phrase **rare variant association study** (**RVAS**) has been suggested (Auer and Lettre 2015).

A different approach is thus required to discern the impact of rare alleles. Note that we are now in the realm of **deep resequencing** (high coverage sequencing of a small region),

and potentially even **whole-genome sequencing** (**WGS**), as rare alleles are typically not scored by standard DNA chips (Asimit and Zeggini 2010), and accurately imputing them requires a very large reference collection of full sequences (for the region, exome, or genome, of interest). The basic logic behind the menagerie of rare-allele approaches (summarized in Table 20.3) is to pick a candidate region (often a single gene), and then test for an excess of rare alleles in this region in the cases over that seen in the controls (or in high versus low trait-valued individuals; Example 20.10).

As we detail shortly, a variety of approaches and weighting schemes have been proposed to quantify such comparisons (reviewed by Asimit and Zeggini 2010; Bansal et al. 2010; Basu and Pan 2011; Chen et al. 2011; Pan and Shen 2011; Stitziel et al. 2011; Lee et al. 2012, 2014; Ionita-Laza et al. 2013; Derkach et al. 2014; Moutsianas and Morris 2014; Pan et al. 2014; Moutsianas et al. 2015; Nicolae 2016; Santorico and Hendricks 2016; Povysil et al. 2019). Basic design issues for detecting rare variants are discussed by Li and Leal (2009), Li et al. (2011), Lee et al. (2014), and Auer and Lettre (2015).

Example 20.10 The basic logic of rare-allele mapping traces back to Cohen et al. (2004). These authors were interested in the impact of three potential human candidate genes (*ABCA1, APOA1,* and *LCAT*) on HDL cholesterol (HDL-C) levels. The exons of these genes were sequenced in two different high-low group comparison designs. Variants were classified as either nonsynonymous (NS) mutations that changed the amino acid sequence of the encoded protein or synonymous (S) variants that did not. The first comparison involved 128 low individuals (HDL-C score less than the fifth percentile) and 128 high individuals (HDL-C are least in the 95th percentile) in the Dallas Heart Study (a diverse collection including both European and African ancestry). The second study was a corresponding group of 155 lows and 108 highs from a more homogeneous Canadian population. Lumping the variants over the three genes, 10 NS and 19 S mutations were found in both the high and low Dallas groups, 15 NS and 7 S mutants restricted to the low group, and 3 NS and 6 S variants were only in the high group. Similar distributions of variants were found in the Canadian study. Hence, NS variants were significantly more common in the low than in the high group. Of the 18 NS variants restricted to one group, all but two were singletons, with these latter two alleles present at four copies each. The PolyPhen program (Ramensky et al 2002; Adzhubei et al. 2010), which attempts to predict the functional impact of a specific amino acid replacement (benign, possibly damaging, or probably damaging), suggested that these latter two mutations had a benign impact on function. A brief overview of bioinformatical approaches for functional annotation of sequence variation can be found in Auer and Lettre (2015).

A similar study by Ahituv et al. (2007) examined sets of candidate genes in 379 obese and 378 lean individuals. One set consisted of 21 genes with known mutations impacting obesity in mice or humans, and the second set had 37 genes from weight-related pathways. The first group of genes had 46 NS mutations restricted to the obese cohort, but only 26 restricted to the lean group, a significant difference. Further, 19 of NS mutations in the obese cohort were predicted to be deleterious, as opposed to 4 in the lean group. In contrast, the pathway gene group had 72 NS in the obese group, and 69 in the lean group, which was not significantly different.

These two examples highlight the logic for many rare allele tests: contrast the distribution of rare variants between case/control or high/low (**extreme phenotype**) groups. The above examples focused on **private alleles**, those restricted to just one group, and hence are often called **private allele tests**. More generally, one can also incorporate shared alleles. Further, mutational classes could be differentially weighted (such as a focus on only NS mutations). One could go further, scoring the severity of a NS mutation (or other functional change), and include this information in a weighted comparison statistic.

One important source of bias when using rare alleles in a case-control (or extreme phenotype) design was noted by Li and Leal (2009). If rare variants were discovered by

exclusively resequencing only in *cases* and then searching for these same rare variants in controls, the result is a large inflation in false positives. The reason is simple: if an allele is rare, an allele *unrelated to the trait* may by chance be found only in cases. Conversely, a rare allele may only be found in controls. More formally, given a random sample of n (diploid) individuals, for an allele with frequency p,

$$\Pr(\text{at least one allele in sample}) = 1 - (1-p)^{2n} \simeq 1 - e^{-\gamma} \tag{20.23a}$$

where $\gamma = 2np$ is the expected number of allelic copies in the sample, with $e^{-\gamma}$ the probability that no copies are seen. Rearranging Equation 20.23a, the required sample size to ensure a probability π that at least one copy of the rare allele is present is

$$n = \frac{-\ln(1-\pi)}{2p} \tag{20.23b}$$

For example, if $p = 0.001$, the sample size to have a 95% chance of seeing at least one copy is $n = -\ln(0.05)/[0.002] = 1498$. The probability that *no* copies of an allele are found in a sample of n_0 controls while at least one (or more) copies are found in a sample of n_1 cases becomes

$$e^{-\gamma_0}(1 - e^{-\gamma_1}) \tag{20.23c}$$

where $\gamma_i = 2n_i p$. For example, with 3000 cases and 3000 controls, and a rare allele ($p = 1/3000$), the expected number of copies in both samples is 2. The probability that at least one allele is found in the cases is $1 - e^{-2} = 0.8647$, while the probability of at least one copy in the cases and none in the control is $e^{-2}[1 - e^{-2}] = 0.117$. Hence, a number of rare alleles with no impact on the trait would exclusively be found in the cases. More generally, the probability of exactly k copies in the sample follows from the Poisson (Equation 2.21) as $\gamma^k \exp(-\gamma)/k!$, for $k = 0, 1, \cdots$.

Burden and Collapsing Tests

The central problem with detecting rare alleles in a standard GWAS framework (such as using a chi-square or regression test) is that the very large sampling variance associated with rare alleles implies very low power, even when the allelic effect is large. A growing number of approaches have addressed this concern building on the basic logic of Example 20.10: contrasting the *total frequency* of rare alleles between comparison groups. These are often called **burden** or **collapsing tests**: *burden* because the presence of rare alleles might imply a fitness burden, and *collapsing* because rare alleles are consolidated into one (or a few) classes, increasing the effective sample size, reducing the sampling variance and thus increasing power. Collapsing tests can be classified as **indicator** (present/absence of one, or more, rare alleles), **counting** (how many rare alleles are carried by an individual), and **data-adaptive** (scores are weighted by some feature of the data, such as MAF or functional information).

The biological assumption for aggregating rare alleles at a locus into a single class is that **extreme allelic heterogeneity** (**EAH**) is present, where a disease (or extreme trait value) is generated via independent rare mutations. If each mutation has roughly the same effect, then the power of a method that treats them all as a single class is a function of the frequency of that class. For example, Bansal et a. (2010) considered a rare allele with a frequency of 0.01 in the controls and 0.02 in the cases, where 80% power required 250,000 cases and control. However, if there are five such alleles, then when treated as a single class, the frequencies becomes 0.05 in the controls and 0.10 in the cases, now requiring only 3,000 cases and controls for the same power.

Beyond deciding the appropriate scope of the **testing unit**, or **genomic unit of analysis**, (a single gene or set of genes connected in a pathway), a related issue is whether resequencing in the search for rare alleles uses an **exon-only** or a **whole-gene** approach. The logic for exon-only is that variants in the coding region are more likely to be impactful, and less likely to have no effect (often referred to as **neutral** or **null** variants), with the latter diluting a

rare-variant signal (Zawistowski et al. 2010; Pan et al. 2014). Conversely, an exon-only focus assumes that coding mutations are the major source of rare disease/trait variants, ignoring contributions from regulatory effects in noncoding regions (e.g., Haller et al. 2009). For more general genome-wide scans, one might consider an **exome scan** over all coding regions in the genome, or using a genomic sliding window, or scanning all evolutionary conserved regions (which are often noncoding). The problem with such global scans is their relatively modest power within any tested region, coupled with a very high multiple-comparison penalty.

Early versions of burden tests include the **cohort allelic sums test** (**CAST**) of Morgenthaler and Thilly (2007) and the **combined multivariate and collapsing** (**CMC**) method of Li and Leal (2008). CAST simply sums the total number of rare alleles at a target gene (or more generally, region or pathway) and compares the mean number of rare alleles per individual in cases versus controls. Li and Leal (2008) independently proposed a similar collapsing method that compares the fraction of cases carrying rare alleles in a target with the same fraction in controls, which can be tested by using a Fisher's exact test. Li and Leal also suggested an improvement by collapsing variants into a small number of classes based on their allele frequency, and then using a Hotelling's T^2 test (Equation 20.5) over these classes to test significance. This binning approach can accommodate both rare and common alleles, such as collapsing all the rare alleles into one class, and then either considering each common allele separately, or using some collapsing/exclusion scheme on them as well.

While the choice of using only NS mutations (or only private alleles) implicitly assumes a weighting scheme, the next generation of burden tests assigned more formal weights to variants. Madsen and Browning (2009) proposed a **weight sum statistic** (**WSS**) where each variant was weighted by the standard deviation of its frequency,

$$w_j = \sqrt{n_j \widetilde{p}_j (1 - \widetilde{p}_j)}, \quad \text{with} \quad \widetilde{p}_j = \frac{c_j + 1}{2(n_{0,j} + 1)} \tag{12.24a}$$

Here n_j is the total number (cases plus controls) of individuals scored for SNP j, and $\widetilde{p}_j$ corresponds to the **pseudocount** frequency in the controls (adding one to the numerator and denominator to adjust for small-sample issues), where c_j is the number of j alleles seen in the $n_{0,j}$ individuals in the control scored for SNP j. If the allele is absent in the control, then $\widetilde{p}_j = 1/[2(n_j + 1)]$. A **burden score** is then computed for individual i over the m chosen SNPs in region of interest, with

$$S_i = \sum_{j-1}^{m} \frac{I_{ij}}{w_j} \tag{12.24b}$$

where I_{ij} is the indicator function, being one if the minor allele at SNP j is found in i and zero otherwise. One then ranks of the S_i and uses a Wilcoxon test for whether the ranks in cases and controls are nonrandom (such as an excess of higher ranks in cases). Zawistowski et al. (2010) suggested a **cumulative minor-allele test** (**CMAT**) which contrasts weighted minor and major allele counts in cases and controls. Finally, in the **Accumulation of Rare variants Integrated and Extend Locus-specific test** (**ARIEL**) of Asimit et al. (2012), weights on variants are given by a called-sequence quality score (such as a phred score), with rare variants called with higher confidence given more weight.

Using similar logic, burden tests can be constructed for continuous traits. Consider the trend test (Equation 20.1a), but now express the trait value for individual i as a function of gene (or marker region) k as

$$z_i = \mu + b_k \phi_{i,k} + e_i \tag{20.24c}$$

where $\phi_{i,k}$ is some collapsing of the total rare variants at region k in individual i. Morris and Zeggini (2009) suggested using either (i) the proportion of scored rare alleles in the region for individual i (if $n_{i,k}$ rare SNP sites in region k are scored in individual i and $r_{i,k}$ of these contain the rare allele [a count method], then $\phi_{i,k} = r_{i,k}/n_{i,k}$) or (ii) assigning ϕ a value of one whenever *any* rare allele in the region is seen in i, else it has value zero

[an indicator method]. These suggestions are often referred to as the **MZ method** in the literature. As noted by Zawistowski et al. (2010) and Asimit et al. (2012), ϕ can also use **probabilistic genotype calls** (such as the posterior mean allelic dosage discussed above), when rare alleles are either called using low-coverage sequencing (which can be error prone) or by imputation (Mägi et al. 2012).

Significance of any of the above methods can be accessed using standard permutation techniques: keep the genotype information for any individual intact, and then shuffle the phenotypic labels (cases/controls or continuous values). There are two limitations with this approach. First, if population stratification is present, the shuffling has to be done in such a way that stratification information is retained. Second, when multiple comparisons are performed, the resulting smaller p values per comparison (to control the FWER) requires a substantial increase in the number of permutations per region to obtain sufficiently stable estimates of empirical significance thresholds. If R multiple (independent) regions are being compared, a Bonferroni FWER of γ requires testing each region using γ/R. For testing 10 regions, a 5% FWER requires testing each region at $p = 0.005$, which requires at least 10^4 permutations per region. Hence, scaling this approach to more than a small number of candidate regions is both computationally challenging and results in greatly reduced power. For example, with an exome-wide scan scoring roughly 20,000 genes separately, a FWER of 0.05 requires testing each gene at $\alpha = 2.5 \times 10^{-6}$. Estimating stable empirical thresholds to achieve this level of α requires on the order to 10^7 to 10^8 permutations *per gene*.

Variance Component (Dispersion) Tests

Neale et al. (2011) noted that unweighted collapsing methods (treating all chosen rare alleles as functionally equivalent) can be underpowered. In particular, they noted that in a given gene, one could have neutral rare variants with no impact on risk (with the expectation of being equally represented in cases and controls), variants that increase risk (overrepresented in cases), and **protective variants** that *decrease* risk (underrepresented in cases). Early collapsing methods attempted to account for neutral variants by focusing on variants with apparent functional differences (such as using only nonsynonymous, NS, mutations; Example 20.10). However, by lumping potential protective variants with risk variants, power is decreased. In effect, early collapsing methods required a mean directionally of effect for rare variants. In contrast, Neale et al. proposed using *variance*, rather than *mean*, comparisons between cases and controls. Their approach was motivated observations on the *APOB* gene and triglyceride levels. They examined NS *APOB* mutations in a sample of 96 high and 96 low individuals, finding 18 segregating NS mutations in the combined sample. One variant was present as six copies in the high cohort but absent in the low, while a second variant was absent in the high but with six copies in the low. Hence, the first mutation likely increased risk, while the second was likely protective. If both of these are collapsed into the same category, the result is six copies in both the high and low, and no signal.

To deal with this concern, Neale et al. proposed using the **C-alpha**, **C(α)**, test for **overdispersion** (excess variance) in binomial samples due to an underlying mixture of effects (Neyman and Scott 1966). The idea is that the between-group variance would be larger than expected under binomial sampling (assuming equal chances of an allele being in the cases or controls). More formally, let $p_1 = n_{case}/(n_{control} + n_{case})$ be the fraction of the total sample that are cases ($p_1 = 1/2$ when the cases and controls have the same sample size). Suppose there are n_i total number of copies (in cases plus control) of variant i, with y_i of these in cases. For a variant that has no impact on the trait, $y_i \sim B(p_1, n_i)$, a binomial with sample size n_i and success probability p_1 (Equation 2.19a). In this case, Equation 2.19b gives the expected sample variance as $n_i p_1(1 - p_1)$. However, suppose that the true underlying distribution is a mixture,

$$y_i \sim \pi_r B(n_i, p > p_1) + \pi_p B(n_i, p < p_1) + (1 - \pi_r - \pi_p) B(n_i, p = p_1) \tag{20.25a}$$

where π_r and π_p are, respectively, the probabilities of a risk or protective allele. Because of the between-class variance (p varying over classes), such a mixture results in a variance

larger than the expected binomial variance. The C-alpha test statistic, which tests for such an overdispersion over m groups, is given by

$$T = \sum_{i=1}^{m} \left[(y_i - n_i p_1)^2 - n_i p_1 (1 - p_1) \right] \tag{20.25b}$$

Here $(y_i - n_i p_1)$ is the observed deviation from the neutral expected value, namely a standard burden test. The C-alpha test instead uses the squared deviation and contrasts it with its expected value under neutrality (the binomial variance). For example, for the two *APOB* variants discussed above, $y_1 = 6$ and $y_2 = 0$, their two deviations cancel

$$(y_1 - n_1 p_1) = 6 - 3 = 3, \qquad (y_2 - n_2 p_1) = 0 - 3 = -3$$

Conversely, $n_1 p_1 (1 - p_1) = n_2 p_1 (1 - p_1) = 1.5$, yielding $T = 2[9 - 1.5] = 15$. For large samples $T/\sigma(T) \sim N(0, 1)$ under the null, being one-sided test of excessive dispersion (the expression for $\sigma(T)$ can be found in Neale et al.). Note that singletons, by themselves, do not provide information on overdispersion, but the entire collection of singletons can be lumped to form a new category. The *APOB* data found 6 high group and 4 low group singletons, giving $y_i = 6, n_i = 10$ in the summation term in Equation 20.25b. Finally, one can use the EM method (Appendix 4) to estimate the mixture proportions (π_r and π_p) in Equation 20.25a, as well as the average value of p for each category (Neale et al. 2011).

A somewhat related approach that also allows for both protective and risk variants is the **kernel-based adaptive cluster** (**KBAC**) test of Liu and Leal (2010). One unusual feature of KBAC is that it uses the frequencies of *multilocus genotypes* as the unit of analysis, and hence allows for the potential of detecting epistatic interactions that could easily be missed in standard burden tests. The basic structure of KBAC is that one considers all of the observed multilocus genotypes in the region of interest that *contain at least one rare allele*. Suppose there are k such combinations, which we label as $G_1, \cdots, G_k$, while G_0 is the collapsed set of all multilocus genotypes *lacking* any rare alleles. KBAC then computes a weighted sum of the *squared* differences between the frequency of G_i in the cases versus controls. The weights are based on the likelihood (under a neutral model) of the observed number of G_i in the cases. For example, if 7 of 10 copies are found in cases, the weight is the probability (under a neutral model) of seeing 7 or fewer copies in the cases. By considering the squared difference, the impact of directional effects is reduced, and Liu and Leal noted that KBAC is also fairly robust to the inclusion of variants with no functional impact.

As noted above, when most of the variants included in the test set are not causal (the **sparse alternative** setting), methods can lose power. Most of the above models arrive at a final summary statistic by taking some weighted sum of individual deviations, or squares of such deviations, e.g., $S = \sum w_i s_i$. Chen et al. (2012) noted an **exponential combination** (**EC**) procedure, using sum of the *exponent* of these individual statistics, e.g., $\sum \exp(w_i s_i / 2)$, is a much more robust approach under the sparse alternative setting.

Regression-based General Framework

The connections between simultaneously testing multiple markers in a normal GWAS (e.g., Equation 20.5), burden tests, and variance component tests—as well as the relative roles of common versus rare alleles—can all be seem by using a multiple regression framework. We do so by extending the trend test (Equation 20.1a) to m markers, whose inclusion is based on some criteria (such as all being within a target gene or candidate pathway). Let $N_{i,k}$ denote the copy number for the minor allele at marker k in individual i, and let b_k be the associated regression slope. In its simplest form (assuming only additivity), the multiple-marker trend test becomes

$$z_i = \mu + \sum_{k=1}^{m} b_k N_{i,k} + e_i \tag{20.26a}$$

Note that the multiple regression framework accommodates correlations among markers, and (at this point) makes no distinction between common versus rare alleles. Equation

20.26a can be modified to explicitly model dominance or recessivity by replacing $N_{i,k}$ by $X_{i,k}$, with

$$X_{i,k} = \begin{cases} 0 & \text{homozygous recessive} \\ 1 & \text{otherwise} \end{cases} \quad \text{or} \quad X_{i,k} = \begin{cases} 1 & \text{homozygous recessive} \\ 0 & \text{otherwise} \end{cases} \tag{20.26b}$$

for a dominance or recessive, respectively, assumption.

We can further extend Equation 20.26a by adding cofactors (e.g., Equation 20.18a), and likewise to a logistic regression on binary traits by replacing z_i by logit(z_i) (Equation 20.2b). The null hypothesis is that the vector $\mathbf{b}^T = (b_1, b_2, \cdots, b_m)$ of regression coefficients is zero. This can be tested using Hotelling's T^2 statistic, resulting in a χ^2_m, and hence a degree of freedom for each marker. If the chosen alleles are common, then there is often sufficient power, *in isolation* (e.g., testing one marker at a time), to detect those with an effect. However, when testing for the significance of a *region* (as opposed to an *individual* maker), inclusion of markers with no impact on the trait reduces power. This occurs by increasing the degrees of freedom associated with the test statistic (χ^2_m versus a χ^2_1) without a corresponding increase in the noncentrality parameter.

Hence, when testing for the significance of a region we would like some criteria to enrich the fraction of markers with true effects. Collapsing methods, with their general focus on rare alleles, represent an attempt to improve power by both increasing the fraction of potentially causal variants and by reframing Equation 20.26 as single degree-of-freedom statistic. They test the significance of a shared marker effect b_c by rewriting this regression as

$$z_i = \mu + b_c X_i + e_i, \quad \text{with} \quad X_i = \sum_{k=1}^{m} w_k N_{i,k} \tag{20.27}$$

Here X_i is the burden score for individual i, and w_k is the weight for tested marker k. For the set of included markers, one could define X_i in Equation 20.27 has having value one when *any* marker in the testing set has a rare allele, else it has value zero (an indicator approach). One could also set X_i as the total copy number of rare alleles over all scored markers in i (a counting approach). Both these approaches amount to collapsing the multilocus marker genotype into a set of synthetic "super alleles" at a single locus and then performing a standard trend test.

One can view the weights as an attempt to increase the signal from potentially causal variants while decreasing the signal for neutral variants. Thus, at one level, the assumed weights provide a simple **masking scheme** for which variants within a region to include ($w_k = 1$) or exclude ($w_k = 0$). As we saw in Example 20.10, one could focus on only private alleles, or on a specific type of mutation (such as NS vs. S, motivated by the assumption that NS mutations are more likely to be causal than S mutations). One could go further, and only include NS alleles that are predicted to be possibly or probably damaging, or even just probably damaging. While inclusion of variants is often based on such *a prior* information, Hoffmann et al (2010) noted that this approach is only optimal when the information is very accurate. They proposed their **step-up approach**, akin to stepwise regression, for determining the optimal set of variants within a region to include based on model fit alone (and hence is agnostic to any assumed biological knowledge).

More generally, one could base weights on other features, such as the sampling variance given the marker frequency in controls (Equation 20.24). Several **adaptive burden tests** have been proposed, where the weights are informed by single-marker regression slopes of each variant. The **data-adaptive sums test** (**aSUM**) of Han and Pan (2010) assigns weights of $w_i = -1$ when the univariate marker effect is sufficiently negative and $w_i = 1$ when it is positive. The **estimated regression coefficient** (**EREC**) **test** of Lin and Tang (2011) bases weights on the estimated single-marker regression slopes (adjusted for estimate instabilities when alleles are rare). A number of other adaptive tests have been proposed (e.g., Pan and Shen 2011; Pan et al. 2014), again using the logic of upweighting variants with causal signatures (such as being NS) and downweighting those without such signatures. Bayesian

approaches have also been suggested to filter out noncausal variants (e.g., Quintana et al. 2011; Logsdon et al. 2013). A related approach is the **RareCover** method of Bhatia et al. (2010), which starts with a set of k candidate markers within a region, and extracts the set of these giving the highest correlation with phenotypic value.

A basic issue with the collapsing approach that we have not explicitly considered is what exactly constitutes a rare variant. Ionita-Laza et al. (2013), using results from large sample theory, suggested setting the allele frequency threshold as $\tau = 1/\sqrt{2n}$, with n the sample size. For example, for $n = 500$, the threshold is 0.03, while for $n = 10,000$, the threshold becomes 0.007. This is a statistical definition, while the threshold is usually set for biological reasons (enhancing the chance that included markers are causal, or linked to causal sites). Operationally, if the allele frequency threshold is set too high, we may include too many noncausal variants (reducing power), while if set too low we may exclude causal variants. One strategy is to use a **variable threshold** (**VT**) test (Price et al. 2010b), which computes the test statistic over a range of thresholds, and then takes the largest score as the final test value. Again, significance is typically assessed via permutation tests.

An alternative approach to control degrees of freedom is to use a random-effects model for the slopes of marker-trait regression (Dandine-Roulland and Perdry 2015). Here the b_k are drawn from a normal distribution with mean zero and variance $w_k\sigma^2$, where the weights are assigned by some criteria. The test for no marker effects reduces to whether the variance component σ^2 is zero, a one degree of freedom test. This is the approach used in the **sequence kernel association test**, or **SKAT**, of Wu et al. (2011), which is a generalization of the C-alpha test, and allows for both positive and negative effects. Indeed, it can be shown that C-alpha is a special case of SKAT, but the latter method has greater flexibility in that cofactors are easily incorporated. The **SumSq** (also referred to as the **SSU**) test of Pan (2009), based on $\sum b_k^2$ (the sum of squared estimated slopes), can also be shown to be a special case of SKAT.

Finally, Xu et al. (2012) noted that the machinery of penalized regressions (Example 20.4; Chapter 31) offers a very powerful approach for incorporating rare variants (also see Zhou et al. 2010). Penalized regression methods are designed to both handle over-parameterized models and models with sparse data (Hastie et al. 2009). Xu et al. showed that both LASSO and RR tended to outperform both burden (VT) and variance tests (SKAT), unless causal alleles are extreme rare (or singletons). Further, both LASSO and RR allow for the joint test of common and rare variants, with Chen et al. (2011) noting that using the LASSO to choose which common variants to include greatly improves power in the setting of many rare, but few common, causal variants.

One closing comment about the regression framework. As stressed above, correction for population structure is often a serious concern in a traditional GWAS, and this is generally accomplished by adding cofactors to Equation 20.26a (e.g., Equations 20.14b and 20.17a). One unresolved issue is that ***the population structure for rare alleles can be rather different from that for common alleles*** (Mathieson and McVean 2012). Common alleles tend to have long evolutionary histories, whereas most rare alleles are rather recent events, potentially leading to different population structures. Further, because of their recent history, rare variants can be spatially clustered (for example, due to different histories of recent expansions, some subpopulations may harbor many more rare variants than others). Hence, using common alleles when constructing cofactor corrections for structure can bias rare-allele tests (but see Listgarten et al. 2013). In theory, one could construct corrections using only low-frequency to rare alleles (MAF less than 0.5%), but these can potentially be unstable and may have a smoother eigenvalue distribution (so that more PCs must be included to capture a stratification signal). Zaidi and Mathieson (2020) further examine PC corrections based on only rare alleles. An especially problematic issue was noted by Zhang et al. (2013). They found that PC-based structure correction based on low frequency variants tended to overadjust in the absence of structure, leading to a substantial loss of power.

Omnibus Tests

A number of authors have examined the power of rare-allele tests (e.g., Basu and Pan

2011; Ladouceur et al. 2012; Lee et al 2012; Moutsianas et al. 2015). When all of the chosen variants have roughly the same effect and act in the same direction, burden tests are the most powerful. Conversely, when the chosen variants are a mixture of neutral, protective, and risk variants, variance component methods (e.g., SKAT) are more powerful. Given that the optimal method depends upon unknown details of the genetic architecture of the trait, several **omnibus tests** have been proposed that combine burden and variance-component tests. The logic is that by combining information from two (or more) tests, there will be more robust power under a random genetic architecture. Conversely, an omnibus test might loose power if one of the extreme architectures is correct (e.g., equal effects in the same direction).

Derkach et al. (2013) proposed using Fisher's method (Equation A6.1a) to combine the p values from a burden test and the SKAT test,

$$T_{Fish} = -2\ln(p_{SKAT}) - 2\ln(p_{burden}) \tag{20.28a}$$

where T_{Fish} follows a χ^2_4 distribution. A related approach is the **Mixed effects Score Test** (**MiST**) of Sun et al. (2013). Their method starts by considering the slope b_k in Equation 20.26a as $b_k = b_t + \delta_k$, where b_t is a fixed effect for the slope when the variant is of type t (such as a NS mutation), assumed to be the same for all type t variants, and δ_k is a random effect unique to variant k. Burden tests assume a constant value for the fixed effect b_t, while variance component tests use the variance of δ_k. Both SKAT and standard burden tests are special cases of this more general slope model, with MiST based on a joint test of the mean effect of a slope (b_t nonzero) and the variance effect of the slope ($\sigma^2[\delta_k] > 0$). They cleverly constructed these two test to be independent, so that Fisher's test can be used to combine their two probability values (as in Equation 20.28a). They also used **Tippett's method** (1931): for k independent test, their combined p value is given by

$$1 - [1 - \min_k(p_k)]^k \tag{20.28b}$$

The Tippett threshold for an overall level of α is the

$$\min_k(p_k) \leq 1 - (1-\alpha)^{1/k} \tag{20.28c}$$

As noted by Westberg (1985), neither Fisher's or Tippet's method is uniformly superior to the other over all settings. Sun et al. (2013) noted that Fisher's procedure seems to be more powerful when *both* the mean and variance effects of b_k are nonzero, while Tippett's is more powerful when only one component is nonzero. Another common method for combining p values from independent tests is Stouffer's Z score (Equation A6.2). Using the same logic as above, tests could be constructed using Stouffer's method in place of the Fisher or Tippet method. More generally, one can combine the p values from a chosen set of k different tests (e.g., Moutsianas et al. 2015; Liu et al. 2019).

The limitation of standard p-combining methods (Fisher, Tippett, Stouffer) is the requirement that tests are independent. The recently proposed **aggregated Cauchy association test** (**ACAT**) approach of Liu et al. (2019) relaxes this assumption. The key is that they translated p values into Cauchy random variables. The Cauchy (the distribution of the ratio of two unit normals) is unusual, in that its density function integrates to one (making it a proper distribution), but none of its moments are finite, because its tails are sufficiently heavy (do not decay sufficiently fast at large values). As a result of this heavy-tail feature, it is largely insensitive to correlations among p values (especially when the p values are small). Second, ACAT upweights small p values, so that (like the EC method) it is a more robust approach under the sparse alternative setting (many neutral variants in the test set). The ACAT test statistic is

$$T_{ACAT} = \sum_{i=1}^{n} w_i \tan[(0.5 - p_i)\pi] \tag{20.29a}$$

with the associated overall p value being

$$p \simeq 0.5 - \frac{\arctan[T_{ACAT}/w]}{\pi} \tag{20.29b}$$

Table 20.3 Rare allele mapping approaches. *General modification* procedures can be applied to fine-tune most of the tests. See text for details on specific tests.

Unweighted Burden Tests		
CAST	**Cohort allelic sums test**	Morgenthaler and Thilly (2007)
CMC	**Combined multivariate and collapsing test**	Li and Leal (2008)
MZ		Morris and Zeggini (2009)
Weighted (Adaptive) Burden Tests		
WSS	**Weighted sum statistic**	Madsen and Browning (2009)
CMAT	**Cumulative minor-allele test**	Zawistowski et al. (2010)
aSUM	**Data-adaptive sums test**	Han and Pan (2010)
Step-up		Hoffmann et al (2010)
RareCover		Bhatia et al (2010)
EREC	**Estimated regression coefficient**	Lin and Tang (2011)
ARIEL	**Accumulation of Rare variants Integrated and Extend Locus-specific test**	Asimit et al. (2012)
Variance Component		
SumSq		Pan (2009)
KBAC	**Kernel-based adaptive cluster**	Liu and Leal (2010)
C(α)	**C-alpha**	Neale et al. (2011)
SKAT	**Sequence kernel association test**	Wu et al. (2011)
Omnibus Tests (Burden plus variance component)		
SKAT-O	**SKAT-optimized**	Lee et al. (2012)
	Fisher's method	Derkach et al. (2013)
MiST	**Mixed effects Score Test**	Sun et al. (2013)
ACAT-O	**Aggregating Cauchy Asssociation test**	Liu et al. (2019)
Omnibus Tests (Common plus rare variants)		
Burden-F, **Burden-C** **SKAT-F**, **SKAT-C**		Ionita-Laza et al. (2013)
Omnibus Tests (Weighted p values)		
σ-MidP		Cheung et al. (2012)
ADA	**Adaptive combination of *P*-values for rare variant association testing**	Lin et al. (2014)
General Modifications: Variable rare-allele frequency threshold		
VT	**Variable threshold**	Price et al. (2010b)
General Modifications: Sparse alternative setting		
EC	**Exponential combination**	Chen et al. (2012)
ACAT-V	**Aggregating Cauchy Association test**	Liu et al. (2019)

where $w = \sum w_i$. Liu's **ACAT-V** test examines each rare variant within the target region, weighting then by a function of their frequency (a modification of Equation 20.24a), and then combines the individual p values using Equation 20.29a. Given the robustness of ACAT under the sparse alternative, ACAT-V easily accommodates regions with a large number of neutral variants. Liu et al. also proposed an omnibus test by combining the p values from six separate tests: ACAT-V, SKAT, and a burden test, where each test is performed assuming equal weight and then recomputed using weights that place more emphasis on rare alleles. The p values from these six test are then combined using ACAT,

$$T_{ACAT-O} = \frac{1}{6}\sum_{i=1}^{6} \tan[(0.5 - p_i)\pi] \tag{20.29c}$$

The idea of using this combination is to construct a test that it relatively robust over different

weights, varying directionality of variants, and fraction of neutral variants.

Lee et al. (2012) proposed their **SKAT-optimized** (**SKAT-O**) test, which constructs a linear combination of the a Burden and a SKAT test statistic

$$T_{SKAT-O} = \rho\, T_{SKAT} + (1-\rho)\, T_{burden} \tag{20.30}$$

where ρ, determined via a grid search over [0,1], is the value that maximizes Equation 20.30. Lee et al. showed how the corresponding p value for this test statistic can be obtained via numerical integration.

As noted by Ionita-Laza et al. (2013), the idea of omnibus tests can be extended to jointly accommodate common and rare alleles. Their **Burden-F** tests computes p values for a burden test using rare alleles and then applies the same test using common alleles, combining the resulting p values using Fisher's method. Similarly, their **SKAT-F** tests Fisher-combines the p values from separate SKAT tests based on rare and common alleles. They also constructed **Burden-C** and **SKAT-C** akin to Equation 20.30 by weighted combinations of Burden and SKAT statistics for a test using only common and a test using only rare alleles.

Finally, variations on combining p values have been proposed that weight individual variant $\ln(p_i)$ terms (from a case-control contrast) in Fisher's method by their MAF (using Equation 20.24a). The ***σ*-MidP** approach of Cheung et al. (2012) used this weighting scheme, but excluded rare alleles with roughly equal counts in cases and controls (to control for the reduction in power from neutral alleles). The **Adaptive combination of P-values for rare variant association testing** (or **ADA**) method of Lin et al. (2014) modifies this rare alleles exclusion criteria. They do so by (i) using a variable MAF threshold for inclusion of a variant, and (ii) computing separate sums for putative risk alleles (more frequent in cases) and protective alleles (more frequent in controls), with the larger of these two sums being the test statistic for that threshold. This approach is performed over a range of threshold values, with the largest of the test values being the final test statistic. Again, permutation tests are used for significance testing in both of these methods.

Closing Remarks

As summarized in Table 20.3, a number of rare variant tests have been proposed, built around different assumptions about the distribution of variant effects within the region being considered. As such, in the absence of any knowledge about the trait architecture in the region, there is no uniformly most powerful test. Omnibus tests are a bet-hedging approach, sacrificing a little power under specific extreme architectures, while having more power under random architectures. The other issue is that the amount of signal within a given region could easily generate a significant p value, but one not large enough to persist under the much more stringent multiple comparison value required for an exome-wide or genome-wide scan. Unless sample size is massive, rare-variant approaches are best considered as tests over a small to modest number of regions, rather than the whole genome scan of a typical GWAS.

META-ANALYSIS

GWAS projects, especially in humans, are often done under a **consortium model**, wherein k groups perform independent GWAS on a given trait. If one has access to the individual data from each study, then, in theory, one could combine all of the data sets and run a single analysis. Such an approach is often called a **mega-analysis** or a **joint-analysis**. However, this is often not possible. Medical privacy concerns usually prevent members of a consortium from sharing their raw data. Further, for a variety of reasons, the study designs may not be compatible, such as involving different confounding variables or using different markers. The latter issue is typically dealt with by using imputation (de Bakker et al. 2008; Zaitlen and Eskin 2010), creating a set of shared markers over all the studies (this assumes that each study sampled from a population where accurate imputation is possible). Hence, for each shared marker, the consortium typically has only **summary statistics** from each of the k studies

(rather than individual data), such as marker p values, estimated effects, and standard errors. A final potential roadblock to performing a full mega-analysis is computational (Panagiotou et al. 2013). While each of the k individual studies may be computationally feasible, a mega-analysis of a full mixed-model may be less so.

The field of **meta-analysis** (the analysis of analyses; Appendix 6) deals with the analysis of such a collection of summary statistics. While the p values for a given marker may be combined over the studies, for a variety of reasons (see Appendix 6), a formal meta-analysis typically deals with the *estimated effects* of a given marker. Lin and Zeng (2009, 2010; also see Olkin and Sampson 1998) showed under rather general conditions that the standard error for a meta-analysis estimate is approximately the same as a mega-analysis estimate using the same data. Hence, in most settings, little efficiency is lost when moving from an analysis based on the individual data to an analysis based on summary statistics from each of the underlying studies. There is a rich, and growing, literature on the application of meta-analysis to GWAS data. de Bakker et al. (2008), Zeggini and Ioannidis (2009), and Evangelou and Ioannidis (2013) offer nice overviews of some the practical aspects of gathering, and **harmonizing**, the data for a GWAS meta-analysis. General reviews of meta-analysis methodology as applied to GWAS are given by Munafó and Flint (2004), Kavvoura and Ioannidis (2008), Trikalinos et al. (2008), Begum et al. (2012), Panagiotou et al. (2013), and Dudbridge and Newcombe (2019).

Meta-analysis Basics: Fixed Versus Random Effects Analysis

We begin with a quick overview of the machinery of meta-analysis (more fully covered in Appendix 6) before examining specific applications to GWAS. Let T_i be some estimate of the marker effect in study i (such as the odds ratio or regression slope), s_i^2 its sample variance, and θ_i the true value for that study. Because of sampling error,

$$T_i = \theta_i + e_i \tag{20.31a}$$

where we assume that the residuals are independent but heteroscedastic, as $\sigma^2(e_i) = s_i^2$. Under a **fixed-effects (FE) meta-analysis**, we assume that the actual effect size is the *same* over all studies ($\theta_i = \theta$). Recalling generalized least-squares (GLS; Equation 10.13a), the meta-analysis estimate of θ becomes

$$\overline{T} = \frac{\sum_{i=1}^k w_i T_i}{\sum_{i=1}^k w_i}, \quad \text{where} \quad w_i = \frac{1}{s_i^2} \tag{20.31b}$$

In other words, we use a weighted average, with each study weighted by its precision (studies with smaller standard errors receive larger weights). Assuming a similar individual variance (σ^2) over studies, $E[s_i^2] = \sigma^2/n_i$, so that this scheme places more weight on studies with larger sample sizes. However, the individual variance could easily vary over studies. For example, a smaller study incorporating appropriate cofactor corrections, or with more accurate scoring of the phenotype, may have a smaller standard error than a more poorly designed, but larger, study. When imputation is used, Zaitlen and Eskin (2010) noted that imputation accuracy is likely to vary over studies, and suggested a modification of the weights to reflect this.

The meta-analysis standard error, $s_{\overline{T}}$, for the global (meta) estimate, $\overline{T}$, is

$$s_{\overline{T}}^2 = \frac{1}{\sum_{i=1}^k w_i} \tag{20.31c}$$

For the situation where we assume that each individual observation in a given study has the same variance (σ^2), then for k studies with size n_i,

$$\sigma^2(\overline{T}) = \frac{\sigma^2}{\sum_{i=1}^k n_i} = \frac{\sigma^2}{nk}, \quad \text{if} \;\; n_i = n \tag{20.31d}$$

The assumption of a single common value for the treatment mean over all studies is often unrealistic, as we might except the true marker effect to vary, at least somewhat, over studies. In this setting, our interest shifts to the variance *among* the actual effects over studies. This leads to the **random-effects** (**RE**) **meta-analysis** model

$$T_i = \mu + u_i + e_i \tag{20.32a}$$

where $u_i \sim (0, \sigma_u^2)$. Typically, the effect sizes ($\theta_i = \mu + u_i$) are assumed to be drawn from a normal, $\theta_i \sim \mathrm{N}(\mu, \sigma_u^2)$. In addition to estimating the grand mean (μ), under the RE framework our interest also extends to the **heterogeneity** of the studies, measured by the variation (σ_u^2) among the realized effects. The estimate of μ is also of the form of Equation 20.31b, but with a critical difference. Under a random-effects model, the weights are now given by

$$w_i = \frac{1}{s_i^2 + \widehat{\sigma}_u^2} \tag{20.32b}$$

where $\widehat{\sigma}_u^2$ is the estimate of σ_u^2. Again, the standard error of the meta estimate is given by Equation 20.31c, but now using the RE weights (Equation 20.32b).

This difference in study weighting between FE and RE analyses has several important consequences. First, when $\sigma_u^2 = 0$, FE and RE estimates are identical. Second, when $\sigma_u^2 > 0$, RE standard errors are larger, and as a result, an RE analysis almost always has lower power than an FE analysis. Third, a subtle feature of RE weights arises when σ_u^2 is on the same order as an average value (s^2) of s_i^2 (namely, the between-study variance is at least large as the within-study variance). In this setting, the RE w_i tend to ***be more similar over studies***. The result is that, relative to an FE analysis, studies with less precision (for example, due to smaller size) are ***given more weight*** under RE than under FE. In the extreme where $\sigma_u^2 \gg s_i^2$, the weights are roughly the same over all studies, independent of their individual precision.

Example 20.11. A modification of the basic random effects (RE) approach has been suggested by Han and Eskin (2011) and Lee et al. (2017). They noted that one reason for the loss of power under an RE analysis (relative to an FE analysis) is testing against an *incorrect null hypothesis*, and suggested an improved likelihood ratio test (Appendix 4) to account for this. Their argument is as follows. As above, let T_i and s_i^2 denote the estimate and its variance. For large sample size, T_i is distributed as $N(\theta_i, s_i^2)$, giving the meta-analysis likelihood as the product of k normals. Under the FE model ($\theta_i = \mu$), the null ($\mu = 0$) and alternative ($\mu \neq 0$) likelihoods become

$$L_{FE,null} = \prod_i^k \frac{1}{\sqrt{2\pi s_i^2}} \exp\left(-\frac{T_i^2}{2s_i^2}\right) \tag{20.33a}$$

$$L_{FE,alt} = \prod_i^k \frac{1}{\sqrt{2\pi s_i^2}} \exp\left(-\frac{(T_i - \mu)^2}{2s_i^2}\right) \tag{20.33b}$$

Substituting the MLE ($\widehat{\theta}$) for θ (which Han and Eskin show is given by Equation 20.31b) into the likelihood for the alternative and computing the likelihood ratio gives a test for the significance of θ (Equation A4.9a).

Under the standard random-effects model, $T_i \sim N(\mu, s_i^2 + \sigma_u^2)$, giving the classic RE null and alternative likelihoods as

$$L_{RE,null} = \prod_i^k \frac{1}{\sqrt{2\pi(s_i^2 + \sigma_u^2)}} \exp\left(-\frac{T_i^2}{2(s_i^2 + \sigma_u^2)}\right) \tag{20.33c}$$

and

$$L_{RE,alt} = \prod_i^k \frac{1}{\sqrt{2\pi(s_i^2 + \sigma_u^2)}} \exp\left(-\frac{(T_i - \mu)^2}{2(s_i^2 + \sigma_u^2)}\right) \tag{20.33d}$$

Han and Eskin noted that Equation 20.33c is for the null hypothesis of a *mean* zero effect ($\mu = 0$), but still allows for a *variance* in the mean effect ($\sigma_u^2 > 0$). Framed this way, the "null" hypothesis here is a mean effect of zero, but allowing for heterogeneity.

The **RE2 model** of Han and Eskin takes the null as $\mu = 0$ with *no heterogeneity* ($\sigma_u^2 = 0$), resulting in the same null as under the FE model (Equation 20.33a), and the alternative as under the RE model (Equation 20.33d), giving the RE2 likelihoods as

$$L_{RE2,null} = \prod_i^k \frac{1}{\sqrt{2\pi s_i^2}} \exp\left(-\frac{T_i^2}{2s_i^2}\right) \tag{20.33e}$$

$$L_{RE2,alt} = \prod_i^k \frac{1}{\sqrt{2\pi (s_i^2 + \sigma_u^2)}} \exp\left(-\frac{(T_i - \mu)^2}{2(s_i^2 + \sigma_u^2)}\right) \tag{20.33f}$$

As above, the ratio of these two likelihoods (with the MLEs substituted into Equation 20.33f) generates the test statistic, which becomes

$$LR_{HE} = \sum_{i=1}^k \ln\left(\frac{s_i^2}{s_i^2 + \widehat{\sigma_u^2}}\right) + \sum_{i=1}^k \frac{\widehat{\theta_i}^2}{s_i^2} + \sum_{i=1}^k \frac{(\widehat{\theta_i} - \widehat{\mu})^2}{s_i^2 + \widehat{\sigma_u^2}} \tag{20.33h}$$

which has a large-sample distribution that is a weighted sum of chi-square distributions, namely, $(1/2)(\chi_1^2 + \chi_2^2)$. Equation 20.33h is partitioned into a heterogeneity component (the first term, testing $\sigma_u^2 = 0$), an FE component (the second term, testing $\mu = 0$), and a final term considering both. Han and Eskin showed that Equation 20.33h is more powerful than the FE test when sufficient heterogeneity is present.

Meta-analysis Basics: Heterogeneity

Despite its lower precision and power, the strength of an RE analysis is in capturing any heterogeneity (variance in true effect sizes) over studies. As we will see below, in many GWAS studies, this is *as important*, indeed in some cases *more important*, than estimating an average effect size. The simplest test for variance in effect sizes is the **Cochran *Q*** test of heterogeneity,

$$Q = \sum_{i=1}^k \frac{\left(T_i - \overline{T}\right)^2}{s_i^2} \tag{20.34}$$

where (under the null of $\theta_1 = \cdots = \theta_k$, and assuming that the values of T_i are normally distributed) the distribution of Q is χ^2 with $(k-1)$ degrees of freedom. While a standard reported metric in a meta-analysis, Q is *grossly underpowered* when k is small, so that a non-significant value *does not* imply a lack of heterogeneity. Indeed, when k is small, often a less strict standard ($p \leq 0.1$) is used to declare heterogeneity (e.g., Dudbridge and Newcombe 2019). Conversely, for large k, Q may be overpowered in the sense of declaring biologically trivial heterogeneity as being statistically significant.

While Q is a test statistic for heterogeneity, it can also be used to obtain an estimate of the between-sample variance, σ_u^2. The **DerSimonian-Laird estimator** is given by

$$\widehat{\sigma}_u^2 = \frac{Q - (k-1)}{S_1 - (S_2/S_1)}, \quad \text{where} \quad S_j = \sum_{i=1}^k s_i^{-2j} \quad \text{for} \quad j = 1, 2 \tag{20.35}$$

which is set to zero if negative (DerSimonian and Laird 1986; 2015). Other estimation approaches (e.g., REML; Chapter 32) have been proposed (Appendix 6).

Higgins and Thompson (2002) noted two problems with quantifying the amount of heterogeneity using either Q or an estimate of σ_u^2. First, the expected value of Q is $k-1$, a function of the number of studies. Second, the value of σ_u^2 is dependent on the scale

of measurement and test statistic used, and hence not readily comparable over different meta-analyses. Hence, neither is an optimal measure for general heterogeneity, and they proposed two related metrics for this task. Their ***H*** statistic is given by

$$H = \sqrt{\frac{Q}{k-1}} \tag{20.36a}$$

and measures the excess in Q over its expected value. A modification is the **Mittlbock-Heinzl** (2006) statistic,

$$H_M^2 = \frac{Q-(k-1)}{k-1} = H^2 - 1 \tag{20.36b}$$

Equation 20.36a is the basis for the **Higgins-Thompson index** of heterogeneity

$$I^2 = \frac{H^2-1}{H^2} = \frac{Q-(k-1)}{Q} \tag{20.36c}$$

where

$$E[I^2] = \frac{\sigma_u^2}{\sigma_u^2 + s^2} \tag{20.36d}$$

with

$$s^2 = (k-1)\frac{S_1}{S_1^2 - S_2} \tag{20.36e}$$

measuring the average within-sample variation. The nice feature of I^2 is its natural interpretation as the fraction of the total variation ($\sigma_u^2 + s^2$) due to heterogeneity (σ_u^2). Equation A6.36c can be used to construct approximate confidence intervals for I^2 (as the latter is a function of H). While I^2 is widely used, Nakaoka and Inoue (2009) noted that there are settings where H_M^2 might be a slightly better metric.

Meta-Analysis of a GWAS Collection: Basic Issues

A meta-analysis of a collection of independent GWAS has three aims: (i) the discovery of new marker-trait associations not found in any single study, (ii) replication of any initial associations, and (iii) detection of any variation in marker effects over studies. An early example of the power of a meta-analaysis is from Crohn's disease (Franke et al. 2010). Six previous GWAS studies (comprising a total of 15,000 controls and 6,300 cases) had localized 32 different markers associated with the disease. A meta-analysis of these studies identified 30 additional associations.

The results of a meta-analysis for a given marker are usually displayed using **forest plots** (Figure 20.5), which provide a visual representation of the individual studies, heterogeneity, and display the final estimates (often both the FE and RE values). These can also be constructed as **cumulative forest plots**, where the top row is the initial study, the second row the meta-analysis using the first two studies, and so on, with the final row giving the full meta-analysis (Figure 20.5). These are usually displayed (from top to bottom) in the chronological order in which the studies appeared. The observation that the first study often has the largest effect has been termed the **Proteus phenomena** by Ioannidis and Trikalinos (2005).

As mentioned, a meta-analysis has multiple objectives. If the goal is simply detecting significant marker effects, then the higher power of an FE analysis is recommended, especially given the very stringent p values required to control the FWER (Begum et al. 2012; Evangelou and Ioannidis 2013; Panagiotou et al. 2013; Dudbridge and Newcombe 2019). Use of the FE model is appropriate because the null hypothesis of no marker effect *jointly* implies $\mu = 0$ (no average effect) *and* $\sigma_u^2 = 0$ (no heterogeneity). In theory, one could have $\mu = 0$ (the *average* marker effect is zero), but still have considerable marker heterogeneity (*individual* studies have significant marker effects). If there is a concern of heterogeneity,

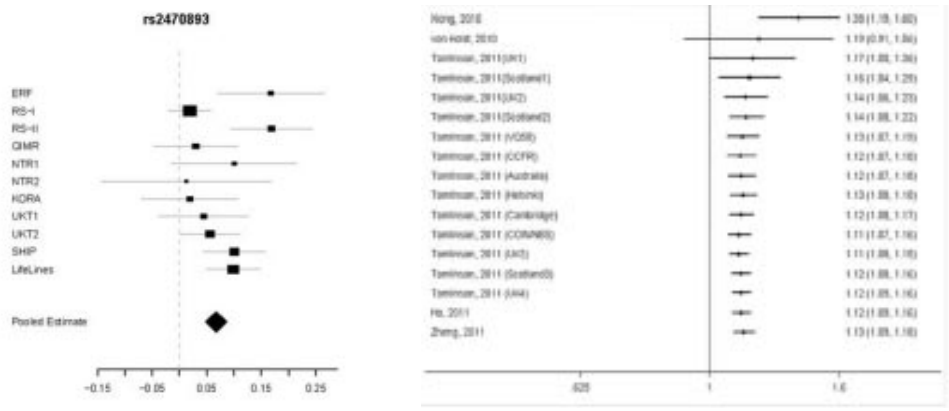

Figure 20.5 **Left:** A meta-analysis **forest plot** for a particular SNP (rs2470893) for a series of GWAS examining coffee consumption. Each row represents a particular study, with middle of the box locating the study estimate, the volume of the box representing the weight given to that study, and the lines on both sides of the box (often called **whiskers**) represent the 95% confidence interval. The vertical dotted line denotes the **line of no effect**. The diamond at the bottom of the plot represents the meta-analysis estimate, with the peak of the diamond denoting the estimate, and the edges denoting the width of the confidence interval. (After Amin et al. 2012.) **Right:** A **cumulative forest plot** for a meta-analysis of rs961253 on colorectal cancer risk. The top row represents the odds ratio estimate in the initial association, and each subsequent row is the meta-analysis of all of the preceeding studies. Note the Proteus phenomena effect, with the initial study having the largest effect. (After Zheng et al. 2012.)

then the RE2 model of Han and Eskin (Example 20.11) should also be used, as it has greater power than FE when sufficient heterogeneity is present. While an FE analysis is typically used for detection, if heterogeneity is present, the resulting FE confidence intervals for effect size are too narrow, and an RE analysis is more appropriate. Recall that an RE analysis usually downweights larger studies and upweights smaller studies (relative to an FE analysis), which may of concern in some analyses.

One complication that can arise in a GWAS meta-analysis is when the same subjects overlap in one (or more) studies (such as when shared controls are used). This creates correlations among the study estimates, which is problematic, as standard meta-analysis models assume these are independent. Lin and Sullivan (2009) addressed this concern by developing a GLS estimate that accounts for such correlations. For studies i and j, the expected correlations are

$$r_{ij} \simeq \begin{cases} n_{ij}/\sqrt{n_i n_i} & \text{population sample} \\ \left(n_{ij0}\sqrt{\dfrac{n_{i1}n_{j1}}{n_{i0}n_{j0}}} + n_{ij1}\sqrt{\dfrac{n_{i0}n_{j0}}{n_{i1}n_{j1}}} \right) \Big/ \sqrt{n_i n_j} & \text{case-control} \end{cases} \tag{20.37}$$

where n_{ij}, n_{ij1}, and n_{ij0} are, respectively, the total number of shared subjects, shared cases, and shared controls for studies i and j, while n_i, n_{i1}, and n_{i0} are, respectively, the sample size, number of cases, and number of controls in study i. When studies are independent, the covariance matrix used in the GLS meta-estimate is diagonal, leading to Equation 20.31b. When studies are correlated, the among-study covariance matrix now has off-diagonal elements from these correlations, with estimates following from standard GLS expressions (Equation 10.13a). An analogous situation occurs when a mega-analysis is performed over a series of studies that contain relatives (such as mouse inbred lines). Furlotte et al. (2012) show how to use an estimated relationship matrix to appropriately weight the studies.

While meta-analyses are typically framed in terms of testing the effects of a single marker on a single trait over a collection of studies, their utility is much more general.

For example, one could conduct a meta-analysis of a *single marker* over different *traits*, namely a ***search for pleiotropic effects*** of a marker on multiple traits. Cotsapas et al. (2011) examined the impact of 107 immune disease-risk SNPs detected for one disease on their impact on other immune-mediated diseases. To do so, they developed a **cross phenotype meta-analysis** (**CPMA**) test. Their concern was, given a detected marker for one disease, how can we test for an association with some, but not necessarily all, of the other diseases. The logic of their test (once again) follows from the assumption of a uniform distribution of p values under the null (Appendix 6). This can be restated as $-\ln(p)$ following an exponential distribution (Appendix 7) with decay rate $\lambda = 1$. Their CPMA test statistic is simply a likelihood ratio of the data under the null ($\lambda = 1$) versus the likelihood under the MLE ($\widehat{\lambda}$) for the fitted λ given the p values for the other diseases. Using this approach, they found that 47 of the 107 SNPs (44%) were associated with some, but not necessarily all, of the other diseases, so that (at least) 44% had pleiotropic effects.

Finally, an efficient meta-analysis heavily relies on stable marker estimates from each individual study, and hence requires that the minor allele frequency (MAF) at the focal marker (in each GWAS) is not too small. We have previously discussed various rare-allele approaches for when this assumption fails (Table 20.3), and these statistics for each study can be used as the entries in a meta-analysis, see Lee et al. (2013) for details. Alternately, a mega-analysis (where possible) might result in a rare allele being sufficiently common in the full dataset to allow for stable estimates of individual effects (Panagiotou et al. 2013). Ma et al. (2013) showed that a joint analysis based on testing single markers can be more powerful than a meta-analysis when the total **minor allele count** (**MAC**) is less than 400. In this setting, they found (with a case-control design) that using logistic regression with a penalized likelihood (the Firth biased-corrected test; Firth 1993; Heinze and Schemper 2002) over the joint data was the most appropriate analysis.

Meta-Analysis of a GWAS Collection: Heterogeneity

As noted by Ioannidis et al. (2007), heterogeneity in GWAS is both very common and "*is a useful aspect of the data, rather than a nuisance, as it can often point to leads that can clarify better the nature of postulated association in the context of meta-analysis*." Heterogeneity can be formally quantified using I^2 (and its confidence interval), and visually displayed with forest plots (Figure 20.5). When significant heterogeneity is present, a **sensitivity analysis** may provide insight into its causal sources. Here, one assesses the effects of removing specific studies on the meta estimate and I^2.

At its simplest level, an understanding of potential sources of heterogeneity, and correcting for them, can result in a more powerful meta-analysis (greater power for detection and tighter confidence intervals). Different trait ascertainment criteria, differential scoring of the trait, and testing the same marker with different genetic models (e.g., fitting a recessive model in some studies and an additive in others) are all potential sources of heterogeneity that can often be easily addressed. More subtle issues involves differences between populations in generating different levels of LD between markers and causal loci, genetic heterogeneity (causal alleles vary over studies), and differential exposures (variation in environmental and/or genetic backgrounds). A deeper exploration of causes of heterogeneity can result in significant biological insight (e.g., Example 20.12). For example, Zeggini and Ioannidis (2009) found significant heterogeneity in a meta-analysis for the association between Type 2 diabetes (T2B) and the *FTO* locus known to be involved in obesity. In settings where cases and controls were matched for body mass index (BMI), no association between *FTO* and diabetes was found. Hence, it appears that *FTO* predicted obesity, which in turn predicted diabetes, yielding some of the earlier detected associations.

Magosi et al. (2017) noted that heterogeneity is often better thought of as a *study-wide*, rather than a *marker-specific*, effect, with some feature(s) of the study generating a systematic signal over a large set of markers (**systematic heterogeneity**). Hence, there is potentially more power to detect such an effect by jointly considering an appropriate collection of markers. This is the basis for Magosi's **aggregate heterogeneity *M*** statistic, which is computed

for each study. M is based on on the average of the scaled deviations of observed marker effects from their meta-analysis predicted values. These scaled deviations are referred to as **standardized predicted random effects** (**SPREs**), with the SPRE score for marker i (of m) in study j given by

$$\text{SPRE}_{j,i} = \frac{T_{j,i} - \widehat{\theta}_{j,i}}{\sqrt{\sigma^2_{u,i} + \sigma^2_{j,i} - E^2_{j,i}}} \tag{20.37a}$$

where $T_{j,i}$ is the association statistic for marker i in study j, $\sigma^2_{j,i}$ its sample variance, and $\widehat{\theta}_{j,i}$ its predicted realization. The other variance terms are $\sigma^2_{u,i}$, the between-study variance for marker i, and $E_{j,i}$ the prediction standard error (Chapter 31). The resulting M value associated with study j is given by

$$M_j = \frac{1}{m}\sum_{i=1}^{m} \text{SPRE}_{j,i} \tag{20.37b}$$

Extreme values of M_j indicate potential outlier studies, whose systematic effects result in marker values being consistently overpredicted or consistently underpredicted, relative to the rest of the studies. Statistical tests follow as M_j is approximately distributed as a unit normal. Magosi et al. applied their approach to a coronary artery disease (CAD) meta-analysis involving 48 different GWAS studies which involved a mixture of ethnicities. Based on M scores, they found that studies with early-onset cases, those that used family history for ascertainment, and those with individuals of East Asian ancestry explained a significant proportion of between-study variation.

While systematic study effects represent one end of the spectrum of heterogeneity sources, at the opposite end is marker-specific heterogeneity. For a given marker, this could be generated by variation in causal alleles over studies, epistatic interactions over different genetic backgrounds, or G x E interactions caused by major environmental factors varying over studies. Kang et al. (2014) exploited this idea by noting that marker-specific heterogeneity can provide insight into G x E when study interactions correlate with environmental features. They examined 17 mouse studies (with a total of roughly 5000 mice), performing separate meta-analyses on 26 markers showing significant effects on HDL cholesterol levels. By considering different mouse lines under different environment conditions, these studies varied in major environmental features (such as diet), sex, and in genetic background (the presence/absence of knockout genes impacting other cholesterol pathway genes). For markers with highly significant I^2 values, they attempted to correlate heterogeneity with specific study factors.

At such high-heterogeneity markers, some studies are expected to have interactions while others are not. To determine which particular studies show an effect, Kang et al. used the **m statistic** developed by Han and Eskin (2011). This estimates the posterior probability that an effect exists in a particular study, given that the collection as a whole is significant. This is an extension of the notion of the posterior probability of association (Equation 20.7b). Han and Eskin suggested that studies with $m < 0.1$ are predicted to have no effect, studies with $m > 0.9$ are predicted to have an effect, and intermediate m values are ambiguous. Han and Eskin (2012) and Kang et al. (2016) proposed that using a **P-M plot** (Figure 20.6)—plotting the p value for a given study and its associated m value—was much more informative as to which studies have effects. Their reasoning was that the m value for a given study borrows cross-study information, while its p value uses only within-study information. For example, they note that markers can have $m > 0.9$, and yet not have very significant p values (e.g., study 2 in Figure 20.6).

Example 20.12. In one of the first human obesity GWAS, Herbert et al. (2006) found a significant signal at the rs7566605 SNP associated with the *INSIG2* gene, with the CC genotype

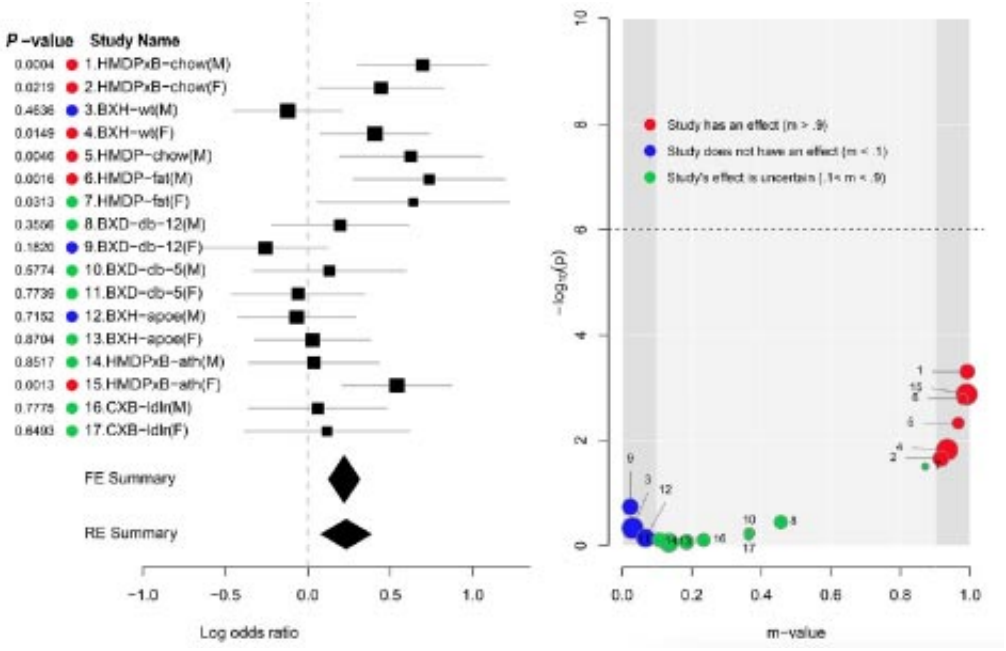

Figure 20.6 **Left:** A forest plot for the *Fabp3* gene in a meta-analysis on mouse HDL levels. Note studies 3 and 4, BXH-wt(M) and BXH-wt(F), which differ only in sex (males vs. females). The impact of this marker on females (study 4) was significantly positive (the confidence interval was entirely greater than zero), while the mean impact on weight in males (study 3) was negative, but not significantly so, showing a clear sex effect of this marker in this background (the BXH wildtype strain) and a common environment (high fat diet). **Right:** The associated **P-M plot** for these studies. For each study, $-\log_{10}(p)$ is plotted on the vertical axis and m (posterior probability that the effect is present) on the horizontal axis. Note that while the confidence intervals for studies 3 and 4 overlap (left panel), so that their heterogeneity might simply be the result of sampling noise, their relative locations on the P-M plot offers some clarity (right panel). Study 3 had a low m value (0.03), while study 4 had a high m value (0.93), supporting very little (if any) effect in males, but a strong effect in females. (After Kang et al. 2016.)

increasing obesity relative to the CG or GG genotypes (a recessive model). This observation proved hard to replicate, occurring in some, but not other, followup studies. Heid et al. (2009) examined if this lack of repeatability was due to an initial false-positive or was a consequence of study heterogeneity. A meta-analysis of 27 studies encompassing 66,000 Caucasians using a case (body mass index, BMI, ≥ 30) versus control (BMI < 30) design found a significant effect (estimated odds ratio, OR, of 1.076, $p = 0.023$) under a fixed-effect analysis, but not under a random-effects analysis (OR of 1.051, $p = 0.268$). For the RE analysis, I^2 was 41% with a confidence interval of 6.6% to 62.8%, indicating significant heterogeneity (a p value of 0.015 for Q). The OR and significance values increased as the case BMI threshold increased (using controls with BMI < 25), with OR values of 1.16, 1.18, 1.22, and 1.27 for BMI cutoffs of 32.5, 35, 37.5, and 40.0. They then broke the 27 studies into three sets: 16 general populations (GP), six obese populations (OP), and five healthy populations (HP). In the GP GWAS set, there was a significant effect (OR = 1.092, p = 0.035) in a random-effects model, and a reduction in I^2 (down to 10.9%, confidence interval of 0% to 48.1%). No significant effect was seen in the OP analysis, which showed a high level of heterogeneity (I^2 of 63.2%, Q with a p value of 0.018). Surprisingly, in the HP population, the CC genotype had a significant *protective* effect, with an odds ratio of 0.796 and with no heterogeneity ($I^2 = 0$). The authors suggested that the *INSIG2* gene is associated with *extreme* obesity, a signal that can be masked by the study design when an insufficient number of such individuals are sampled.