21

Quantitative Genomics and Probing the Nature of Quantitative Genetic Variation

Whether the goal is discovery of rare variants or common variants, sample sizes are a key limiting factor for furthering our understanding of polygenic diseases, and increasing sample size remains a research priority needed to further ... genetic discoveries. Wray et al. (2018)

Version 20 Nov. 2022

The previous chapter examined the powerful GWAS approach for detecting associations between SNPs (and potentially other markers, such as CNVs) and trait values at a target phenotype. In the vast majority of cases, detected markers (GWAS "hits") are not themselves functional, but rather "tag" nearby causal variants via linkage disequilibrium (LD). A GWAS amounts to a very high precision QTL mapping experiment, offering insight into the **genetic architecture** of traits (the number of causal variants and their joint distribution of effects and frequencies). What a GWAS leaves unresolved is the actual molecular nature of causal variants, and how this variation imparts developmental and physiological effects on a target trait or disease. The very nature of high LD that allows one to tag causal variants also impedes their very-fine mapping (i.e., detection of causal **quantitative trait nucleotides**, **QTNs**), as numerous variants around a causal site may be in almost complete LD.

GWAS was fueled by the rise of high-throughput DNA scoring platforms, culminating in rare allele methods using whole-genome sequence data (Chapter 20). Concurrent with this rise in DNA technologies was the development of other high-throughput functional genomics platforms. These scored intermediate features in the pathway between a DNA sequence and the final trait phenotype, such as **epigenetic modification** of the DNA (changes in chromatin status and configuration) and levels of downstream transcripts, proteins, and metabolites. This chapter examines what has been called the **post-GWAS era** (Zhang et al. 2014; Gallagher and Chen-Plotkin 2018), combining information from GWAS studies and functional genomics to probe the nature of quantitative variation. We attempt to weave together a diverse collection of topics, ranging from a more detailed description of gene regulation to analytic methods for exploiting high-dimensional data sets. Here we conclude with an overview of our current understanding of the evolutionary forces that have shaped existing genetic variation, how this translates into patterns of genetic architectures, and offer some glimpses into the molecular mechanisms underpinning such variation.

We start with high-throughput studies of gene expression (microarrays and RNA-Seq) that provide a snapshot of the **transcriptome** (the genomic set of all transcribed regions). Genome-wide expression studies have had three major impacts on the field of quantitative genetics. First, the amount of transcript for any given gene is a *quantitative trait*, as we can quantify its value, and variation in this value typically has both genetic and environmental components. Thus, we can perform standard quantitative-genetic (QG) tasks such as estimating the heritability of expression levels, search for genetic correlations in expression over different genes, probe its variance components (e.g., how much nonadditive genetic variance is present), test for G x E, and so on. The thousands of features scored in a single experiment by high-throughput expression platforms thus represent a collection of thousands of quantitative traits, allowing one to explore a distribution of trait architectures. Second, when marker data is jointly gathered with expression data, one can use methods discussed in Chapters 17–20 to map **expression QTLs** (**eQTLs**, or **eSNPs** when tagged by SNPs), markers tagging genetic regions that influence the expression level of a target gene.

Finally, if expression levels and trait values are both scored, one can search for **quantitative trait transcripts** (**QTTs**), whose expression levels influence downstream trait values.

While our initial presentation is focused on the transcriptome, it is important to stress that *essentially any functional genomics feature of interest can be treated in a similar manner*. The QG framework used for analysis of the transcriptome easily extends to other genomic features, such as the **proteome** or **metabolome** (the sets, respectively, of all proteins or metabolites within a focal cell or tissue). It also applies to various regulatory intermediates, such as chromatin structure and mRNA processing, leading us to next consider more general **regulatory QTLs** (**regQTLs**) impacting these features, which, in turn, leads to discussions on the method of **transcript-wide association studies** (**TWAS**) and statistical methods to fine-map actual causal sites.

One of the early results from microarray studies was the generation of long lists of **differentially expressed genes** (**DEGs**) between two different cell types (such as normal and diseased). The resulting massive datasets sparked the explosive growth of statistical methods—loosely called **gene-set** and **pathway analyses**—developed to search for patterns among the lists of DEGs, such as enrichment of specific functional classes of genes and/or pathways. We present a modest overview of this vast literature, which offers a number of approaches for extending a GWAS analysis from the SNP or gene level (Chapter 20) to the level of user-defined gene sets (such as known biological pathways or genes in the same functional category). Our increasing ability for high-throughput analyses of genomic features raises considerations about how to model and estimate these fine-scale interactions. We briefly introduce the emerging field of **systems biology**, which attempts to model these highly complex systems. As the analysis of gene expression shows, quantitative genetics and systems biology represent natural starting units for a more holistic fusion of genome biology.

We conclude by examining the current picture on quantitative variation offered by existing GWAS and functional genomics studies. We start with an older debate as to whether the bulk of quantitative genetic variation is due to common alleles of small effect or rare alleles of large effects. We then turn to the observation of "missing heritability," wherein the variance accounted for by using just the significant GWAS hits is only a small fraction of the value seen using phenotypic correlations among relatives (Chapters 22–24). Finally, we consider current hypotheses, such as the **omnigenic model**, that attempt to provide a general framework for the nature of quantitative variation.

As with Chapter 20, much of our focus here is on human studies, because this is where the largest investment has been made in developing genomic resources and tools. The logic and methodologies presented here apply equally well to most other study organisms, and can often be further leveraged by exploiting unique biological features that a focal species might possess. For example, Hymenoptera have haploid males and diploid females, a feature that can be exploited by a clever investigator. Likewise, the ability in some species to generate clones can result in much more accurate measurement of functional genomic features by averaging over a population of clones (Example 21.1), as opposed to using less precise measurements based on single individuals, or more derivative measurements based on tissue cultures developed from these individuals.

PROBING THE TRANSCRIPTOME

Whole-genome Expression Analysis

Paralleling the rise of rapid whole-genome sequencing approaches was the development of a number of **high-throughput profiling technologies** that simultaneously score very large collections of cellular macromolecules, starting with whole-genome transcription analysis (the transcriptome). First-generation transcriptome approaches were based on **microarrays**, which used RNA-DNA or cDNA-DNA hybridization to short gene-specific probes (spotted in an array) to simultaneously score thousands to tens-of-thousands of transcripts (Brown and Botstein 1999; Hedge et al. 2000; Berrar et al. 2003). This was an analog tech-

nology, scoring the amount of hybridization by the brightness of a given probe (spot) when made to fluoresce under a laser. This pioneering approach has been slowly replaced by a digital second-generation technology, **RNA-seq** (**RNA sequencing**). This method uses next-generation sequencing to count the actual number of copies of a given transcript in the sample (Mortazavi et al. 2008; Nagalakshmi et al. 2008; W. Wang et al. 2009). Besides being more accurate, RNA-seq also has more discrimination than a microarray, such as being able to quantify the amounts of alternately spliced products (**isoforms**) of a given primary transcript or detect allele-specific expression (e.g., Battle et al. 2014). The resolution of this technology is such that one can perform **single cell RNA sequencing** (**scRNA-seq**) to generate cell-type specific transcription profiles, which can then be tested for trait predictability (e.g., Watanabe et al. 2019; Võsa et al. 2021). Sources of systematic errors and quality control issues in RNAseq are discussed by Li et al. (2014) and Feng et al. (2015)

Gene expression is expected to be exquisitely tailored to specific tissues, environments, and developmental windows. As such, even when examining a genetically homogenous collection of a unicellular organisms (such as yeast), care must to take to standardize environmental conditions and development stages (such as rapid growth versus stationary phases). For metazoans, the situation is even more delicate. At a very crude level, one could perform whole-organism RNA extraction, again under the caveat of using consistent environments and developmental stages. However, in metazoans, the focus is usually on a particular tissue or set of tissues deemed relevant to the trait or question under study, such as endosperm in seeds, muscle in cattle, or immortalized cell lines or blood from humans. Accomplishing this standardization limits the tissues/environments/conditions examined, which can skew the biological interpretation of any results by the bias introduced by the choice of tissues and conditions. For example, seed yield may be constrained by root uptake of limiting nutrients, which could be missed by a focus on expression in flower or seed tissues. Even if the transcript change is similar across all tissues, a modest to small change in one tissue may have a more dramatic effect in another.

One of the very surprising results from genomics is that the transcriptome is a much more wild and unruly universe than was perceived just a few decades ago (Johnson et al. 2005; Carninci 2006; Gustincich et al 2006; Kapranov et al. 2007; Amaral et al. 2008; Kapranov and St Laurent 2012; Deniz and Erman 2017). The historical view was that the vast bulk of the transcriptome was mRNA transcribed from protein coding genes, along with a few other specialty RNAs (tRNAs, rRNAs) critical for translation. As such, extraction and scoring methods tended to focus on mRNAs, for example by exploiting the presence of a poly-A tail in most processed transcripts. It is now apparent that, at some level, most of the human genome appears to be transcribed. Some sources for these transcripts are obvious, such as those associated with mobile genetic elements and noise from leaky or read-through transcription. However, it is also clear that there are numerous, and very diverse, classes of **noncoding RNAs** (**ncRNAs**; also denoted as **transcripts of unknown function**, or **TUFs**), that play critical roles in gene regulation, roles that are still being resolved. These RNAs are often partitioned by size into either **short ncRNAs** (**sncRNAs**; transcript less than 200 nucleotides) or **long ncRNAs** (denoted as either **lncRNAs** or **lincRNAs**, for long intergenetic noncoding RNAs). Important classes of sncRNAs include **small nuclear RNAs** (**snRNAs**) involved in regulating gene splicing, **small nucleolar RNAs** (**snoRNAs**) involved in modification of functional RNAs (e.g., rRNAs, tRNAs, and snRNAs), and **micro-RNAs** (**miRNAs**) and **small interfering RNAs** (**siRNAs**) involved in post-transcriptional regulation of gene expression. While the roles of lincRNAs are less resolved (Deniz and Erman 2017), given that there are over 15,000 lincRNAs in the human genome their impact is likely nontrivial. Given the unknown roles of many ncRNAs, in a bit of physics envy Johnson et al. (2005) referred to these as **transcriptional dark matter** (or the **dark transcriptome**). It is highly likely that some of this "dark matter" is relevant to quantitative trait variation (St Laurent et al. 2014; Issler and Chen 2015; Li et al. 2016; Gamazon et al. 2018; X. Liu et al. 2019), especially under the view that much of phenotypic variation within a population is due to regulatory variation. Given that many biologists hold the RNA world view of

the first protocell (with RNA performing both coding and metabolic functions before being displaced by DNA and proteins), this greatly expanded functional role for RNAs in current cells should really not be that surprising.

Genetical Genomics and eQTLS: Basics

As was discussed in Chapter 18, the first mapping of regulatory QTLs was by Damerval et al. (1994), who detected QTLs influencing the spot volume for anonymous proteins in maize (Figure 18.13), which was followed by similar studies in mice (Klose et al. 2002; Hartl et al. 2008). Jansen and Nap (2001; de Koning and Haley 2005) coined the term **genetical genomics** for this marriage of QTL mapping and expression of genomic features. Jansen (2003) further suggested that functional analysis in a segregating population can be a more powerful tool than more traditional single-gene perturbations (such as gene modification/knockouts or iRNA silencing). With the development of high throughput RNA expression platforms, the initial focus of genetical genomic studies was on the transcriptome. Most of the early studies used classical linkage-based QTL mapping approaches (Chapters 18 and 19), with the trait value being the amount of transcript from a focal gene, an integrated measure of both its transcription rate and message stability.

The study and mapping of such **eQTLs** (**expression QTLs**; Jansen and Nap 2001; Schadt et al. 2003) that influence the mRNA level of a target gene is of great interest for several reasons. First, as noted by Rockman and Kruglyak (2006), "the road from genotype to phenotype runs through gene expression." Most traditional trait-based GWAS SNPs map to noncoding regions (e.g., Edwards et al. 2013 found that 85% mapped to noncoding regions in humans), and hence likely represent regulatory, as opposed to structural, variants. While regulatory changes can occur at many levels—transcriptional, post-transcriptional, translational, post-translational, cell-cell or tissue-tissue interactions, etc.—many of these regulatory variants are likely eQTLs. As a result, the location of GWAS SNPs can be very misleading as to the location of causal coding regions (e.g., Example 21.5). Suppose that variation in the amount of transcript in gene Q impacts trait value/disease status. Variation in regulation in transcript abundance might be governed by sites at some distance from the coding region, with the resulting GWAS hits drawing attention away from Q. Conversely, eQTLs can also map close to the coding region for the transcript they influence (see below). In such cases, a correlation between trait value and transcript abundance of a focal gene can provide support for that gene influencing the focal trait. Thus, eQTLs and GWAS SNPs can jointly provide support for causal genes. Finally, the ability to score thousands of transcripts in a single experiment offers a large, relatively unbiased, set of characters upon which to draw influence about the distribution of genetic architectures over this class of traits. Overviews of eQTLs are given by Rockman and Kruglyak (2006), Gilad et al. (2008), Nica and Dermitzakis (2013), Albert and Kruglyak (2015), and Hill et al. (2021).

Early eQTL mapping studies included linkage-based designs using line-crosses (F_2 and RILs; Chapter 18) in yeast (Brem et al. 2002; Yvert et al. 2003; Brem and Kruglyak 2005), mice (Schadt et al. 2003; Bystrykh et al. 2005; Chesler et al. 2005), rats (Hubner et al. 2005; Petretto et al. 2006), maize (Schadt et al. 2003), and *Arabidopsis* (West et al. 2007). Early work in humans was based on outcrossed pedigree designs (Chapter 19), with expression levels being scored using either cell lines or blood extracted from each sampled individual (Schadt et al. 2003; Monks et al. 2004; Morley et al. 2004; Deutsch et al. 2005). As with any linkage-based design, the level of resolution of eQTLs was poor (typically on a megabase scale), especially given the small sample sizes of most of these studies (usually < 100 individuals/lines). These initial linkage-based studies were soon followed by LD-based GWAS in humans (Cheung et al. 2005; Stranger et al. 2005, 2007a, 2007; Emilsson et al. 2008; Dimas et al. 2009; Dixon et al. 2009) and maize (Fu et al. 2013; H. Liu et al. 2017), moving from eQTLs to eSNPs. While such **expression-wide association studies** (**EWAS**) offered greater mapping resolution, early GWAS studies also suffered from very small sample sizes (again, typically less than 100), and thus were dramatically underpowered.

As discussed in Chapter 20, structural DNA variation—such as copy number variation

(CNV), insertions/deletions (indels), and presence/absence of mobile elements—can be used in place of SNPs as markers in a GWAS. Often these structural variants are tagged via LD to nearby SNPs (**tCNVs**, for **tagged copy number variations**; Gamazon and Stranger 2015), and hence their effects are largely captured by a SNP-based GWAS. However, many structural variants are not well tagged by SNPs, and there are reasons to suspect that such variants often have a direct impact on expression levels of nearby transcripts (leading to **eCNVs**). Their impact could simply be due a duplication resulting in double the amount of transcript, or through more subtle effects such as duplication or deletion of control elements (e.g., enhancers). Consistent with this argument, Gamazon et al. (2011) found that tCNVs were enriched for eQTLs compared to SNPs with matching allele frequencies. Stranger et al. (2007) found that CNVs accounted for around 18% of the total detected variation in gene expression in their study, which is likely an underestimate, as other CNV effects could have been captured by nearby SNPs. Bryois et al. (2014) found that CNVs were more likely to be eQTLs than were SNPs, and that CNVs can act in *trans* as well as *cis*. They also found (as did Chiang et al. 2017 and Uzunovíc et al. 2019) that the effects of structural variants on expression were usually larger than SNP effects. One of the most detailed analysis of the relative contributions from SNPs and structural variants was by Jakubosky et al. (2020), who examined *cis* eQTLs (within 1 MB on either side of the coding region) for 7000 genes showing genetic variation in expression. Over 7 million common variants (SNPs, indels, and other structural variants) were scored, finding 11,000 lead makers impacting expression, with 72% being SNPs, 24% being indels, while the rest were other structure variants.

The major difference between an eQTL/EWAS and a QTL/GWAS experiment is that, in the former, thousands of traits (gene-specific expression levels) are scored at once. Further, many of these traits are highly correlated, reflecting coordinated expression over sets of genes. This trait dimensionality structure introduces two issues. First, it creates a heavy multiple-comparison burden (testing over thousands of individual traits). Coupling this burden of higher stridency levels for each test with the typically small sample sizes of many eQTL experiments made these early studies very underpowered (Gibson and Weir 2005). One consequence of lower power is poor mapping resolution and potentially unstable estimates of eQTL locations (Pérez-Enciso 2004). A second concern are Beavis (winner's curse) effects (Figure 18.8), wherein detected effects are substantially overestimated due to low power (de Koning and Haley 2005).

While many eQTL studies map one transcript at a time (or, conversely, test all transcripts over a given marker, one marker at a time), greater power can be achieved by leveraging information from these correlated values (Kendziorski and Wang 2006). When the number of transcripts greatly exceeds the number of scored samples, methods for mapping low-dimensional correlated traits from Chapter 18 are not applicable. Extensions of standard QTLs methods to handle these high-dimensional, but correlated, traits were developed by Kendziorski et al. (2006), Chen and Kendziorski (2007), Gelfond et al. (2007), Jia and Xu (2007), Chun and Keleş (2009), Zou and Zeng (2009), Wang et al. (2011), Shabalin (2012), Flutre et al. (2013), Davis et al. (2016), and Ongen et al. (2016). An alternative approach has been to use principal components (or, more generally, the singular value decomposition; Appendix 3) to reduce the dimensionality of the transcript data, and then map the resulting PC/SV values as composite traits (Alter et al. 2000; Lan et al. 2003; Biswas et al. 2008). Sul et al. (2013) suggested meta-analytic approaches for combining results over multiple tissues.

Given the potentially complex nature of the transcript correlation structure, permutation tests (Chapter 18) remain the gold-standard for significance testing for eQTLs, but the exchangeable (shuffled) unit is no longer single traits. Permutation tests should instead be constructing by *keeping the vector of expression data intact*. If $\mathbf{e}_i$ and $\mathbf{g}_i$ denote the vectors of expression and marker data (respectively) for individual i, then the randomization should be $\mathbf{e}_i$ over $\mathbf{g}_j$, namely holding each vector intact, but shuffling their association.

Example 21.1 Several features about the biology of Baker's yeast (*Saccaromyces cerevisiae*)

make it an excellent a model system for classical and molecular genetics. Following the cross of two strains, recombinant haploid spores are produced, each of which can be rapidly grown into pure colonies (generating **segregant lines**, the haploid version of RILs). Because each line is a haploid, no dominance is present and any epistasis is additive (e.g., AA, AAA, etc.; Equation 5.9). Finally, yeast has a very high recombination rate, offering increased mapping resolution. Geneticists have long exploited these features, but the lack of any significant morphological variation focused early quantitative-genetic studies on physiological traits such as growth rate. More recently, high throughput systems to phenotype yeast morphological traits have been developed (e.g., Ohya et al 2005), but these still only score a limited number of features. Whole-genome expression analysis introduced thousands of new traits that could be cheaply and rapidly scored, allowing for eQTL mapping. The resulting large sample of eQTLs for different transcripts provides insight into the distribution of genetic architectures underlying expression variation.

An early exploration of yeast expression architectures was done by Brem and Kruglyak (2005), who scored 112 haploid segregant lines for expression level at 5727 transcripts. These lines were also genotyped using 3000 markers, with eQTL mapping performed using the standard RIL mapping framework (Chapter 18), modified for haploid lines. While the clonal nature of each segregant allowed phenotypes (expression levels) to be measured with some precision, the power of this design was still modest given the small number of lines. Over 3500 transcripts showed a significant among-line variation, indicating heritable variation for expression levels. Of these, 2000 had at least one detectable eQTL, with the authors noting that the other 1500 heritable transcripts likely had eQTLs with effects too small to be detected. A similar observation was made in mice by Schadt et al. (2003), who detected eQTLs for only about a quarter of the transcripts that showed significant among-line expression differences.

Brem and Kruglyak noted that the difference between the observed among-line variation and the variation attributed to detected eQTLs provides information on the effect size of undetected eQTLs. To remove any Beavis effects (overestimation of effect size when power is weak), they split their 112 lines into a random set of 56 lines for detection, with the remaining set of 56 used to estimate the eQTL effects (this procedure is called a **subsampling approach**). They found that only 3% of highly heritable transcripts were consistent with a single controlling locus, roughly 20% were consistent with one or two locus control, and over half required at least five eQTLs. The latter bound assumed that undetected eQTLs had equal effects, so that the actual number is likely much higher. Transgressive segregation (Chapter 18) was common, where expression levels in some of the lines exceeded the values seen in their parents. Finally, because no dominance is present in the haploid lines, departure of the mean of the segregant lines from the midparent value (line cross analysis; Chapter 11) indicates (additive) epistasis, which was seen in roughly 20% of the highly heritable transcripts. A much more powerful follow-up study by Albert et al. (2018) using over 1000 segregant lines was able to account for over 70% of the heritability in expression with mapped eQTLs. As predicted by Brem and Kruglyak, control of transcript expression was highly polygenic, with a median number of 6 eQTLs per transcript, where the majority of variation was accounted for by eQTLs acting at some distance from their target transcript coding region.

Example 21.2 Battle et al. (2014) performed an EWAS using RNA-seq on 922 human samples of whole blood, finding detectable eSNPs for over ten thousand transcripts. Using this large sample of traits, a number of patterns consistent with selection against variants with a large effect on expression emerged. As in Example 21.1, subsampling was used to remove Beavis effects, with the resulting estimated eSNP effect sizes tending to decrease as the variant frequency increased. This same eQTL variant size–frequency pattern was detected in the mustard *Capsella grandiflora* by Josephs et al. (2015). Battle et al. found two other patterns suggesting that transcript level changes were more moderated for genes with potentially large, or important, roles in cellular function. Reduced eQTL variation was seem for both highly conserved genes and for genes whose products were hubs in protein-protein interaction (PPI) networks (Appendix A2). Further, they showed that eQTL effect size was negatively correlated with the number of PPIs for the product of that transcript.

A followup study on eQTLs in *C. grandiflora* by Uzunovíc et al. (2019) found that CNVs, in the form of transposable elements (TEs), made a significant contribution to expression

variation. They observed that rare TE eQTLs tended to strongly downregulate expression, in contrast to rare SNPs that showed no net directionality in expression levels. Conversely, common TE eQTLs were more likely to increase expression. These observations suggest TE insertions that downregulate expression are generally selected against, but variants that up-regulated expression were often less deleterious.

Genetical Genomics and eQTLS: *Cis* Versus *Trans* Effects

The poor resolution of linkage-based eQTL locations has an important implication in their functional interpretation. In classic genetics, regulatory factors are typically classified as either having ***cis*** or ***trans*** functionality (Haldane 1942; Lewis 1945). A *cis* regulatory site only impacts a gene residing on the same DNA molecule, and typically represents regulatory binding sites such as promoters or regional features such as enhancers, silencers, or insulators. While *cis* sites are usually thought of as being adjacent to the coding sequence they impact, in theory they act at some distance (especially given the tendency of chromatin looping, wherein sites megabases apart may still find themselves in very close proximity in a cell; Rao et al. 2014). *Trans* regulatory factors, on the other hand, are generally envisioned as diffusable products (such as proteins or RNAs) that can impact genes residing on the same, or different, chromosomes (for example, by binding to *cis* sites). The *cis* terminology was used in the early days of eQTL mapping to refer to an eQTL location that mapped very close to the coding region of the target transcript site, typically within a megabase from the transcription start (**TSS**), or end (**TES**). *Trans* referred to an eQTL that was either at some distance away on the same chromosome, or on an entirely different chromosome, from the target transcript coding region. Given that these are *distance*, as opposed to *functional*, metrics, they are often replaced in the literature with the terms of **local** (or **proximal**) and **distant** (or **distal**) eQTLs (Rockman and Kruglyak 2006; Gilad et al. 2008). Given the poor resolution from linkage mapping, "local" can operationally refer to regions on a megabase scale, which could contain multiple eQTLs.

Note that local eQTLs could functionally act in *trans*. One classic example would be regulatory feedback loops, where the amount of mRNA or protein product from a gene self-regulates its own expression. One can formally distinguish between *cis* versus *trans* functionality using **allele specific expression**, **ASE** (Wright and Moyer 1966; Knight 2004; Wittkopp et al. 2004, 2008; Battle et al. 2014; Glassberg et al. 2019; Hill et al. 2021). Suppose allele *A* shows a higher level of expression than allele *B* on their original backgrounds. If this is due to a *cis* effect, then the level of expression of *A* should still exceed that of *B* in *A/B* heterozygotes. However, if this is due to a *trans* effect from a factor closely linked to *A*, then both *A* and *B* should show similar expression levels in heterozygotes, as any *trans*-acting factors will operate equally on both. Using this approach, studies in yeast and mice showed that most (but not all) local eQTLs are due to *cis* effects (Doss et al. 2005; Ronald et al. 2005). This pattern was also seen using RNA-seq to score expression in human lymphoblastoid cell lines (Prickrell et al. 2010). Wittkopp et al. (2008) and Emerson et al. (2010) discuss separation of *cis* and *trans* effects when both impact ASE. As noted by Battle et al. (2014), one can take this analysis a step further and map SNPs in LD with causal sites that directly influence ASE, detecting **aseQTLs** (or **aseSNPs**). They did so by treating the ratio of the two transcript types in a heterozygote as a quantitative trait. They used this same logic to map **sQTLs** (or **sSNPs**) that influence the ratio of two different splicing product from a transcript. Unexpectedly, some of their detected sSNPs acted in *trans*.

Given that a focal transcript could be impacted by both local and distant factors, rather than provide a profile (or Manhattan) plot (Figure 20.1) for each of the thousands of transcripts, the typical representation of an eQTL experiment is in the form of a **transcriptome map** (Chesler et al. 2005). This is a two-dimensional plot, where each displayed point represents a significant association (which could also be color- or size-coded to represent the strength of the effect; e.g., Jiang et al. 2013). As shown in Figure 21.1, one axis gives the

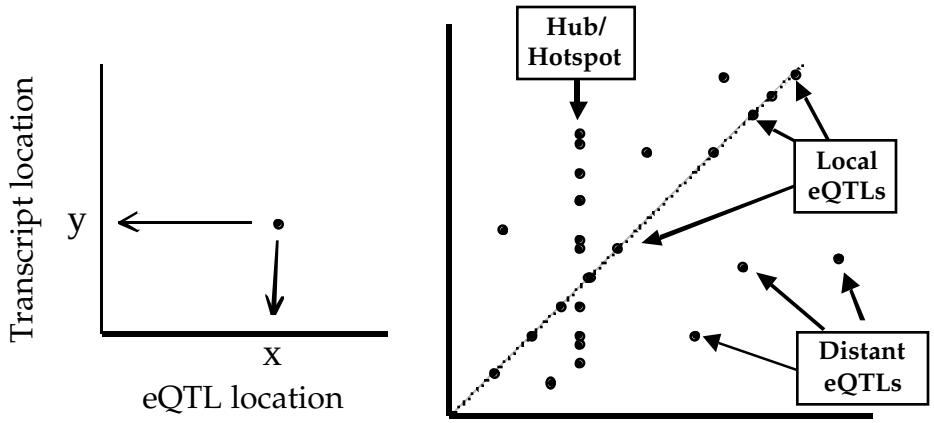

Figure 21.1 A stylized **transcriptome map**, plotting eQTL locations versus the location of the coding region for a transcript. Both axes correspond to genome position, with the horizontal (x) axis denoting a region/marker being tested as an eQTL and the vertical (y) axis the location of the coding region for a transcript (occasionally in the literature these two axes are reversed). A point or pixel at position (x, y) on this map indicates a significant association between a transcript whose coding region is at genomic position y and a marker/region at genomic position x. Points falling on the diagonal correspond to eQTLs that map very close to, or at, the same location as the coding region for the transcript they influence. These have been called ***cis* eQTLs**, but as discussed in the text are better referred to as **local** (**proximal**) **eQTLs**. Points falling off the diagonal correspond to eQTL locations that influence transcripts whose coding regions are at a different location from the eQTL. These have been called ***trans* eQTLs**, but are better referred to as **distant (distal) eQTLs**. A vertical stack of points correspond to a (small) genomic region that is enriched for eQTLs, and is called a **hotspot** or **hub**, with the eQTLs in that region impacting numerous transcripts.

genomic location of the eQTL/eSNP, while the other axis corresponds the genomic location of the transcript coding region that it impacts. If an eQTL maps close (local) to the coding region of the transcript it impacts, this generates a point on the diagonal of the plot, while if the eQTL maps distal, the point will be off the diagonal (Figure 21.1). When eQTL location is plotted on the horizontal axis (as in Figure 21.1), the presence of **hubs** or **hot spots** (distant eQTLs that impact numerous transcripts) will appear as vertical lines, while they appear as horizontal lines when the roles of axes are reversed (eQTL locations are mapped on the vertical axis). As observed by Lutz et al. (2019; 2022), the genetic architecture of variants at a hub can be rather complicated. In yeast crosses, *trans*-eQTLs mainly occur in hubs, a rather different situation from humans, perhaps reflecting coordinated control of a number of genes when the single-celled yeast experiences an environmental challenge. For example, the protein encoded by the yeast *IRA* gene, a key regulator in the RAS signaling pathway, is a *trans*-eQTL, impacting the level of expression at over a thousand transcripts. The coding region of this gene is over 9,000 bases long, and contains at least seven causal nonsynonymous variants displaying complex epistatic interactions with each other.

Bryois et al. (2014) noted multiple *cis*-eSNPs can impact the same coding region, which can be missed if the lead SNP has a large effect. Akin to removal of the effect of a major gene in standard QTL linkage analysis (Chapter 18), they suggested that the genotype of the first detected *cis*-eSNP be used as a cofactor to search for additional *cis* sites. Using this approach, they found that 20% of the genes in their modest sample of 870 human cell lines had least two detectable *cis*-eSNPs, a percentage that is certain to rise with larger sample sizes (Gusev et al. 2016). Any collection of such *cis* sites impacting the same transcript should be subsequently tested for epistatic interactions (Chapters 18 and 20).

A common observation (Rockman and Kruglyak 2006; Breitling et al. 2008; Albert and Kruglyak 2015) is that individual *cis* effects tend to be both larger, and more common,

that individual *trans* effects. There was considerable debate as to whether this represented reality or was simply a reflection of the low power of most early designs. The biological notion for stronger *cis* effects is that they are likely to be more direct, often focused on just a single transcript, than are *trans* effects, which are diffusive and may only become apparent after interaction with *cis* sites. There is clearly an ascertainment bias in favor of *cis* sites in that many studies only searched for associations with markers immediately adjacent to the transcript coding region. When power is low, the Beavis effect implies that any such sites declared to be significant will likely have an overestimated effect size (de Koning and Haley 2005). Note that tests of only transcript-adjacent (local) markers results in a much lower multiple comparison burden that testing every genomic location (distant) against the focal coding region. If indeed most *trans* sites have weaker effects, their heavier multiple comparison burden will further make then difficult to detect. Note from Example 21.1 that the yeast data suggest eQTLs of modest to weak effect are likely to be the norm, and that most of these are likely *trans*. Conversely, Stamatoyannopoulos (2004) noted a bias in diploids that can result in an *underestimation* of *cis* effect sizes, in that expression level measures incorporate values from both alleles (unless allele specific expression is scored), potentially diluting a strong *cis* effect from one of the alleles.

Given their weaker power of detection, Kendziorski and Wang (2006) suggested that one approach in a scan for *trans* sites is to test each marker separately, and look for accumulating evidence over the entire set of transcripts. Under the null hypothesis of no marker-expression effect, the distribution of p values follows a uniform, so a marker-specific histrogram of the p values for each transcript could be used to estimate the number of significant effects at each marker (e.g., Cotsapas et al. 2011). Appendix 6 discusses a number of these methods, but the complication to applying them to expression data is that we *expect* some transcripts to be highly correlated (Brynedal et al. 2017 suggest adjustments based on the decompositon of the transcript correlation matrix). This also raises concerns about the validly of apparent hubs (Figure 21.1), as Pérez-Enciso (2004) showed via simulation that the correlated structure of transcripts can easily create false positive (or *ghost*) hubs. A related issue is that subtle environment changes can result in expression shifts of a number of genes, potentially creating ghost hubs if individuals in the sample vary over these conditions (Stamatoyannopoulos 2004).

As mentioned, there is a huge disparity in testing local versus distal effects. Peterson et al. (2016a) noted that with their human date set of roughly 6.8 million SNPs and 30,100 transcripts, they had 142 million local tests (as an average of 21 genes were within a MB of a typical SNP), and over 200 billion distal tests. As a result, they suggested a hierarchical structure for hypothesis testing, first separating local and distal tests. For local tests, they focused on the entire collection of local SNPs for a given gene (using gene-based GWAS approaches; Chapter 20). Among those genes that showed a significant effect on a local transcript, they then tested each SNP within the gene separately. By first focusing on *sets* of SNPs, the multiple comparisons burden is greatly reduced. Similarly, for distal tests they first tested significance of a given SNP over the entire collection of distal transcripts (consided as a single set; again, modification of gene-based tests can be used), and, if significant, subsequently tested this SNP against each distal transcript.

Recalling the lessons from early GWAS studies (Chapter 20), wherein increasing sample sizes resulted in the discovery of more associated SNPs and greater replication over studies, we expect this same pattern applies to early eQTLs/eSNPs studies given their very small sample sizes. Even more recent studies are typically of rather modest relative to a modern GWAS, with samples sizes often no greater than one to a few thousand individuals (e.g., Zeller et al. 2010; Westra et al. 2013). As would be expected for eQTLs of small effect, Bryois et al. (2014) found that eQTL detection improved with increasing sample size. Indeed, Võsa et al. (2021) were able to detect *cis*-eQTLs for roughly 90% of the genes in a study of almost 32,000 individuals. Further, most of these eSNPs replicated over several tissues.

An important breakthrough in the analysis of eQTLs was the hypothesis of ***cis*-mediated *trans* effects**, or ***cis*-mediation** for short (Fehrmann et al. 2011; Jiang et al. 2013;

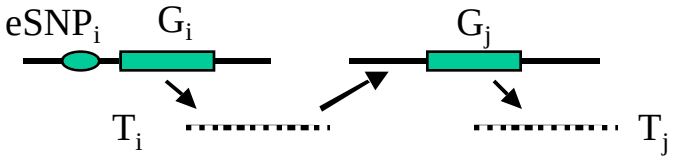

Figure 21.2 The concept of ***cis*-mediation**. The observation is that eSNP i acts at some distance away from a coding region (G_j) to regulate the level of its transcript T_j (eSNP i is a *trans*-eSNP for transcript j). The **mediation hypothesis** is that the impact of eSNP i is through a *cis* effect on the transcript from (local) gene G_i, whose transcript T_i then influences the regulation of transcript T_j of a distant gene G_j. Path analysis methods (Figure 21.3; Appendix 2) allow this idea to be extended over much more complex regulatory networks, as well as providing a framework for estimating direct and indirect effects of any component player (Example 21.3).

Battle et al. 2014; Bryois et al. 2014; Pierce et al. 2014). As shown in Figure 21.2, this idea postulates that many of the observed *trans* effects are the result of a *cis* effect at the eSNP that impacts the transcript of a local gene, with that transcript then having a *trans* effect on distant transcripts. Formally, suppose that SNP i is observed to have a *trans* effect on the transcript T_j from a distal coding region j, then the causality path is eSNP$_i \rightarrow T_i \rightarrow T_j$, in that SNP i *cis*-regulates the level of transcript T_i, which in turn influences transcript T_j. As an aside, the causalty notation $x \rightarrow y$ implies that a change in x while holding all other variables (other than y) constant results in a change in the distribution of y, but a change in y (holding all other variables constant) does not change the distribution of x.

The *cis*-mediation model explains, in part, why *trans* effects appear to be weaker that *cis* effects, as we are scoring a secondary effect. It also suggests strategies to improve *trans* detection. For example, Bryois et al. (2014) focused on only those eSNPs showing a local effect. These were then tested for *trans* effects over the rest of the transcriptome (an idea loosely akin to searching for epistasis by first starting with a QTL having a marginal effect; Chapter 18). This strategy significantly reduces the number of comparisons, resulting in improved power, and allowed Bryois et al. to detect addition *trans*-eSNPs. As outlined in Example 21.3, and discussed in more detail in Appendix 2, multiple regression and path-analytic models can be used to detect, and quantify, mediation effects. In humans, the observation is that *cis*-mediation is common, but usually is only partial (Example 21.3). One explanation is **mediator confounding** (Figure 21.3), wherein some unmeasured variable impacts both the *cis* mediator (T_i) and the *trans* transcript (T_j). This can arise when the eSNP used has imperfect LD with the causal SNP, or when measurement error impacts estimates of T_i and/or T_j (Pierce et al. 2014; Yang et al. 2017). Pierce et al. (2014) and Yao et al. (2017) found that many human *trans*-eQTLs are also *cis*-eQTLs for local genes. While partial mediation was common for many of these, complete mediation was rare.

What is the current big-picture view of the quantitative genetics of RNA expression levels? The bulk of work on eQTLs comes from two very distinct biological systems: multicellular humans and unicellular yeast (recent summaries in Albert and Kruglyak 2015, Albert et al. 2018, and GTEx Consortium 2020). Given the rather modest sample sizes of most studies, a substantial number of additional eQTLs, especially *trans*-eQTLs, are expected to be detected as sample size increases. Despite this limitation, the conclusion is inescapable that eQTLs are ubiquitous, and control of expression for almost all transcripts is polygenic, with *trans* effects generally contributing between two- to four-fold more variation than *cis* effects (Liu et al. 2019). In both humans and yeast, almost all major transcripts have associated *cis*-eQTLs. In humans, this is greater than 95% for protein-coding regions and 67% for lincRNAs (GTEx Consortium 2020), while most yeast protein-coding genes have detectable *cis*-eQTLs. Further, in humans allelic heterogeneity is common for *cis* sites, with multiple independent *cis* variants being the norm.

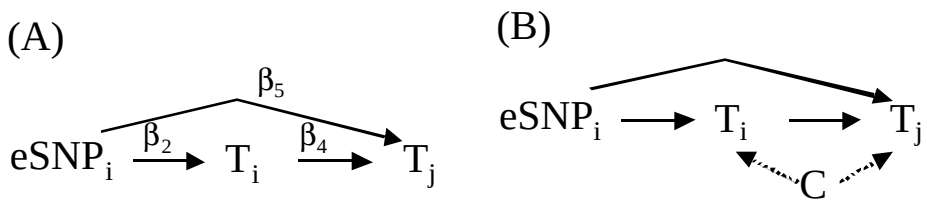

Figure 21.3 Path-analysis of a trio (eSNP$_i$, T_i, and T_j) to separate direct and indirect effects. **(A):** The path diagram when only these trio elements are involved (Example 21.3). The direct effect eSNP$_i$ $\rightarrow$ T_j (avoiding T_i) is given by β_5, and the indirect (**mediated**) effect via T_i, eSNP$_i$ $\rightarrow$ T_i $\rightarrow$ T_j, is $\beta_2 \cdot \beta_4$. The total effect is given by $\beta_1 = \beta_5 + \beta_2 \cdot \beta_4$. This same logic can be applied to a trio of an eSNP, a transcript it impacts, and a trait value (i.e., replacing T_j with z_k, the value for trait k), or other more complex regulatory pathways (e.g., Example A2.2). **(B):** Mediator confounding occurs, in the simplest case, when an unmeasured factor (C) impacts both T_i and T_j. In this setting, estimates of the direct and indirect effects can be biased.

The situation for *trans*-eQTLs is a bit more complex. These tend to have much weaker individual effects than *cis*-eQTLs, and are currently vastly undercounted. In humans, roughly one-third of detected *trans*-eQTL show some *cis*-mediation (GTEx Consortium 2020). Further, *cis*-eQTLs tend to be more **tissue-sharing**, while *trans*-eQTLs are more **tissue-specific**, indicating that even more are awaiting discovery as the tissue pool expands. Finally, *trans*-eQTLs are more enriched for known GWAS hits than are *cis*-eQTLs, which themselves are enriched relative to random frequency-matched SNPs. Hubs (*trans*-eQTLs that impact numerous transcripts) are seen, but generally impact only a modest number of transcripts. In contrast to humans, the vast majority (90%) of *trans*-eQTLS effects in yeast map to just around 100 hubs. The median number of transcripts impacted by a yeast hub is 425, with a range of 26 to 4600 (Albert et al. 2018). Four of these hubs impact over half of the genes showing variation in expression. In contrast to the highly-structured *trans* hubs of yeast, the hub structure of humans is far weaker and much more diffuse. Part of our current view of the human hub structure is likely impacted by low power. Indeed, highly structured hubs have been seen in other line-crossed eQTL mapping experiments with model organisms (mice, rats, *Arabidopsis*). In the line-cross setting, segregating alleles have equal frequencies (Chapter 18), as opposed to a GWAS setting where there is lower power to detect effects associated with rare alleles (Chapter 20). Despite these issues, at present it appears that humans and yeast may have rather different *trans* structures.

Example 21.3 As developed by a number of investigators, one can use conditional regressions (path analysis methods; Appendix 2) to both detect, and quantify, the amount of mediation that gene i has on transcript j (Chen et al 2007; Jiang et al. 2013; Pierce et al. 2014; Yang et al. 2017; Yao et al. 2017; Shan et al. 2019). This is done using a nested series of regressions to establish causality. Using the notation in Figures 21.2 and 21.3, first consider the association between the dosage of SNP i (N_i) and the transcript associated with coding region j (T_j),

$$T_j = \alpha_1 + \beta_1 N_i + e_1 \tag{21.1a}$$

One declares SNP i to be a *trans*-eSNP for coding region j when the slope β_1 is significant. This slope measures the **total effect** of SNP i on T_j, the contributions from both direct effects and indirect effects (such as through T_i). Next, we declare SNP i to be a *cis*-eSNP for coding region G_i when the regression

$$T_i = \alpha_2 + \beta_2 N_i + e_2 \tag{21.1b}$$

has a significant slope. Similarly, we declare the T_i has an effect on T_j when β_3 is significant

for the regression

$$T_j = \alpha_3 + \beta_3 T_i + e_3 \tag{21.1c}$$

Significant slopes in the above three regressions establish that (i) SNP i is associated with T_j; (ii) SNP i is associated with T_i, and (iii) T_i is associated with T_j. These univariate regressions, by themselves, do not separate direct from indirect effects. To do so, a multiple regression of T_j is constructed based on both N_i and T_i,

$$T_j = \alpha_4 + \beta_4 T_i + \beta_5 N_i + e_4 \tag{21.1c}$$

If $\beta_5 = 0$, then any effect from SNP i on T_j is simply through its effect on T_i, namely, **full mediation** (the effect of SNP i on T_j is entirely through its *cis*-effect on T_i). When both β_4 and β_5 are significant, then **partial mediation** occurs, where *both* T_i and SNP i (through a path independent of T_i) impact T_j. Note that this logic need not be restricted to just transcripts, one could measure (say) P_i, the level of protein from gene i, or some other regulatory measure such as methylation, splicing, etc. Modifications of permutation tests to accommodate the correlation structure of mediation analysis are discussed by Jiang et al. (2013) and T. Wang et al. (2020).

From the theory of path analysis (Appendix 2), the indirect effect of SNP i on T_j through the path given by T_i, is just the product of the path coefficients, which turns out to be $\beta_2 \cdot \beta_4$ from the above regressions. As shown in Figure 21.3A, the total path effect β_1 assumes the potential of a direct effect β_5 from eSNP$_i$ to T_j ($N_i \rightarrow T_j$) and an indirect of eSNP$_i$ via paths through T_j ($N_i \rightarrow T_i \rightarrow T_j$) with effect $\beta_2 \cdot \beta_4$. Hence, the proportion of the total effect on T_j from eSNP$_i$ mediated via T_i is

$$(\beta_1 - \beta_5)/\beta_1 = \beta_2\beta_4/\beta_1 \tag{21.1d}$$

If there are no unscored correlated factors that impact members of this trio, then the relation $\beta_1 = \beta_5 + \beta_2 \cdot \beta_4$, namely total effect = direct effect plus indirect effect, should hold. If it does not, one is likely missing correlated elements (confounders). Figure 21.3B shows one example. Such confounding could be caused by the focal eSNP$_i$ being in LD with different causal SNPs for the *cis* effect on T_i and the *trans* effect on T_j (Pierce et al. 2014).

Negative values of Equation 21.1d are commonly observed (e.g., Yang et al. 2017), indicating mediator confounding (Figure 21.3B). Even when one has a candidate lists of potential confounders to test (e.g., age, sex, environment risk factors, etc.), two variable selection issues arise. First, even assuming that the correct potential confounder variables have been identified, testing all of them in a single regression greatly lowers power. Second, and more problematic, confounder variables can vary over both trios (eSNP$_i$, T_i, T_j) and the tissues in which a trio is tested. Yang et al. (2017) developed **Genomic Mediation analysis with Adaptive Confounding adjustment** (**GMAC**) to address this concern, a trio-tissue specific variable selection approach to choose the appropriate confounder variables (from some candidate list).

The Epigenome and More General Regulatory QTLs (regQTLs)

The same logic used for mapping QTLs for protein- and transcript-level regulation can be applied to other regulatory features, such as methylation (Gibbs et al. 2010; McRae et al. 2018; Wu et al. 2018), DNAse I sensitivity (Degner et al. 2012), and mRNA splicing (Battle et al. 2014). Indeed, any genomic or cellular feature that can be quantified (assigned a numerical value, such as a binary 0/1 on-off score) can be treated as a quantitative trait. We refer to QTLs (or SNPs) associated with a scored regulatory feature as a **regulatory QTLs** (**regQTLs** or **regSNPs**). Table 21.1 lists a some of the different types of QTLs/SNPs that have been mapped using functional genomics features. Note that these classes are not exclusive, as an eQTL might be the result of regulatory variation at splicing (sQTL), DNAse I sensitivity (dsQTL), methylation (methQTL), or any number of other steps (Example 21.4). regQTLs represent a key step in the merging of quantitative genetics and functional genomics, a union we refer to as **quantitative genomics**, the natural extension of genetical genomics. The machinery of quantitative genomics provides a powerful analytic framework for extracting signals from the growing tsunami of functional data.

Table 21.1 A few of the different classes of QTLs (SNPs). The general terminology is to use QTLs generically, especially in a linkage-based analysis to indicate a region, and SNP in a GWAS setting to refer to a SNP showing an association. The QTL/SNP terminology is a bit idiosyncratic, with different versions for some of these abbreviations appearing in the literature.

acQTL/acSNP	Chromatin acetylation QTL/SNP
aseQTL/aseSNP	Allele-specific expression QTL/SNP
caQTL/caSNP	Chromatin accessibility QTL/SNP
cis-xQTL/*cis*-xSNP	*Cis* (local) QTL/SNP for feature x
dsQTL/dsSNP	DNAse I sensitivity QTL/SNP
eQTL/eSNP	RNA expression QTL/SNP
hQTL/hSNP	Histone QTL/SNP
haQTL/haSNP	Histone acetylation QTL/SNP
hmQTL/hmSNP	Histone methylation QTL/SNP
meQTL/meSNP	DNA methylation QTL/SNP
methQTL/methSNP	DNA methylation QTL/SNP
miR-QTL/miR-SNP	MicroRNA QTL/SNP
pQTL/pSNP	Protein expression QTL/SNP
pb-xQTL/pb-xSNP	Population-based QTL/SNP for feature x
QTN	Quantitative trait nucleotide
QTT	Quantitative trait transcript
rQTL/rSNP	Ribsome occupancy QTL/SNP
regQTL/regSNP	Regulatory QTL/SNP
sQTL/sSNP	Splicing QTL/SNP
sb-xQTL/sb-xSNP	Sex-based QTL/SNP for feature x
tQTL/tSNP	Trait QTL/SNP
trans-xQTL/*trans*-xSNP	*Trans* (distal) QTL/SNP for feature x
vQTL/vSNP	Variance QTL/SNP

As with eQTLs, regQTLs are conditional, potentially having tissue or developmental-state specificity. For example, even when using the same tissues, one could be scoring a functional feature under steady-state behavior in some settings and under the dynamic response following some environmental perturbation in others. Similarly, ascertaining the direction of causality for regQTLs remains a major problem. A phenotype could *cause* a change in some functional feature, rather than being the *result* of that feature. **Mendelian randomization** (Example 21.10; Appendix 2) offers one approach for assigning causality.

An especially interesting set of regulatory features is the **epigenome** (Susuki and Bird 2008), the structure of the chromatin (and associated binding proteins), as well as other features, that distinguish different tissues between cells with otherwise identical DNA sequences. Examples include methylation of DNA CpG sites, acetylation and methylation of histones, and DNAse I sensitivity regions (Strahl and Allis 2000; G. Wang et al. 2007; Chi et al. 2010; ENCODE Project Consortium 2012, 2020; Romanoski et al. 2015). A growing number of studies highlight the importance, and potential use, of the epigenome. For example, Rakyan et al. (2011b) examined differences in the DNA methylation patterns (the **methylome**) between monozygotic twins (MTZ) that were discordant for childhood-onset Type 1 Diabetes (T1D). Despite being genetically identical, in discordant twins, one sib displays the disease while the other does not (discordant MTZ are expected to be common unless the disease prevalence is high, Yang et al. 2010c; Visscher et al. 2012). Rakyan et al. were able to detect 132 CpG sites, differentially methylated between discordant pairs, that correlated with T1D status. Such an association does not imply causality, as cellular perturbations from T1D might *induce* these changes, rather than these changes *causing* T1D. Rakyan et al. then used an independent data set to show that many of these methylation differences could be seen *prior* to the expression of the disease in case individuals.

If environmentally induced methylation events are persistent, then (provided the correct tissue is chosen), one could (in theory) search the methylome for signatures of past environmental exposures. Human epigenetic markers can persist over a lifetime (Shah et

al. 2014). For example, Heijmans et al. (2008) found that, even after 60 years, Dutch exposed prenatally to famine during a food embargo near the end of World War Two still showed reduced methylation (in peripheral blood) at the insulin-like growth factor II gene relative to appropriate control groups. Breitling et al. (2011) found a single locus that was under-methylated (again, using blood) in smokers (relative to nonsmokers) in an initial sample of 177, and were able to replicate their finding using an independent sample of 316 individuals. Hence, in at least some cases the epigenome may display persistent signals of past environmental events and exposures. A genome-wide scan for epigenetic features associated with particular traits has been refered to as a **epigenome-wide assocation study**, or **EpWAS** (EWAS is also widely used in the literature, but we reserve this for expression-wide studies) or a **methylome-wide association study** (**MWAS**). Issues when conducting an EpWAS are reviewed by Rakyan et al. (2011a).

Example 21.4 Two interesting studies attempted to dissect the impact of *cis* variation at various stages in the regulatory pathway (from gene to transcript to protein) using a collection of lymphoblastoid cell lines extracted from 72 Nigerian Yorubas (Battle et al. 2015; Li et al. 2016). The use of cell lines allowed for the detailed measurement of a number of genomic and cellular features, while SNP genotyping of these lines allowed for a GWAS to be conducted for each scored feature. The GWAS was restricted to *cis* (local) QTL by restricting the search to a 20-kb window around the target gene. Battle et al. used RNA-seq to measure standing mRNA levels, ribosome profiling (**ribo-seq**) to measure translation (via occupancy rates, a measure of transitional rates and efficiency; Ingolia et al. 2009), and mass spectroscopy to measure the associated protein levels. Despite the very low power of their design (given their very small sample size), they still found numerous regulatory QTLs at all three steps: 4,400 genes were tested for protein abundance, detecting 300 *cis*-pQTLs; 15,000 genes were scored for ribosome occupancy, yielding 930 *cis*-rQTL; and 16,600 genes were scored for mRNA levels, finding 2400 *cis*-eQTLs. Among the 4300 genes scored for all three regulatory phenotypes, 66% of eQTLs had some overlap with both downstream rQTLs (66%) and pQTLs (35%), while 52% of rQTLs overlapped with their corresponding pQTLs. While the eQTL and rQTL effects (when both were present) tended to have a similar effect size, pQTL effects tended to be diminished relative to their upstream eQTL effect. Thus the signal from a change in mRNA expression tends to be at least somewhat buffered when it reaches the protein level. Finally, they detected a class of *cis*-QTLs that influenced protein levels with essentially no effect on mRNA levels (pQTLs that were not eQTLs), suggesting that these might arise via posttranslationa regulation.

Li et al. (2016) conducted a more granular analysis of using these same lines, adding several additional regulatory steps in addition to those measured by Battle et al. (2015). They examined potential regulatory steps upstream from the TSS by considering several markers for chromatin modification relating to transcription factor accessibility. The first was a putative enhancer signal, acetylation of the lysine at position 27 on Histone 3 (H3K27ac) around the TSS, which is associated with enhanced transcriptional activity (Wang et al. 2016). [Methylation of this lysine, H3K27me3, and of the lysine at position 4, H3K4me1 and H3K4me3, are associated with silencers, enhancers, and promoters, respectively.] They also scored methylation levels and DNAse I sensitivity. While RNA-seq measures the steady-state level of transcript, the actual *rate* of transcription was scored (**4sU-seq**) by using pulse-labeled uridine (s4U). Their RNA-seq analysis included isoform ratios for alternatively spliced products. Regulatory QTLs were detected at each of these steps. They found that around 65% of all eQTLs have effects on chromatin. They also found an uncoupling of expression level and splicing in that sQTLs and eQTLs tend to be independent (only 14 of 275 genes shared the same lead SNP for both). Further, eQTLS are enriched around the TSS, while sQTLs are enriched within the body of the coding sequence (in particular, within the introns that they regulate). While not influencing expression level, 90% of sQTLs generated variation in the final protein sequence. Finally, they noted that GWAS SNPs for four disease and two traits were enriched for both eQTLS (as noted above) *and* sSNPs, and that the latter appeared to have effect at least as great as the former. Hence, variation in alternative splicing rates may be an important player underlying quantitative variation.

What is the Nature of GWAS Noncoding SNPs?

One of the surprising observations from GWAS studies was that the vast majority of significant "hits" (SNPs/CNVs statistically associated with trait variation) fall into noncoding regions, and thus are not the result of structural changes (variation in the amino acid sequence of coded proteins). While, by default, these GWAS hits are likely regulatory in nature, can we associate noncoding hits with *specific* regulatory features? One could examine one or a few model traits for such associations, but these results might be somewhat idiosyncratic, and not that representative of traits in general. A much more powerful approach is to uses catalogues of noncoding SNP hits assembled over a wide array of traits and diseases to see if these SNPs are enriched for regulatory features (such **enrichment tests** will be discussed shortly). For example, Maurano et al. (2012) showed that a collection of over 5600 noncoding human SNPs for 200 different diseases and 450 different traits preferentially mapped to DNase I hypersensitive sites (**DHS**) as scored using the ENCODE collection of 85 cell-tissue types. DHS sites are a strong regulatory signal, indicating an open chromatin structure that allows the transcriptional machinery to have greater accessibility to genes. Trynka et al. (2013), Pickrell (2014), Farh et al. (2015), and Gusev et al. (2018) reported similar observations using other human chromatin markers.

An even more direct regulatory connection is the observation in humans that collections of GWAS SNP hits (involving numerous traits and diseases) are significantly enriched for eQTLs. This is especially remarkable in that most of the early studies scored eQTLs using just a single cell line/tissue type, yet the detected eSNPs were enriched for GWAS hits over a wide variety of traits (e.g., Nicolae et al. 2010; Fehrmann et al. 2011; Westra et al. 2013; Battle et al. 2014; Bryois et al. 2014; GTEx Consortium 2017, 2020; Yao et al. 2020; Barneira et al. 2021). One might imagine that focusing attention on only a single tissue (such as lympholbastoid cell lines) or peripheral blood (a mixture of a limited number of cell types) would only capture a small fraction of eQTLs over the whole organism, and yet even with this restriction, a strong enrichment emerged. Later studies used eQTLs detected over a variety of tissues, similarly showing strong enrichment for GWAS SNP hits over a variety of diseases and traits. For example, Gamazon et al. (2018) used a bank of 44 human tissues in 450 individual, finding that 60% of known trait-associate SNPs were in LD with *cis*-eQTLs detected over one (or more) of the cell lines. They classified detected eSNPs as either tissue-shared, or tissue-specific, eSNPs, finding that for most (but not all) traits, tissue-shared eSNPs (detected in more than one tissue) accounted for great proportion of trait associations than did tissue-specific eSNPs. Further, at least in humans, local eQTLs tend to be more tissue-shared, while distal eQTLs are often tissue-specific (GTEx Consortium 2017; Liu et al. 2017).

Example 21.5 An important cautionary tale on fine-mapping was offered by Smemo et al. (2014). A set of roughly 90 variants in very high LD that map within a 47 kb region spanning introns 1 and 2 of the *FTO* gene had very strong, and highly reproducible, GWAS hits for human obesity (measured by body mass index, BMI). Individuals homozygous for risk alleles averaged more than 3kg heavier than individuals homozygous for non-risk alleles. Deletion of *FTO* in mouse models results in leaner mice, while mice overexpressing *FTO* are heavier. Finally, this 47kb region is heavily enriched with *cis*-acting control factors (enhancers, repressors, DNAse I sensitivity sites, TF binding sites). However, none of the variants within this region map as eQTLs for *FTO* expression. Smemo et al. found that this regions is involved in chromatin looping to a region over a megabase away containing the gene *IRX3*. In a human EWAS using brain tissue, 11 of the *FTO* SNPs associated with BMI were also eSNPs for *IRX3*, but not *FTO*, expression. Further, of the eSNPs associated with *IRX3* expression in either brain or mature adipose tissue, only those expressed in the brain showed highly significant associations with BMI. Hence, the *FTO* GWAS hits appear to be distal eSNPs that impact expression levels of *IRX3* in the brain. The apparent colocalization of *FTO* GWAS hits and mouse knockout effects gave a misleading picture of how these specific causal sites influence human body mass. Further, focusing expression studies solely on one obvious target, mature adipose

tissue, would have missed this signal.

An independent study by Claussnitzer et al. (2015), using gene editing in human tissue cultures, offered a rather different finding, highlighting the subtleties of tissue choice. They found strong effects of a particular SNP variant (rs1421085) within this *FTO* region on the expression of *IRX3* and the nearby *IRX5* gene in *precursor* adipocyte cells, resulting in a switch from fat burning to fat storage. This variant disrupted a repressor within this region (*ARID5B*), resulting in the activation of a rather potent early adipocyte enhancer and a doubling of *IRX3* and *IRX5* expression early adipocyte differentiation. Thus, there appear to be potentially several different gene circuits (with different tissue specificity) influencing BMI from genes some distance from the location of the GWAS hits. The different, but not necessarily exclusive, conclusions from these two studies highlight the concern stressed by Barbeira et al. (2018) that researchers need to adopt a more **agnostic scanning** approach when assessing correlations between expression levels and trait values.

A variety of approaches have been proposed to functionally associate a SNP to a particular gene (**S2G**, for **SNP-to-gene**), such as whether the SNP is within an exon, distance to the TSS, fine-mapped *cis*-eQTL, promoter capture or assessable chromatin information. These approaches are all restricted to detecting relatively closing-acting *cis* regulatory effects. A number of approaches are reviewed by Gazal et al. (2022), who propose an optimal weighting for combining several S2G metrics into a single index.

TRANSCRIPTOME WIDE ASSOCIATION STUDIES (TWAS)

Quantitative Trait Transcripts, QTT

The logical complement to an eQTL study is to directly search for **quantitative trait transcripts**, **QTTs** (Passador-Gurgel et al. 2007), whose expression levels are correlated with trait values. At first blush, this idea seems very straightforward. Starting with a sample consisting of both expression and trait data, simple modifications of a variety of GWAS approaches (Chapter 20) can be used to search for QTTs. The resulting analysis is often called a **transcriptome wide association study** or **TWAS**. We note that there are two flavors of TWAS, based on either using *observed* transcript values (which we denote as an **oTWAS**) or using *predicted* transcript values (given some reference set of eQTLs), a **pTWAS**. A pTWAS tests association based on *genetically predicted* expression levels, while an oTWAS tests association using the *total expression* level (genetic plus environmental contributions). Most TWAS studies, especially in humans, are pTWAS, and indeed the common literature use of TWAS refers to this class of analysis. We examine pTWAS in detail shortly, framing our initial discussion by assuming we have observed transcript values in some appropriate tissue (or set of tissues) given our trait.

Regardless of how transcript values are obtained, a TWAS scores each primary transcript (as opposed to a GWAS scoring each SNP), with the resulting Manhattan plot showing expression level-trait association p values at each transcriptional coding location. One simply changes the predictor variables, replacing the GWAS discrete variable of gene dosage (minor allele copy number, N) with the continuous variable of expression level (T) of a given transcript (which can be extended to the constellation of splicing variants for a given coding region). For example, standard linear regressions can used for continuous traits, with

$$z_i = \alpha_k + \beta_k T_{i,k} + e_i \tag{21.2}$$

where z_i is the phenotype of individual i and $T_{i,k}$ is the transcript amount of gene k in i. As in Chapter 20, the simple intercept α_k can be extended to a set of fixed effects to account for any relevant cofactors. Similarly, with binary traits, one could use logistic regressions (Equation 20.3). The dimensionality of the TWAS test set is on the order of the number of transcripts (in the low tens-of-thousands), as opposed to the number of SNPs (often in the millions), and thus has a much lighter multiple-comparison burden (essentially a gene-based GWAS; Chapter 20). This burden is likely even smaller than the number of transcripts, due to their

high correlation structure (e.g., Peterson et al. 2016b).

Unlike a traditional GWAS, a significant transcript effect in a TWAS has immediate biological interpretation. In a SNP-focused GWAS, the sign of a significant SNP effect typically does not convey much of a biological meaning. However, a positive β_k from Equation 21.2 implies that upregulation of transcript T_k increases the trait value (or disease risk), while a negative β_k implies that downregulation of T_k increases the trait value (or risk). As noted by Gamazon et al. (2015), this has immediate therapeutic implications, as targeted downregulation of a gene is usually easier to achieve that targeted upregulation. Hence, genes whose transcripts have positive β_k are better potential targets of therapeutic intervention, as reducing the level of these transcripts would decreases disease risk.

Several important statistical considerations arise when moving from a GWAS to a TWAS. Like SNPs in LD, the predictor variables in a TWAS (transcripts) can be correlated, and one must account for their covariance structure. In the case of SNPs, LD generates these correlations, which do typically not extend outside of a block of tightly-linked sites. Trimming SNPs within such a block is often used reduce the impact from high SNP colinearity. With transcripts, however, there is no genomic locational impact on the correlation structure, as transcripts from unlinked genes can be very highly correlated. There are two other, rather subtle, complications when moving from SNPs to transcripts. First, while the SNP genotype of an individual does not change over the environment, *this is not the case with transcripts*. Different environmental exposures can impact which genes are expressed ($E \rightarrow T$), and also impact the trait, *independent* of transcript levels ($E \rightarrow z$). Such an environmental effect would be a confounder (Figure 21.3B), generating a false associate between transcript level and trait value. An appropriate choice of cofactors can often mitigate this effect.

The second (closely related) TWAS complication is that while the direction of GWAS causality between SNP and trait is always unidirectional (SNP $\rightarrow$ trait), such that an individual's trait value cannot change their SNP value, *causality can flow in both directions for a trait-transcript pair* ($T_k \rightarrow z$ and $z \rightarrow T_k$). Flow in the latter direction is often called **reverse causality**. When the SNP is correlated with both T and z, three base models are possible (Schadt et al. 2005): **causal** (SNP$\rightarrow T_k \rightarrow z$), **reactive** (SNP$\rightarrow z \rightarrow T_k$), and **independent** ($z \leftarrow$ SNP $\rightarrow T_k$). Consider weight. An environment feature (such as diet) may have a strong impact of trait value, independent of any underlying transcript values. Conversely, as weight increases, the expression levels of genes can change *as a result of the trait value*. The resulting trait-transcript association is driven by the *trait*, not the *transcript* (the reactive model).

The impact of high correlations among predictor variables differs between a GWAS and a TWAS. In a GWAS, a number of significant hits can be generated over a set of SNPs that are all in high LD with the same causal site. Fortunately, the locational clustering of these SNPs usually suggests this possibility. Conversely, a TWAS can have noncausal *transcripts* showing significant associations with a trait because of tight coregulation with a true causal transcript (for the trait). Unlike the case with SNPs, there is no obvious locational clustering to suggest that most are false positives (arising from correlations with a true positive). One potential approach for decoupling the effects of correlated transcripts is to construct the covariance matrix of all n_k transcripts correlated with a focal transcript T_k, and then extract the leading $m_k < n_k$ PCs from this matrix, and use these in a regression,

$$z_i = \alpha_k + \beta_k T_{i,k} + \sum_{j=1}^{m_k} \beta_{k,j} PC_{i,j,k} + e_i \tag{21.3}$$

Here $PC_{i,j,k}$ is the jth PC score for individual i of the covariance matrix for all transcripts associated with the focal transcript k. One could proceed in this manner over each member of the set of correlated transcript (using some threshold value for a minimum absolute correlation) to extract a subset of leading predictors, which could be then be simultaneously fit in a multiple regression. Penalized regressions, such as the LASSO (Example 20.4), could also be used for model selection. Given the potential of very high transcript colinearity, these

approaches are not fool-proof. Further, if there are additional, but unmeasured cofactors (such as important environmental factors that influence both trait and transcript values), false associations can arise. The above-mentioned methods of path and mediation analysis can be used to probe the causality structure (Appendix 2; Example A2.2), but such an analysis is only as good as the variable selection of possible confounders. As we will see next, the solution to this concern is to use *eQTLs as proxies for transcript values*, as changes in the trait value, or the value of other transcripts, will not change the eQTL genotype. This is the instrumental variable approach that Mendelian randomization (Appendix 2) uses to test the impact of some factor (such as high blood pressure) on a disease, free of the effect of confounders or reverse causality (Figure A2.7). If the factor has a direct effect on the trait, genotypes that influence that factor should also influence the downstream trait value.

Example 21.6 Kirst et al (2004) measured growth rate in clones of 91 backcross individuals from a *Eucalyptus* cross (F_1 offspring of *E. grandus* $\times$ *E. globulus* were backcrossed to a *E. grandus* parent). This set of clones was also scored for expression levels at 2600 genes via a microarray. After adjusting for multiple comparisons, 26 transcripts were significantly correlated with growth, with an addition 11 transcripts added using a slightly less stringent threshold. All of these transcript levels were negative correlated with growth (lower mRNAs levels were seen in faster growing individuals). A single transcript in the lignin (a major wood component) biosynthetic pathway explained 38% of the growth variation, and the majority of the other significant transcripts coded for enzymes in this pathway. An independent analysis found lower amounts of lignin in the faster-growing clones. All, save one, of the significant transcripts associated the lignin pathway were influenced by a single *trans*-acting eQTL on linkage group 9. A second *trans*-eQTL on linkage group 4 also influenced the majority of these transcripts. These two eQTLs colocalized with two QTLs independently mapped for growth (however, confidence intervals for QTL/eQTL locations were very large due to small sample sizes).

TWAS Using SNP-predicted Transcript Values (pTWAS)

The term TWAS as it commonly appears in the human literature refers to a GWAS setting where one has SNP and trait, but no expression, data. Hence, to be consistent with the more common use, in what follows we use TWAS in place of pTWAS. The idea is to leverage the SNP-trait data (without performing any additional genomics) by using the eQTLs previously detected in some **reference transcriptome** population, and then (based on the observed SNP genotypes of an individual) predict the expected values for their unmeasured transcripts (Gamazon et al. 2015; Gusev et al. 2016). In essence, this is a form of *imputation* (Chapter 20). Instead of using the observed SNP genotypes and some larger SNP reference collection to infer *unscored genotypes*, one instead uses the observed SNP genotypes and a reference collection of eQTLs, and then infers the *unscored transcript* values (Schadt et al. 2012 discuss the converse issue of predicting an eSNP genotype from an observed transcript value). For example, in humans one could use eQTLs detected using the **Genotype-Tissue Expression Project** (**GTEx**) set of roughly 50 different tissues from 900 individuals (GTEx Consortium 2017, 2020) or the **Genetic European Variation in Health and Disease** (**GEUVADIS**) set of 460 lymphoblastiod cell lines (Lappalainen et al. 2013). As noted by Huang et al. (2018), one concern is that eSNP effects are overstaimted when power is low,, and suggest using bootstrap appraoches (CHapter 18) to obtain less biased estimates. One concern is that eSNP effects are overestimated when power is low, and Huang et al. (2018 suggested using bootstrap approaches (Chapter 18) to obtain less biased estimates. Wainberg et al. (2019) presents an overview of TWAS advantages, limitations, and best practices.

The massive leverage of a TWAS occurs because once a model for predicting transcript values is built, it can be applied to *any* GWAS study lacking expression data, *provided* two strong assumptions hold. First, that the LD patterns are very similar in the GWAS sample

and the transcriptome reference populations. Note that this is the same general restriction as with imputation of missing SNP. Second, that the appropriate tissue type(s) are used as the reference for the trait of interest, or expressions levels in the target tissues have a high level of genetic correlation with in the proxy tissue(s) scored. Provided that these assumptions hold, one can take previous (and future) GWAS studies and reanalyze each using TWAS. The resulting shift from testing all the SNPs to testing just the (predicted) transcripts is akin to moving from a SNP-based, to a gene-based, GWAS (Chapter 20), with a huge reduction in the multiple comparisons burden, and, as a result, a potential increase in power. However, any power gain from reduction in the burden of multiple tests could be more than offset by prediction inaccuracies of the expression levels (Huang et al. 2018). It is also worth noting that while we frame TWAS in terms of using local eSNPs to predict the impact of transcript levels on a trait, the exact same approach would be used for other regulatory features. For example, prediction the methylation levels from local meSNPs and then examining the impact of methylation site status on the trait.

The starting point for a TWAS is to first map eQTLs using the reference transcriptome data in the desired tissue(s), using SNPs proximal to the coding region in a statistical model to predict its transcript level (i.e., *cis*-eSNPs). Let $N_{i,k,j}$ be the number of reference alleles (typically the minor allele count, 0, 1, or 2 at each SNP) for individual i in the reference sample at a proximal SNP $1 \leq j \leq n_k$ for the coding region of transcript k. One then fits the linear model

$$T_{i,k} = \alpha_k + \sum_{j=1}^{n_k} \beta_{k,j} N_{i,k,j} + e_{ik} \tag{21.4a}$$

using a variety of approaches, such as using the **lead** (or **sentinel**) **SNP** (that with the largest effect), or more generally penalized regression/model selection approaches such as LASSO (Example 20.4). Distal eSNPs are generally not fitted because of model instability issues given their generally weaker effects coupled with the massive increased in model dimensionality. This statistical justification also has some biological justification, in that local eQTLs tend to be tissue-sharing, while distal eSNPs are often more tissue-specific (GTEx Consortium 2017; Liu et al. 2017; Urbut et al. 2019). Hence, one can use a proxy tissue when the real focal tissue (or tissues) for a trait/disease is unknown, as the proxy may often capture much of the local eQTLs (but could easily miss important distal eQTLs). For example, Qi et al. (2018) examined the correlation of both *cis*-eSNPS and *cis*-meSNPs effects between blood and brain tissue. Blood is far easier to obtain and score for genomic features than brain tissue, and thus much larger sample sizes can be used for the regulatory reference panels (in their study, 500 to 1000 for brain tissues versus 2000 to 14,000 for blood). They found a high correlation in effect sizes between the two tissue types, 0.70 for *cis*-eSNPs and 0.78 for *cis*-meSNPs. Hence, the reduction in accuracy in using blood for brain expression is more than offset by far greater reference population sample sizes. Ongen et al. (2017) discussed how to leverage information from an expression panel of different tissues (e.g., GTEx) to determine the causal tissue(s) for a given trait.

Letting n_{k*} denote the number of proximal SNPs retained in the final model for transcript k, the predicted value of transcript k in individual i from the GWAS (with no expression data) becomes

$$\widehat{T}_{i,k} = \alpha_k + \sum_{j=1}^{n_{k*}} \beta_{k,j} N_{i,k,j} \tag{21.4b}$$

with this value then substituted into an oTWAS ($\widehat{T}_{i,k}$ replaces $T_{i,j}$ in Equation 21.3). Gamazon et al. (2015) called this approach **PrediXcan**. More generally, the $\beta_{k,j}$ weights can vary over tissue type. For tissue type h, one can compactly write the vector of n_T predicted transcript values for individual i, $\mathbf{t}_i^{(h)}$, as the product of an tissue-specific $n_s \times n_T$ weight matrix, $\mathbf{W}^{(h)}$, and an n_s-dimensional column vector of SNP genotype scores (the $N_{i,k,j}$) for individual i, $\mathbf{g}_i$, e.g., $\mathbf{t}_i^{(h)} = \mathbf{W}^{(h)}\mathbf{g}_i$. Because of the use of only local SNPs in the weight matrix, it is very sparse. The power of the TWAS approach is that $\mathbf{W}^{(h)}$ is estimated just

once from the transcriptome reference populations, and can then be applied to all relevant GWAS studies (different sets of $\mathbf{g}_i$ vectors). Ideally, a separate TWAS is performed over each of the tissue-specific matrices. Xu et al. (2017) discussed the connections between TWAS and multiple-SNP gene-based GWAS tests (Chapter 20).

There are two shortcomings with this basic TWAS approach that were addressed by subsequent investigators. The first is that any uncertainties in the estimates of $\beta_{j,k}$ are not incorporated into the resulting TWAS (the $\widehat{T}_{i,k}$ values are assumed to be predicted without error). This limitation was addressed by the **collaborative mixed model** (**CoMM**) approach Yang et al (2018) and Yeung et al. (2019), which jointly estimates the effects of a SNP on transcript levels in the transcriptome reference set and its effects on the focal trait in the GWAS. This allows for uncertainty in the predicted transcript levels to be accommodated for in the final TWAS. The cost of this approach is that $\beta_{j,k}$ values must be (jointly) estimated for each new GWAS, as compared to a single estimation that can be used on all future GWAS (using their $\mathbf{g}_i$ values). The second shortcoming is that Equation 21.4b assumes one has access to the full GWAS data set (in particular, the genotypes of each individual). More generally, one may only have access to the summary statistics, not the individual genotype data. As outlined in Example 21.7, a number of investigators have extended TWAS to the summary data setting.

While using predicted transcripts values might appear to be less optimal than using their observed values, as noted by Gamazon et al. (2015) they have two important advantages. By predicting transcription levels solely on the basis of the genotype at the eSNP, what Gamazon et al. (2015) refer to as the **genetically regulated expression** component (**GReX**) of the transcript (eSNP $\rightarrow T_k \rightarrow z$), any contribution from trait feedback ($z \rightarrow T_k$), which is included in the observed transcript value, is not in the TWAS predicted value. Similarly, by conditioning on the eSNP, any environmental contribution is also removed. Note that only common local eSNPs appear in the weights in Equation 21.4b, ignoring any genetic contribution from rare local eSNPs or distal eSNPs. TWAS is thus a test of a nonzero genetic correlation between local control of expression of a target transcript and a focal trait. As mentioned, TWAS is an application of the method of Mendelian randomization (MR), testing for the effect of a factor (transcript T_k) on an outcome (trait value z), where the use of an instrumental variable (*cis*-eSNPs) that predicts T_j controls for bias from reverse causality and confounders (factors influencing both the transcript and trait, but not the genotype). Appendix 2 examines the MR approach in more detail.

Using the TWAS approach, Gusev et al. (2016) leveraged existing GWAS studies to discover 70 new genes associated with obesity related traits (using expression in blood and adipose tissues as the reference transcriptomes). Their simulation studies showed that when transcript level was due to multiple causal local eSNPs, TWAS offered greater power than either a standard GWAS or an **eGWAS** (a GWAS with increased power by only testing for associations using significant eSNPs).

Unfortunately, the TWAS approach based upon (predicted) genetically control transcript levels can also generate false positives in many settings (Mancuso et al. 2019; Wainberg et al. 2019). For example, suppose local eSNP$_j$ impacts both transcripts T_i and T_k (**pleiotropy**), but only T_k impacts trait value ($T_i \leftarrow$ eSNP$_j \rightarrow T_k \rightarrow z$). Gene i would still be declared as significant in a TWAS. An alternative setting (**linkage**) occurs when distinct SNPs (j and ℓ) impact the two different transcripts, but these SNPs are in high LD ($T_i \leftarrow$ eSNP$_\ell \leftrightarrow$ eSNP$_j \rightarrow T_k \rightarrow z$). Again, gene i would be declared significant under a TWAS. The **FOCUS** (**fine-mapping of causal gene sets**) method of Mancuso et al. (2019) attempts to removed correlations due to LD among eSNPs from different transcripts. Wainberg et al. (2019) suggested that best practices are to followup significant TWAS results by using colocalization methods to control for linkage (to be discussed shortly).

Example 21.7 A common setting is that one has access to just the summary statistics from

a trait GWAS, rather than the full SNP genotypes of all the study individuals (i.e., we do not know $N_{i,k,j}$). Similarly, the amalgamation of a number of individual GWAS into a single meta-populaion study usually returns just summary statistics for each SNP (Chapter 20). As shown by a number of investigators, one can still use the weights $\beta_{k,j}$ from Equation 21.4a on the local SNPs around a target transcript T_k to perform a TWAS. For example, the summary statistics extension of PrediXcan (**S-PrediXcan**) by Barbeira et al. (2018) proceeds as follows. As in Equation 21.4a, consider $1 \leq j \leq n_{k*}$ local SNPs around a transcript k. From the GWAS summary statistics, we have the estimated effects for these SNPs on the focal trait, $b_{k,j}$, and their sampling variances, $\sigma^2(b_{k,j})$. From the reference transcriptome, one can estimate the variance of transcript k, $\sigma^2(T_k)$, and either from the GWAS, or an equivalent reference population, one can estimate the variance in reference copy number $\sigma^2(N_{k,j})$ for each SNP, which is function of their population frequencies. An approximate z score for a TWAS for the coding region associated with transcript T_k is then given by

$$z_k \simeq \sum_{j=1}^{n_{k*}} \beta_{k,j} \left(\frac{\sigma(N_{k,j})}{\sigma(T_k)} \right) \left(\frac{b_{k,j}}{\sigma(b_{k,j})} \right) \tag{21.5}$$

where $\beta_{k,j}$ are the regression prediction weights from Equation 21.4a. See Barbeira et al. (2018) for a derivation. A number of such summary statistics-based TWAS approaches have been proposed, such as by Mancuso et al. (2017) testing for a genetic correlation between expression and the trait, the **Summary TWAS** (**S-TWAS**) of Gusev et al. (2016), and the **summary statistic collaborative mixed model** (**CoMM-S**^{2}) of Yang et al. (2020). The **UTMOST** (**unified test for molecular signatures**) test of Hu et al. (2019) provides a single unified approach for considering all tissues (in the reference set) simultaneously.

STATISTICAL APPROACHES FOR FINE-MAPPING CAUSAL VARIANTS

The high level of LD among blocks of variants that powers a GWAS is also the major impediment to unambiguously declaring a particular variant (or set of variants) to be causal. Depending on the structure of the study population, an LD block could be less than a kb (wild maize), tens of kb (humans, with Europeans having longer blocks than Africans), or over hundreds of kb (association panels of elite cultivars; Buckler and Gore 2007). A shorter LD block does not necessarily mean fewer candidate variants, as the nature of the evolutionary forces generating LD is such that populations with shorter LD blocks also tend to have higher levels of variation (WL Chapters 3 and 8). A further complication is added by recalling Example 20.1: the strength (r^2) of LD is a ***function of the allele-frequency matching between a marker and a causal site***. A nearby marker can actually have a *lower* r^2 value than a more distant marker. When attention is focused to an **association region** tagged with confidence in a GWAS (also called a **risk region** when mapping a disease gene), one is still left with tens to thousands of potential candidate variants. Within this region, p values likely do *not* monotonically become increasingly significant as one approaches the causal site. The tagged region could also contain multiple causal variants (such as a collection of rare variants), further complicating the use of spatial distribution of p values as an aid for fine mapping of the causal site(s).

Even after sequencing the entire association block (so that *all* variants, including those that are causal, are scored), the **lead**, or **index**, **SNP** (that displaying the most significant p value) within that block is ***likely not the causal variant***, especially when power is low and LD is extreme (Ledur et al. 2010; Udler et al. 2010; van de Bunt et al. 2015; Wu et al. 2017; Huang et al. 2018; Schaid et al. 2018). Simulations by van de Bunt et al (2015) assuming **whole-genome sequencing** (**WGS**) data still found that the lead SNP corresponded to the causal SNP only 80% of the time when the allele had high frequency and a strong effect (p = 0.5, OR = 1.5), and less than 3% of the time when the allele was less common and of modest effect (p = 0.05, OR = 1.1). Hence, even with WGS data and a large population sample, determining the causal variants is far from trivial. The term **QTN** (**quantitative**

trait nucleotide) has been used to declare a clear demonstration of a causal SNP (or some other variant, such as a CNV), but this has been very challenging to accomplish in most settings (see Example 21.8 for some exceptions).

Simulations by Wu et al. (2017) highlighted the impact of marker-causal allele frequency mismatch on fine mapping. They examined the distance between the lead SNP and the actual causal site under a variety of assumptions and genotyping schemes (WGS and imputation using different reference sets). Their simulations were based on WGS data of 3600 individuals with roughly 18 million (retained) genetic variants. Randomly-drawn common (minor allele frequency, MAF, > 0.01) and rare (MAF ≤ 0.01) variants were chosen to be causal, with an effect size ranging from 0.2% to 3% of the trait variance. GWAS was them performed using this underlying data and different sets of variants (imputed sets and WGS). When the causal allele was common, 95% of the lead SNP–causal site pairs had an MAF difference of < 0.05. For rare alleles, 95% had an MAF difference of ≤ 0.003. Hence, rare alleles are not tagged by common alleles (the LD, r^2, is just too small). In terms of mapping precision, under WGS, 80% of the lead SNPs for common causal variants were within roughly 10kb from the causal site. With imputation (instead of WGS; Chapter 20), this distance increased to between 25 and 34 kb (depending on the reference population used). Hence, for common alleles, the mapping resolution using WGS data was only marginally better than using imputed data, and is usually not a cost-effective approach. In contrast, for rare variants, almost 95% of the lead SNPs were with 5kb of the causal variant using WGS, but only 37% were this close using imputed data. Even assuming WGS data, mapping precision was a function of the causal allele frequency. With very rare causal variants ($p < 0.001$), 98% of the lead SNPs corresponded to causal sites, but this decreased to 30-40% for common causal alleles (MAF > 0.1).

Historically, the hope was that a GWAS would limit causal variants to a region in which obvious nonsynonymous variants were segregating, with these structural changes in the protein sequences of causal genes driving the majority of trait variation and disease risk. Alas, as we have detailed above, the vast majority of tagged SNPs occur in either noncoding (intergenic) or intronic regions. Regulatory, rather than structural, variation appears to underlie most quantitative trait variation (although protein variation can be generated via the regulatory effects of sSNPs). This has enormous implications for determining a potential set of causal variants within a GWAS region. While we have a fairly good understanding of the impacts of amino acid changes (e.g., Cooper and Shendure 2011), this is not the case for regulatory variation. First, the functional annotation of regulatory sites, especially beyond the TSS and intron-exon boundaries, is still an evolving enterprise. Second, for *trans*-acting regulatory variants, genomic location provides little insight as to the target they influence (e.g., Example 21.5). As noted by Spain and Barrett (2015), the "physical distance of a variant to a gene is not substantive evidence of causality." Indeed, in a schizophrenia GWAS, Gusev et al. (2018) found that the gene closest the lead SNP was the eventually implicated gene less than a quarter of the time.

Given these concerns, a number of statistical approaches have been developed to prioritize a smaller set of variants from a tagged region for any future functional studies, such as gene editing in model organisms. These fall into two categories: **agnostic approaches** that simply use the correlation (LD) structure among scored variants along with their marginal association statistics, and **annotation-driven** approaches that attempt to leverage additional functional information. The later could simply be some prior probability of a given class of change having some functional effect (such as a variant in a known promoter box or splicing junction). An important class of annotation-driven approaches searches for colocalizations between a GWAS-detected SNP and a regSNP (such as eQTLs, sQTLs, meQTLs, haQTLs, etc.; Table 21.1). We examine these different approaches in turn. Statistical fine-mapping methods are reviewed by Spain and Barrett (2015), Cannon and Mohlke (2018), Schaid et al. (2018), Sieberts and Schadt (2019), Cano-Gamez and Trynka (2020), and Hutchinson et al. (2020a). While we frame the following discussion on fine mapping in term of SNPs, the logic applies equally well to other classes of variants (e.g., CNVs).

Selecting Variants and Exploiting Local LD

Suppose a GWAS has been performed, and a number of association regions with genome-wide level significant SNP-trait associations have been detected (e.g., Figure 20.1). The next step is to break these regions down further into blocks of SNPs that are highly correlated (in high LD) with each other, but weakly correlated over blocks. Within each such block, there is a lead SNP (or SNPs) and a collection of tightly-linked SNPs in high to very high LD with the lead (e.g., Figure 20.2). The goal is to obtain some minimal **causal set** of SNPs that jointly has a high probability of including the causal SNP (or SNPs), while still containing as few members as possible. In an ideal setting, this set has a membership of one (e.g., Example 21.8).

The simplest analysis, often called the **heuristic approach**, is to consider all SNPs within some (arbitrary) correlation threshold of the lead SNP. Under the heuristic framework, an investigator typically ranks the importance of SNP candidates by their p values, with the largest viewed as the marker most likely to be closest to the causal variant. There are three reasons why this is incorrect. First, as mentioned, a more distal marker can often have a larger r^2 with the causal site than a more proximal marker if the allele frequency match is closer (Example 20.1). Second, p values are a function of the standardized effect, e.g., $\widehat{\beta}/\sigma(\widehat{\beta})$. A marker could have a larger actual effect ($\widehat{\beta}$) but also a larger standard error (e.g., due to having a low-frequency minor allele), resulting in a smaller standardized effect. Finally, there is always statistical noise in the realization of the underlying expected value.

A more formal approach is to use **regression and model selection**: starting with some initial set of SNPs within the association region, a stepwise, multiple, or penalized regression is used to extract the most impactful SNPs after accounting for their LD structure. If the number of SNPs in the region is small, one might fit them all in a multiple regression, but this fails when LD is very high. A more common approach is to use a forward stepwise regression: The best fitting SNP is added to a regression model as a cofactor, and then the next best SNP is selected, and so on until the increase in model fit is no long significant (e.g., Yang et al. 2012). Penalized regressions, such as LASSO (Example 20.4), offer a very flexible model-selection approach, as (depending on the penalty function used; Example 20.4), most SNP effects are shrink to zero. SNPs remaining in the final model form the reduced set of candidate causal sites. Regression and model selection approaches provide a formal framework to make full use of the LD structure (the pairwise correlations among all SNPs in the starting set). When using summary data, typically LD estimates from some reference population are used, which can introduce bias if the LD in the reference and study populations have different structures. While regression-based methods are better than heuristic approaches, if SNPs are very highly correlated, even penalized models can be problematic (Schaid et al. 2018). Suppose that two SNPs are almost entirely correlated. A model-selection approach will reject one of the SNPs, which, by chance, could be the causal one (Hormozdiari et al. 2014).

The final, and best, class of methods are Bayesian (also called **probabilistic methods**), based on simple applications of Bayes theorem (Equation 3.3b). They return **posterior inclusion probabilities** (**PIPs**) for each SNP, allowing for the construction of **Bayesian credible sets** (Example 21.8). The rank of SNPs using PIP is often rather different from their p-value ranks (e.g., Maller et al. 2012). Variants not included in the credible sets can be excluded as being causal, allowing the investigator to focus on a smaller set of candidates in more detailed followup studies. Bayesian methods can use the full SNP correlation structure and can incorporate additional information in their priors, such as differential weighting for different variants (e.g., promoter mutations given higher weight than variants in random intergenic regions). Example 21.9 outlines their basic structure. Another tool, suggested by Udler et al. (2010), is to use information from two (or more) different populations, an approach called **trans-ethnic mapping** in humans. The logic is that if the same causal variants are segregating in all populations, but under different LD structures, a more precise signal can be generated by a combined analysis (van de Bunt et al 2015; Schaid et al. 2018).

Finally, Wang et al. (2020) noted a problem with just reporting *marker* PIP values.

Suppose there are two causal sites within a focal region, sites 1 and 3. Further, suppose that site 2 is in full LD with site 1 ($r^2 = 1$, so that the marker effect sizes using a single-marker regession are equal, $b_2 = b_1$). Likewise, assume noncausal site 4 is in full LD with site 3 ($b_3 = b_4$). Under a model allowing two causal sites, the PIPs for each of these four sites would 0.5, where in reality a more informative metric would be based on *sets* of markers, e.g.,

$$\Pr[(b_1 \neq 0 \text{ or } b_2 \neq 0)] \quad \text{and} \quad \Pr[(b_3 \neq 0 \text{ or } b_4 \neq 0)]$$

This altered credible site would return the PIPs that at least one *set* member is causative (a set, as opposed to single marker, metric) and provides a more informative view of the biological situation. Wang et al. developed their **sum of single effects** (**SuSiE**) model to accomplish this goal of constructing better credible sets by adding contributions from a regression assuming a single marker effect and then computing a posterior distribution over these (see their paper for details).

Example 21.8 Maller et al. (2012) performed a followup investigation on 14 associated regions detected in the initial WTCC study (Chapter 20). They genotyped 5500 SNPs in 8000 individuals across 12 genomic regions for 3 diseases (type 2 diabetes [T2D], coronary artery disease [CAD], and Graves' disease [GD]). Under the (strong) assumption that each association region contained *exactly* one causal variant, they showed that the PIP for SNP i among the candidate set of m SNPs within a region is simply

$$\Pr(\text{SNP}_i \text{ is causal}) = \text{PIP}_i = \text{BF}_i / \sum_{k=1}^{m} \text{BF}_k \tag{21.6a}$$

where BF_i is the Bayes factor for SNP i (Chapter 20 and Appendix 7). [An alternative expression for Equation 21.6a was given by Udler et al. (2010), replacing the BFs by likelihoods, which assumes a prior that all SNPs have an equal chance of being causal.] Wakefield's (2008) approximate Bayes factors are often used, which are given as follows. Let $\widehat{\beta}_i$ be the (regression slope) estimate of the effect of SNP i, with sample variance σ_i^2. Letting $s_i = \widehat{\beta}_i/\sigma_i$ be the standardized effect for SNP i, with a β prior that is normal with variance σ_b^2, then

$$\text{BF}_i = \frac{\sqrt{1-\gamma_i}}{\exp\left[-\gamma_i s_i^2/2\right]} \qquad \text{where} \quad \gamma_i = \sigma_b^2/(\sigma_i^2 + \sigma_b^2) \tag{21.6b}$$

Using the set of PIP values for a given association region, Udler et al. formed 95% (99%) credible sets by choosing the smallest number of SNPs whose PIP values sum to 0.95 (0.99). For 3 of the 14 regions (2 for T2D, one for GD), a single SNP accounted for most of the PIP (>70%), while in four other regions, the number of SNPs in the credible set was small, excluding most of the SNPs from further evaluation. In the remaining 7 regions, the credible sets were large (> 70 SNPs). However, in two of these regions, a few SNPs had a PIP > 20%.

A more powerful analysis was offered by Huang et al. (2017), who examined two subtype of inflammatory bowel (IB) disease: ulcerative colitis and Crohn's disease (we use IB instead of the medical literature abbreviation IBD, reserving the latter to denote identical by descent). Roughly 200 IB loci have been mapped using GWAS, and the authors fined-mapped 94 of these using high density genotyping in 68,000 individuals (roughly 34,000 cases and controls; with the cases consisting of roughly 19,000 Crohn's and 15,000 ulcerative colitis). With these sample sizes, even very tightly linked markers in high LD could be somewhat decoupled. Several of the original 94 regions could be further broken into two or more independent signals, resulting in a total of 139 independent association regions. Three different Bayesian methods were used to construct credible sets, and only elements present in at least two of these were placed in the final set. The resulting set sizes ranged from 1 to over 400 variants. For 18 of the regions, the 95% credible set had just a single member (**single variant credible set**), 24 others had 2–5 variants in the set, and 27 associations had a SNP with a PIP > 0.5. Hence, with sufficiently large sample sizes, very high confidence of causality can be assigned

to a single variant (Udler et al. 2010 presented power calculations for the required sample sizes to exclude a noncausal SNP in high LD). The vast majority of IB signals were associated with both disease subtypes. While several of the regions were enriched for known regulatory signals (e.g., regions with modified histones H3K4me1, H3K4me3, and H3K27ac, indicative of active chromatin), variants with PIP > 0.5 in 21 noncoding regions could not be associated with known regulatory motifs.

Hutchinson et al. (2020b; also Wallace 2013) noted that Beavis effects (inflated effect size estimates for many of the top SNPs) introduced a slight bias on the construction of credible sets, tending to include more members than are needed (in moderate to high power settings). They propose an **adjusted credible set** method to reduce this bias.

Example 21.9 Bayesian methods (Appendix 7) return a posterior probability of inclusion (the probability, given the model assumptions, that a particular SNP is causal). They do so by using Bayes' theorem (Equation 3.3b),

$$\Pr(\text{SNP}_i \text{ is causal}) = \text{PIP}_i = \Pr(c_i = 1 \,|\, D) = \Pr(D \,|\, c_i = 1)\Pr(c_i = 1)/\Pr(D)$$

where c_i is a zero-one indicator variable equaling one when SNP i is causal and zero otherwise. The data, D, consists of the vector $\mathbf{s}$ of standardized marginal association statistics for the k SNPs being considered in the association region and the $k \times k$ correlation matrix $\mathbf{C}$ between these SNPs. There are numerous way to implement this core idea, depending on assumptions about the number of causal variants within a region, their prior distributions (e.g., using functional information to weight variants), different computational approaches, and so on. Commonly cited methods include **BIMBAM** (Guan and Stephens 2008); **BVSR** (Guan and Stephens 2011); **CAVIAR** (Hormozdiari et al. 2014, 2015); **PAINTOR** (Kichaev et al. 2014, 2016); **SSMR** (Wen 2014); **CAVIARBF** (Chen et al. 2015); **FINEMAP** (Benner et al. 2016); **DAP-G** (Wen et al. 2016); and **SuSIE** (Wang et al. 2020).

To illustrate the basic logic, we outline the **CAusal Variants Identification in Associated Regions** (CAVIAR) method of Hormozdiari et al. (2014, 2015), which allows for $\ell \leq k$ of the SNPs in a focal region to be causal. Let $\mathbf{c}$ be a k dimensional vector, whose ith element, c_i, is one if SNP i is causal, and zero otherwise. We make the simplifying assumption that all causal SNPs have the same (standardized) effect β, and that γ is the prior probability that a SNP is causal (both these restrictions can easily be generalized via modifications of the prior). Let s_i denote the standardized marginal association statistic for SNP i. Specifically, the standard gene-dosage regression $z_j = \mu + \beta_i N_{j,i} + e_j$ is fit for SNP i (Equation 20.1a), with $s_i = \widehat{\beta}_i/\sigma(\widehat{\beta}_i)$. Hormozdiari et al. showed that if i is a causal variant in LD with noncausal variant h ($r_{i,h} \neq 0$), then $s_h \sim N(r_{ih}s_i, 1)$ and $\sigma(s_i, s_h) = r_{ih}$. Hence, the vector of association statistics, given the causal SNPs, $\mathbf{s} \,|\, \mathbf{c}$, is MVN with mean vector $\beta\mathbf{C}\mathbf{c}$ and covariance matrix $\mathbf{C}$. As a result, the likelihood of $\mathbf{s}$, conditional on $\mathbf{c}$, follows from Equation 9.24, with

$$\Pr(\mathbf{s} \,|\, \beta\mathbf{c}, \mathbf{C}) \propto \exp\left[-\frac{1}{2}(\mathbf{s} - \beta\mathbf{C}\mathbf{c})^T\, \mathbf{C}^{-1}\, (\mathbf{s} - \beta\mathbf{C}\mathbf{c})\right] \tag{21.7a}$$

Assuming a constant probability γ that any SNP is causal, then the prior on $\mathbf{c}$ becomes

$$\Pr(\mathbf{c}) \propto \prod \gamma^{c_i}(1-\gamma)^{1-c_i} \tag{21.7b}$$

with the restiction that $\sum c_i = \ell$. The resulting posterior for a particular configuration $\mathbf{c}^*$ of causal sites becomes

$$\Pr(\mathbf{c}^* \,|\, \mathbf{s}, \mathbf{C}) = \frac{\Pr(\mathbf{s} \,|\, \beta\mathbf{c}^*, \mathbf{C})\ \Pr(\mathbf{c}^*)}{\sum_{c^* \in C} \Pr(\mathbf{s} \,|\, \beta\mathbf{c}^*, \mathbf{C})\ \Pr(\mathbf{c}^*)} \tag{21.7c}$$

where the set C is all the possible configurations for causal SNPs subject to the constraint that $\sum c_i = \ell$ (ℓ causal variants). For example, assuming a single causal SNP, there are exactly k configurations of one 1 and $(k-1)$ zeros in $\mathbf{c}^*$, with 2 causal SNPs, there are $k(k-1)$ configurations of two ones and $(k-2)$ zeros, and so forth. For a given choice of ℓ, the

posterior value of c_i (the PIP for SNP i) is obtained by summing the c_i values from Equation 21.7c over all of the configuations in C. As an example of how $\mathbf{c}^*$ is constructed with prior information, suppose there are three sites within the focal region, and we assume only one is causal. Further suppose, based on functional annotation (such as a site being an NS variant, near a TSS, etc.), that the three sites show a four-fold, no, and a two-fold enrichment relative to random variant. The resulting weights on these three sites would become (4/7, 1/7, and 2/7).

Colocalization: Exploiting Signals From regSNPs

The TWAS paradigm—a local SNP impacts a transcript, which in turn influences the trait—naturally leads to a deeper discussion of how to leverage functional genomics information to fine-map GWAS hits. With just GWAS summary statistics and SNP LD patterns, one could use functional annotation (when available) to assign weights on the priors of the SNPs in a candidate set when computing their PIPs (e.g., Gagliano et al. 2014; Kichaev et al. 2014; Chen et al. 2016; Wen et al. 2017; Weissbrod et al. 2020). The utility of this approach depends on the annotation accuracy, which is often very poor for many sites with regulatory roles. A biologically more robust approach is an extension of TWAS, namely looking for the **colocalization** between association regions for the trait and association regions for eQTLs influencing a QTT for that trait. More generally, one could use other regulatory QTLs—such as sQTLs, meQTLs, hQTLs, etc.—in place of eQTLs (Table 21.1). There are different levels of granularity associated with a colocalization analysis, depending on both the data set (the amount of recombination in the sample setting the mapping resolution) and the question of interest. Typically, one might use colocalization on a coarse scale to implicate specific genes (as in a TWAS), or on a much finer scale to fine-map specific SNPs that influence both the trait and the molecular feature of interest.

Early searches of candidate loci, even under the crude resolution of linkage-based mapping, looked for loci in a QTL region whose transcript variation was correlated with trait variation. Wayne and McIntyre (2002) used this approach to identify 34 candidate genes for ovariole number in *Drosophilia*. As mentioned above, the problem with simply searching for transcript-trait associations is that the correlation could be generated by the reactive model, where trait variation generates the transcript variation ($z \rightarrow T$), not vice-versa ($T \rightarrow z$). Similarly, both z and T could be impacted by a confounder C ($z \leftarrow C \rightarrow T$), creating a correlation between z and T in the absence of $T \rightarrow z$. Hence, simply finding a GWAS hit close by a coding region whose transcript is a QTT is not sufficient evidence that the causal variant acts through the nearby QTT. A more direct connection is provided by demonstration that the trait GWAS hit is also a regSNP for the QTT that impacts the trait. The initial step in such a demonstration is showing that a GWAS SNP for the trait colocalizes with a regSNP.

There are three different levels of a colocalization analysis. The first is simply determining whether there is support for causal SNPs influencing both the focal trait and molecular intermediates in a given association region. As shown in Figure 21.4, even when a very strong colocalization signal is detected, its interpretation is still unclear. It could be the result of *cis*-mediation, linkage, pleiotropy, or a combination of all three involving multiple, tightly linked, causal variants. Hence, the second step a colocalization analysis is ruling out *linkage* (Figure 21.4B)—assocations generated by tightly linked SNPs with independent effects on the two traits (T and z). The final step, distinguishing pleiotropy (Figure 21.4B) from *cis*-mediation (Figure 21.4A), requires the type of analysis outlined in Example 21.3. One concern is that multiple, tightly-linked *cis*-eQTLs are common (e.g., Zeng et al. 2019), so that allelic heterogeneity (multiple causal variants with the association region) may be the norm (as least for regQTLs).

Early attempts at detecting colocalization between trait and regSNPs were usually visual: looking for alignments of peak SNPs on separate Manhattan plots for the trait and transcript, or using a two-dimensional scatter plot of the trait and transcript $-\log(p)$ values

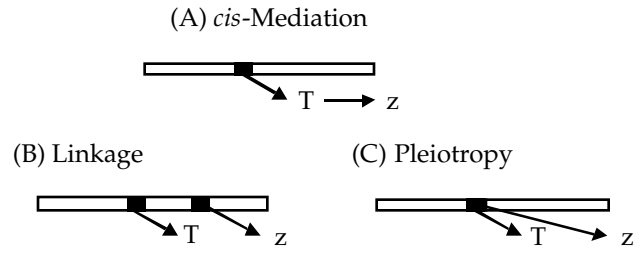

Figure 21.4 An apparent colocalization between a GWAS SNP for trait z (the small black box), transcript (T), and trait (z) could occur through three different pathways. **(A):** Direct *cis*-mediation. The GWAS SNP is an eSNP which directly influences the transcript, which in turn directly influences the trait. **(B):** Linkage. Two tightly linked SNPs are involved. One directly impacts the transcript, the second directly impacts the trait. **(C):** Pleiotropy. The same SNP directly impacts transcript levels and trait values separately, but the transcript level does not impact the trait value. Any combination of these different pathways could be involved, such as a direct *cis*-mediated SNP tightly linked to a separate SNP that only impacts the trait.

for each SNP (e.g., B. Liu et al. 2019). One immediate issue is that the trait and eSNP p values almost always come from different studies, raising concerns (as in a TWAS) of whether the appropriate tissues (given the focal trait) were used in the eSNP mapping and whether the two different samples had the same LD structure. Another complication is stochasticity. Even with the same underlying causal SNP and with appropriate tissue and population match, the sampled lead SNP for the trait and for the eSNP are likely to map in different locations. For example, if the lead SNP and causal site agree 50% of the time (which is often overly optimistic given our earlier discussion), there is only a 25% chance that the lead GWAS and lead eSNP agree, and there is a good chance that the two lead SNPs will map kilobases apart, even if they have the same underlying causal SNP.

As a result of these concerns, a variety of increasingly more formal approaches have been proposed for testing colocalization. When full (i.e., individual) data is available on both traits, **QTLMatch** tests whether regressions of each character (the focal trait and transcript) against multiple candidate SNPs have proportional slopes, as expected if the SNPs jointly tag a common variant (Plagnol et al. 2009). A variant of this approach was used by Barbeira et al. (2021), who examined the correlation between a SNP's regulatory effect (they looked at both eSNPs and sSNPs) and its effect of the focal trait, finding correlations of 0.18 for eSNPs and 0.25 for sSNPs over a set of almost 75 different human traits/diseases. The **regulatory trait concordance** (**RTC**) method of Nica et al. (2010) also requires individual data (at least for expression sample), and looks at the impact on expression data after including the peak GWAS SNP as a cofactor. The **Sherlock** approach of He et al. (2013) requires only summary statistics, and looks for a concordance between GWAS SNPs and eSNP for the focal transcript. Because there is low SNP filtering, even eSNPs with weak effects (e.g., with nominal, but not genome-wide, significance) are included, and as a result Sherlock can incorporate *trans* signals. Sherlock also has an useful evidence asymmetry feature in that an eQTL for the focal transcript not corresponding to a GWAS SNP is taken as support against the regSNP-GWAS colocalization, but a GWAS hit that does not correspond to an eQTL for the focal transcript has a neutral impact on the support for colocalization.

These early methods have largely been replaced by Bayesian approaches (e.g., Giambartolomei et al. 2014; Hormozdiari et al. 2016; Wen et al. 2017; Roytman et al. 2018; Wang et al. 2020). The **coloc** method of Giambartolomei et al. (2014) uses summary statistics (and does not require any LD data), returning the posterior probabilities for five different hypotheses concerning the association of a region with two traits. These are: no association within the region for either trait (PP0); association with one trait, but not the other (PP1 and PP2); association with both traits, but via two independent SNPs (PP3); and association with both

traits due to a shared SNP (PP4). A helpful feature of coloc is that it distinguishes evidence for colocalization (high PP4 value) from a lack of power (PP0, PP1, or PP2 have high values). Further, the values of PP3 (linkage) and PP4 (pleiotropy) provide support for distinguishing between these two sources of a colocalization signal. The **multiple-trait-coloc** (**moloc**) method of Giambartolomei et al. (2018) extends this approach to the colocalization of any k features (such as a trait SNP, an eSNP, and an meSNP), but still under the coloc assumption of (at most) a single causal site (in the association region) for each feature. Wallace (2020, 2021) developed approaches to extend coloc to allow for multiple variants (impacting each trait) within an association region. The TWAS **MetaXcan** framework of Barbeira et al. (2018) uses these PP values following an initial TWAS, upweighting inclusion of regions with high PP4 values and removing regions with high PP3 values.

While coloc tests for colocalization within an association region, a colocalization fine-mapping approach (*which* SNP within that region), even when multiple causal sites are present, is offered by the **eCAVIAR** (**eQTL and GWAS Causal Variants Identification in Associated Regions**) method of Hormozdiari et al. (2016). This extends their single-trait fine mapping CAVIAR method (outlined in Example 21.9) to jointly consider the causal support for a *pair* of traits. The strength of evidence of colocalization at a specific SNP is given by a **colocalization posterior probability** (**CLPP**). For a vector of M candidate SNPs within an association region, CAVAIR returns PIP_i, the probability that SNP i in this region is causal. Basically, eCAVIAR works by computing PIP_i^z and PIP_i^T, the posterior probability that SNP i is causal for the trait (z) and transcript (T), respectively. The resulting CLPP for SNP i becomes $\text{CLPP}_i = \text{PIP}_i^z \cdot \text{PIP}_i^T$, an approximation that does not account for enrichment.

As noted by Wen et al (2017) and Hukku et al. (2021), there are connections between fine-mapping, colocalization, and enrichment (next section) that are apparent when one considers the priors for Bayesian methods. Following Example 21.9, let c_i^z and c_i^T be 0/1 indicator variables that SNP i is causal for the trait and transcript, respectively. From Bayes theorem, the resulting expression for the CLPP, given the data D, can be expressed as

$$\Pr(c_i^z = 1, c_i^T = 1 \,|\, D) \;\propto\; \Pr(D \,|\, c_i^z = 1, c_i^T = 1) \Pr(c_i^z = 1, c_i^T = 1) \tag{21.8a}$$

Focusing on the prior, we can write this as

$$\Pr(c_i^z = 1, c_i^T = 1) = \Pr(c_i^T = 1 \,|\, c_i^z = 1) \cdot \Pr(c_i^z = 1) \tag{21.8b}$$

$\Pr(c_i^T = 1 \,|\, c_i^z = 1)$—the probability that the eSNP is likely to be causal when the trait SNP is causal—represents the enrichment of eSNPs (or any other regSNPs) given a trait GWAS hit. When this conditional probability exceeds $\Pr(c_i^T = 1)$—the chance that a random SNP is an eSNP—then GWAS SNPs are enriched for eSNPs. The **ENLOC** method of Wen et al (2017) models this enrichment using a logistic regression (Equation 20.2b), with

$$\text{logit}\left[\Pr(c_i^T = 1 \,|\, c_i^z = 1)\right] = \alpha_0 + \alpha_1 c_i^z \tag{21.8c}$$

where the α_i can be fitted empirically. A value of $\alpha_1 > 0$ measures the log of the odds ratio increase in enrichment when the SNP is causal ($c_i^z = 1$) for the trait. A variety of enrichment tests are reviewed in the next section.

Example 21.10 Another approach for distinguishing linkage from pleiotropy—the latter lumping both traditional pleiotropy (Figure 21.4C) and *cis*-mediated (Figure 21.4A) effects—was suggested by Zhu et al. (2016). Their **SMR** method is based on Mendelian randomization (MR), using GWAS summary statistics. As is more fully developed in Appendix 2, the concept of MR is essentially a subset of path analysis, built around the concept of mediation (Figures 21.3 and Example 21.3). The methodology comes from epidemiology, where one tunes some **instrumental variable** (here, a SNP genotype) that impacts a mediator, that in turn impacts an outcome. Consider a genotype (g, scored by reference allele copy number, N), transcript (T),

and trait value (z). Let β_{Nz} and β_{NT} be the regression slopes of the trait and transcript level, respectively, on minor allele copy number for a given SNP (obtained from trait and transcript GWAS). If $g \rightarrow T \rightarrow z$, then assuming no additional path from g to z that is independent of T, $\beta_{Nz} = \beta_{NT} \cdot \beta_{Tz}$, so that $\beta_{Tz} = \beta_{Nz}/\beta_{NT}$ is the mediation effect (Example 21.3). The logic of Zhu et al. is that if the same causal variant underlies both z and T (whether through *cis*-mediation or thorough the pleiotropy effects of g on both z and t), then the ratio β_{Tz} should have the same expected value for any SNP linked to the causal site (as the correlation between the SNP and causal site appears in both the numerator and denominator of β_{Tz}, and thus cancels). In their **HEIDI** test (**heterogeneity in independent instruments**, the "instrument" being the genotypes at the focal SNP) of pleiotropy versus linkage tests for *heterogeneity* of the β_{Tz} estimates over a set of linked SNPs in the association region. Such heterogeneity is consistent with linkage (different SNPs), but inconsistent with pleiotropy (assuming a single causal variant). Hence, a significant p value indicates linkage as the source of the colocalization.

An alternative approach to resolving linkage versus pleiotropy when a strong colocalization signal is found was offered by Chun et al. (2017). They reasoned that if the same causal SNP underlies both the trait and expression signal, then this joint evidence should be maximized at markers in the tightest LD with the causal site. Conversely when this association is due to LD between causal sites with independent effects, a different likelihood structure occurs, and they proposed an LR test for this (the same type of linkage versus pleiotropy test have been proposed for linkage-based mapping; Chapter 18).

GENE-SET, PATHWAY, AND NETWORK ANALYSIS

One outcome from the initial wave of whole-genome expression studies was an avalanche of data in the form of lists (often very long lists) of differentially expressed genes between case and control samples. As highlighted above, the development of numerous other high-throughput platforms extends this data to differences in protein levels, methylated sites, splicing isoform ratios, metabolite levels, chromatin modification differences (e.g. the presence of acetylated or methylated histones at a target site), and so on. Hence, while our comments are framed in terms of expression data, the approaches presented here obviously extend over large classes of other genomic and cellular features. In an attempt to extract insight from such data, a number of increasingly sophisticated **knowledge-based approaches** have been developed. These attempt to leverage information from functional genomics databases, such as the functional category of a gene (e.g., a transcription factor, a kinease, etc.), their membership in known pathways, or their interaction partners in networks (e.g., proteins that make physical contact in the cell). Although initially developed for expression data, it was quickly realized that these approaches could be easily modified and applied to GWAS data. We start by briefly considering some of the basic logic of these methods as applied to expression data before focusing on their applications to GWAS. Reviews of expression analysis methods can be found in Drăghici and Krawetz (2003), Allison et al. (2005), Khatri and Drăghici (2005), Curtis et al. (2006), Nam and Kim (2008), Fridley et al. (2010), and Rahmatallah et al. (2016).

Gene Set Analysis of Expression Data: Enrichment Methods

Using a collection of genes as the unit of analysis is loosely called a **gene set analysis** (**GSA**), as opposed to an **individual gene analysis** (**IGA**), which considers genes one at a time. A GSA is often called a **pathway analysis** (not to be confused with the regression-based method of *path analysis*; Appendix 2) when the gene set members are chosen from a known pathway (e.g., $A \rightarrow B \rightarrow C$). A GSA is used for two different, but not necessarily independent, reasons. The first is *power*, in that a collection of individually weak IGA signals—such as those showing nominal (e.g., $p < 0.05$), but not genome-wide ($p < 0.05/n$, with n very large), significance—might be boosted under a GSA (akin to gene-based and rare-allele GWAS approaches considered in Chapter 20). The second is that finding signif-

icant *sets of genes* (such as members of the same functional class or pathway) may provide greater biological insight than a focus on single genes. If a disease is caused by the *disruption of a pathway*, individuals (and populations) may harbor mutations in different genes in that pathway, resulting in a rather weak *gene-based* signal (Wang et al. 2007; Schadt 2009). A *pathway-based* signal, however, may not only be stronger, but could also be more consistent over replicates (e.g., Elbers et al. 2009; Wang et al. 2009).

Example 21.11 shows the initial GSA approach, **enrichment analysis** (**EA**; also called **overrepresentation analysis; ORA**), which asks if there is a modest, but coordinated, shift in expression levels over genes in some known pathway or in some defined category. The basic structure of an enrichment test is that one starts with some bioinformatics database and then examines whether differentially expressed genes are enriched for some feature (Tavazoie et al. 1999; Mootha et al. 2003; Subramanian et al. 2005). In one of the early applications of GSA, Mootha et al. (2003) showed that genes involved in oxidative phosophorylation had coordinately decreased expression levels in the muscle tissue of diabetics. As mentioned in previous sections, enrichment analysis has been widely used to examine whether lead SNPs from a trait GWAS are enriched for SNPs other features (such as regSNPs) or are overrepresented in specific genomic regions (such as sites with more open chromatin).

Example 21.11 Suppose that one has expression data for 5000 genes in normal versus cancer cells, 300 of which are declared to show significant differences (using some threshold critical value, say $\alpha = 0.05$). Now suppose we choose (by some criteria) a set of these scored genes (denoted by $\mathcal{G}$; the remainder are in the complement set, $\mathcal{G}^c$), which consists of 100 members, 20 of which show significant differences. The resulting 2×2 contingency table of expression differences versus set membership becomes

	Significant	Not Significant	Totals
Gene set $\mathcal{G}$	20	80	100
Gene set $\mathcal{G}^c$	280	4620	4900
	300	4700	5000

Using a standard Chi-square test shows that differentially expressed genes are significantly overrepresentated in our gene set ($p = 9 \times 10^{-9}$).

With smaller gene sets, the observed value in one (or more) of the entires in the contingency table can be small, in which case Fisher's exact test is more appropriate. This latter test is based on the **hypergeometric distribution** (Equation 2.25). Under the null hypothesis (no enrichment), the probability of observing at least k genes from a functional category (i.e., $\mathcal{G}$) in a sample of n genes declared to show significant differences is given from the upper tail of the hypergeometric,

$$p = \Pr(X \geq k) = \sum_{i=k}^{n} \Pr(X = i) = \binom{g}{n}^{-1} \left[\sum_{i=k}^{n} \binom{f}{i} \binom{g-f}{n-i} \right] \tag{21.9a}$$

where f is the total number of genes in the functional category and g is the total number of scored genes, where $g \geq f \geq n$ (Drăghici and Krawetz 2003). If n is small relative to f, one can approximate the hypergeometric (sampling without replacement) by a binomial (sampling with replacement) with success parameter $\pi = f/g$ and sample size n (Drăghici and Krawetz 2003), giving

$$p = \Pr(X \geq k) = \sum_{i=k}^{n} \Pr(X = i) = \sum_{i=k}^{n} \binom{n}{i} \pi^i (1-\pi)^{n-i}, \quad \text{with } \pi = f/g \tag{21.9b}$$

There are a variety of enrichment methods based on such 2×2 contingency tables, differing in whether a chi-square, Fisher, hypergeometric, binomial, or other (e.g., normal approximation of a binomial; Doniger et al. 2003) is used to test significance (Khatri and Drăghici 2005; Curtis et al. 2006; Tryputsen et al. 2014).

While Example 21.11 shows the basic logic for enrichment approaches, it leaves open a number of important issues. The first is choosing the genes to include in the tested set ($\mathcal{G}$). This could be *hypothesis-driven*, testing a few candidate gene-sets, akin to the candidate gene approach of testing associations (Chapter 17). Alternatively, one could adopt an *exploratory approach*, testing over a wide number of sets, whose gene elements are chosen by some functional criteria. A major tradeoff between these approaches is correcting for multiple comparisons (tests of specific gene sets), which imposes a mild burden in a candidate approach, but can be large in an exploratory analysis.

One classifier for set construction is to use genes that cluster in the same **Gene Ontology** (**GO**) group (Harris et al. 2004). GO is a hierarchical vocabulary of function (e.g., pigmentation → regulation of pigmentation → regulation of eye pigmentation), with categories becoming larger (and more general) as one moves up the GO hierarchy (toward **parent nodes**, e.g., pigmentation) and narrower and more specific as one moves down the hierarchy (toward **offspring nodes**, e.g., regulation of eye pigmentation). One issue with testing different sets of GO genes is that the hierarchical nature of their labels implies than genes in different groups can be correlated. Zhang et al. (2010) suggested Bayesian approaches to model this dependency structure.

Another criteria is to choose genes from known pathways, for example using the **Kyoto Encyclopedia of Genes and Genomes** (**KEGG**; Kanehisa et al. 2010) database. While deciding which elements constitute a specific pathway might seem straightforward, this is not always the case. A pathway is often treated as a discrete modular item, when, in reality, smaller pathways are often nested within much larger pathways and networks. Hence, the same gene could play key roles in several different user-defined pathways, creating correlations among the analyses. Defining a specific pathway raises issues akin to those of trying to assign SNPs to a particular gene that were discussed in Chapter 20 (in the context of gene-based GWAS). Besides the biological concerns as to what constitutes a reasonable pathway, there are also statistical issues with gene choice (which pathway members to include). As was the case for rare allele methods (Chapter 20), inclusion of too many pathway genes with no effect on a trait in the gene set diffuses any true signal. Conversely, there is a bias towards larger gene-sets being significant (Holmans 2010; Wang et al. 2010; Ramanan et al. 2012) unless set size is appropriately controlled (for example, by using permutation for significance testing).

The strength of any GSA depends on the quality of the data used to choose gene set members. If a gene is incorrectly classified (**annotated**) as to its functional group, this lowers the power of an analysis using that gene. Further, a gene can (correctly) be assigned to multiple functional groups. In theory, one could use Bayesian methods to weight set members by some metric of their perceived accuracy of correct assignment (such as was done with imputed SNPs; Chapter 20). The same issue of quality of the bioinformatics data using GO exists with pathways, with a further complication. The nature of a pathway can easily change over cell type, developmental stage, or environment (such a high versus low sugar diet, or well-watered vs. drought conditions).

A second issue is that the classification of an expression difference as being significant in an EA is *arbitrary*, depending on the critical-value threshold used. As one changes this threshold, the elements in the 2×2 tables change, potentially changing the significance of the gene set. The **gene set enrichment analysis** (**GSEA**) approach of Mootha et al. (2003), Subramanian et al. (2005), and Efron and Tibshirani (2007) avoids the arbitrariness of thresholds by basing tests on the *ranks* of some expression statistic (such as their p values or score statistics). One then tests for an *enrichment of ranks* in the gene set relative to a control set of scored genes. This results in a running **enrichment score** statistic that changes with each new gene added, yielding in a Kolmogorov–Smirnov-like test statistic. At some point the score reaches its maximal value (**maximum enrichment score**, **MES**), after which is starts to decline as further genes are added. The set of genes that yields the MES was called the **leading-edge subset** by Subramanian et al. (2005), and contains the core members of the gene set that provide the greatest signals. While the original GSEA method assumed cor-

related expressions changes occurred in the same direction (e.g., all up-regulated), Saxena et al. (2006) extended it to an **absolute enrichment** (**AE**) approach that scores absolute (as opposed to signed) changes in expression levels between cases and controls.

A final issue is more subtle, but no less critical: ***What is the appropriate null hypothesis?*** Is it that (i) there is *no* differential expression for any of the genes in the candidate set ($\mathcal{G}$), or is it that (ii) the candidate set does not contain *more* differentially expressed genes relative to the rest of the scored genes ($\mathcal{G}^c$). This distinction was first noted by Tian et al. (2005), and formalized by Goeman and Bühlmann (2007), who denoted the first null as **self-contained tests** ($H_{o,sc}$), and the second as **competitive tests** ($H_{o,co}$; such as is performed in a GSEA). Generally speaking, tests against $H_{o,sc}$ are more powerful than tests against $H_{o,co}$, as if $H_{o,sc}$ is true, then so is $H_{o,co}$, while the converse relationship does not hold. As noted by Goeman and Bühlmann, a self-contained test "always has a clear biological meaning. At the same time, it may not always be biologically interesting," as we *expect* some differential expression over the genome. Conversely, a self-contained test does not require information from genes outside of the candidate set.

To examine these different nulls more closely, suppose that the number of differentially expressed genes in $\mathcal{G}^c$ in Example 21.11 was 980 (20%), the same fraction as in $\mathcal{G}$. The resulting 2×2 table now becomes nonsignificant. This is a competitive test ($H_{o,co}$), with the null being that the fraction of differentially expressed genes in the test set is no greater than in the control set. This is a natural comparison in many settings, as the levels of expression may increase over a wide fraction of the genome in (say) diseased versus control individuals. In this situation, the question of interest is whether genes in the control set are *enriched* for differential expression. Conversely, if we restrict our attention to *just* the gene set ($H_{o,sc}$), the expected number of false positives under the null is 5 ($n\alpha = 100 \cdot 0.05$), with the observed value of 20 being highly significant (one can use either Fisher's exact test or a binomial with $\pi = 0.05$ and $n = 100$). Ebrhimpoor et al. (2020) showed that both of these nulls ($H_{o,sc}, H_{o,co}$) are special cases of a more general null that they called **simultaneous enrichment**. Here the null is that the fraction of significant differences in $\mathcal{G}$, $\pi(\mathcal{G})$, is $\leq c$. Under the self-contained null, $c = 0$ ($\pi(\mathcal{G}) = 0$); while under competitive null $c = \pi(\mathcal{G}^2)$, namely $\pi(\mathcal{G}) \leq \pi(\mathcal{G}^c)$.

The structure of permutation tests change under these different nulls (Goeman and Bühlmann 2007). A self-contained test permutes the phenotypic labels while holding the vector of expression data for a single individual constant (often called **subject sampling**). Conversely, a competitive test involves **gene sampling**. Here, the data vector for a given individual consists of phenotypic values, expression data, and labels on which genes (transcripts) are in gene-set $\mathcal{G}$ versus set $\mathcal{G}^c$. Permutation now involves keeping the phenotypic and expression data intact, but shuffling the gene-set membership labels. For example, if our original set consists of 10 genes, a permutation sample performs the analysis using the same phenotypic and expression data, but the ten genes to be labeled over all individuals as being the case set ($\mathcal{G}$) are randomly chosen from the sample of all genes. As noted by a number of authors, gene-sample permutation assumes independence of expression changes, which is often not the case. Further comments on subtle features of the nature of the competitive null, and their permutation tests, are offered by Efron and Tibshirani (2007), Gatti et al. (2010), Maciejewski (2013), and Debrabant (2017).

Gene Set Analysis of Expression Data: Next-generation Methods

Enrichment analysis (EA) represented the first generation of GSA methods for expression data. As noted by Khatri et al. (2012), this early approach had a number of limitations. They essentially treated each member in the class as having equal weight, and assumed that each gene was independent of all other genes in the set. For many EA tests, each member is simply a binary data point, counted as either present or absent in a given catergory.

Second-generation methods, which Khatri et al. refer to as **functional class scoring** (**FCS**), assign weights to each gene in the set, such as their p value or some score statistic based on their amount of differential expression. The problem of how to combine this information, especially when the individual gene metrics are potentially correlated, is one

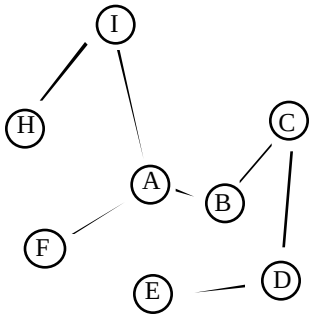

Figure 21.5 A **network** is a collection of **nodes** (the elements A through H), which are connected by **edges**, showing which elements interact with each other. For example, A directly acts with B, F, and I. In a protein-protein interaction network, this graph implies that protein A contacts proteins B, F, and I in the cell, while B contacts A and C, and so on. Appendix 2 examines the toplogy (shape and connectiveness) of networks in more detail.

that we have encountered before. In Chapter 20, we examined a variety of methods for aggregating correlated SNP data from a single gene into a gene-specific statistic (gene-based GWAS). All of these approaches (p-value combining approaches such as Fisher, GATES, or rank-truncation; multiple regressions and Hotelling's T; penalized regressions such as LASSO; random-effect variance component models, etc.) have been used to construct FCS tests, where the statistics being combined are based on the expression data for each gene. The discussion from Chapter 20 on hypothesis testing when gene-specific statistics are correlated also applies to FCS methods. Permutation approaches are the gold-standard, but simulation and large-sample approximations have been proposed (see Chapter 20 for details). As with first-generation methods, null hypotheses can be framed as either self-contained or competitive tests. Given the number of combinations of different single-gene expression statistics, aggregating statistics over the gene set, and different null hypotheses, it is not surprising that a massive number of gene-set methods have been proposed (Ackermann and Strimmer 2009 examined over 250 different combinations of these elements, while Xie et al. 2021 noted over 100 different published methods/programs).

One of the emerging fruits of the functional genomics revolution is a more detailed understanding of cellular pathways and networks. In particular, we often have detailed information about the **topology** (the structure of connections between elements; Appendix 2) of pathways and networks (such as the web of protein-protein interactions). EA and FCS ignore this information when constructing pathway-summary statistics. **Pathway topology** (**PT**) methods attempt to exploit this additional knowledge, and can be considered as the third generation of GSA approaches (Khatri et al. 2012). For example, Pan (2008) gave higher weights to nearby genes (closer links in the pathway/network topology). Similarly, a network (such as a **protein-protein interaction network**; **PPI**) may not have the directionality of a pathway, but still has a complex topology of interactions (e.g., Figure 21.5). Such information can also be used in GSA. An early such example is Baranzini et al. (2009), who found that certain subsets of the human PPI were enriched in individuals with multiple sclerosis. An overview of these topology-informed methods is given by Mitrea et al. (2013).

We can also consider a fourth generation of methods, **dynamical pathway approaches** (**DPA**), that use information from not only the shape (adjacency and connectivity of elements) but also the **flux** (rate of flow) of products over the pathway, such as the **impact analysis** method of Drăghici et al. (2007). These third- and fourth-generation approaches forge connections between classical quantitative genetics and functional genomics.

Genome-wide Pathway Analysis (GWPA)

It was quickly realized that expression-based GSA approaches could be applied to GWAS

datasets (K. Wang et al. 2007, 2010; Torkamani et al. 2008; Elbers et al. 2009; Hong et al. 2009; Cantor et al. 2010; De la Cruz et al. 2010; Peng et al. 2010; Zhong et al. 2010), leading to the notion of a **genome-wide pathway analysis**, or **GWPA** (Ramanan et al. 2012), also called **GWAS pathway analysis**, **GWASPA** (Cantor et al. 2010). A GWPA represents the logical extension of the basic unit of a GWAS from a *SNP* to a *gene* to a *gene set*. As with a gene-based GWAS (Chapter 20), motivating factors for a GWPA are the potential of increased power by combining weaker signals, reduced multiple-comparison burden (reduced complexity by reducing the number of tests), greater replication based on pathways (rather than on specific genes), and the hope for greater biological insight. Before proceeding, the reader might find it useful to review the material from Chapter 20 on gene-based GWAS.

Chapter 20 examined two secondary analysis methods to mine hard-won SNP-based GWAS data: gene-based GWAS and meta-analysis. GWPA represents a third class of data-mining approaches, and can be also viewed as a meta-analysis where data (genes) from *within* a single study is distilled into *summary statistics over gene sets*. A gene-based GWAS can similarly be considered a within-study meta-analysis, combining SNPs into gene-based units of analysis. Gene-based GWAS and GWPA from a single study can themselves be placed in a meta-analysis framework to combine results over multiple studies. While a GWPA can be very useful complement to SNP-based or gene-based GWAS, it is best considered as an exploratory analysis to gain additional insight and to guide future research directions.

While the basic logic, and many of the analysis methods, from an expression-based GSA extend to a GWPA, there are two complications, both based on extracting gene-specific scores. Under an expression-based GSA, one has a single variable (mRNA level) for each gene (ignoring alternatively spliced products). What is the corresponding GWPA metric for a gene, which consists of data from a number (perhaps a large number) of SNPs? Specifically, an investigator must decide (i) which SNPs are assigned to a specific gene, and (ii) how to combine this SNP data into a gene-specific score. There a number of proposed approaches for the latter issue that were examined in Chapter 20 (gene-based GWAS).

The question of which SNPs comprise a gene is less clear, but is usually accomplished by considering a set of SNPs within some defined window around a coding sequence (akin to considering only proximal SNPs when searching for local eSNPs). This set could be all the SNPs in the window, or a trimmed set, such as removing pairs in very high LD, or by extracting PCs (Chapter 20). This window-based approach can result in a single SNP being assigned to multiple genes in a cluster, which can cause complications. An example of this was highlighted by Sedeño-Cortés and Pavlidis (2014), who noted that Dixson et al. (2014) assigned a single SNP to multiple members in a gene cluster, each of which had a similar GO annotation, resulting in a GWPA false-positive. As with rare-allele GWAS methods (Chapter 20), power is improved by excluding SNPs with no apparent functional effect. However, the question of which of SNPs are likely functional (especially in noncoding regions) is, at best, poorly resolved (see the discussion above and in Chapter 20). Further, potential control regions for a gene many occur at some distant from the coding sequence, such as long-range *trans*-acting factors. These can potentially be handled in an eSNP framework, using expression data for the target gene (e.g., Zhong et al. 2010).

Because a GWPA involves two levels of data amalgamation—turning collections of SNPs into a *gene score* and turning collections of gene scores into a *set score*—there are two strategies for performing a GWPA: **one-step** versus **two-step** approaches. A one-step approach simultaneously incorporates all of the chosen gene-set SNPs into a single analysis, using the methods from Chapter 20 for combining results for multiple markers. In a two-step analysis, one first generates summary statistics (such as p values) for each gene (e.g., gene-based GWAS), and then combines these scores over the set to obtain the gene-set-level statistic. Which approach is used depends on both data availability and assumptions about the underlying biology. Based on the nature of the gene set, a one-step approach is favored when most of the signal comes from just a few SNPs. In the extreme, a phenotype could be simply the product of one gene and be largely independent of the rest of the pathway.

Similarly, when multiple sites within a gene contribute to trait variation, collapsing these to a single gene score can lower power and skew the biological interpretation. Conversely, when there are weak signals over multiple genes, then a two-step approach is favored.

In a two-step approach, one can mix and match analysis methods (developed for gene-based GWAS), using one approach within genes and a second for combining the gene-level statistics. For example, the **GRASS** (***g*ene set *r*idge regression in *a*ssociation *s*tudies**) method of Chen et al. (2010) uses the LASSO (Example 20.4) to obtain gene-level statistics, and then uses a p value combining method to obtain a gene-set score. The **HYST** (***hy*brid *s*et-based *t*est**) method of Li et al. (2012) uses GATES (extended Simes method; Equation 20.4f) to compute gene-level scores and then combines these using a modification of Fisher's method to allow for correlation among genes. Many two-step methods use the **Min-p approach** for each gene, namely just using the SNP with the smallest p value (i.e., the lead SNP). This has both statistical and biological limitations. Clearly, the more tested SNPs within a gene, the smaller the expected value of Min-p. This gene-size bias can be overcome by using standard permutation methods of keeping the genotypes intact while shuffling phenotypic labels. The biological concern is that if multiple sites in the focal gene are segregating causal alleles, the full impact of the gene is not captured by only using the largest SNP effect.

As with expression-based GSA, an investigator must decide between a self-contained hypothesis (*none* of the gene set elements have an impact on the trait) or a competitive hypothesis (the gene set is not *enriched* for trait-impacting genes relative to the rest of the genome). While one might consider the self-contained hypothesis to be more natural, de Leeuw et al. (2016) strongly argued that the competitive hypothesis is often more appropriate. Their reasoning is that a highly polygenic trait has a large number of underlying genes spread over the genome, so that any random set of genes has some chance of containing causal loci. In this setting, the more natural question becomes whether there is an *overrepresentation* of causal loci in the candidate set.

A final issue is the choice of members in a gene set. One entree into a potential pathway could be through a SNP showing genome-wide significance in a standard GWAS. If such a SNP can be assigned to a known gene, one could then test known associated pathways. Given the ascertainment bias created by choosing a (genome-wide) significant SNP, the entree gene should be removed from the gene set before proceeding. While it could simply be left out of the gene set, due to potential correlated effects among pathway members, a cleaner approach is to include it as a cofactor (as is often done in linkage-based QTL mapping to remove the effect of the leading QTL; Chapter 18). A related issue is that candidate pathways should only be considered when one (or more) of their associated SNPs show a nominal level of significance in the original GWAS (Sedeño-Cortés and Pavlidis 2014), otherwise one could simply search for trait-pathway associations until one is found by chance. Another issue is the size of the tested gene set, as there is a bias towards larger gene sets being more significant (Holmans 2010; K. Wang et al. 2010; Ramanan et al. 2012). As with many concerns in GSA, this bias can be controlled by using permutation for hypothesis testing. Lastly, beware of false-positives generated by a single SNP. The complex systems nature of biological organisms dictates that components/products/intermediates from one pathway can also be involved in several others. Hence, a SNP with genome-wide significance in a GWAS that is involved in several (known) pathways might falsely generate signals in each, when in reality the impact of the gene associated with the SNP on the focal trait is through yet another, undiscovered, pathway.

As would be expected given our above brief comments, there is a massive literature on GWPA methods, reviews of which can be found in Huang et al. (2008), Yu et al (2009), Holmans (2010), K. Wang et al. (2010), Fridley and Biernacka (2011), Khatri et al. (2012), Ramanan et al. (2012), Schaid et al. (2012), Maciejewski (2013), Mitrea et al. (2013), Tarca et al. (2013), Mooney et al. (2014b), Newton and Wang (2015), de Leeuw et al. (2016), Tamayo et al. (2016), and Xie et al. (2021). Given all of the possible analysis options in a GWPA, there is a real risk of "method-shopping" to find the best p value. Investigators thus must be as transparent as possible about the process of model selection that led to their final result.

Some best practices for reporting GWPA results were suggested by Mooney et al. (2014a) and Mooney and Wilmot (2015).

Given all of these options, how should one proceed with a GWPA given a candidate gene set (or a small number of such sets)? A nice case study was given by Gui et al. (2011), looking at Crohn's disease. They examined seven different GWPA approaches, finding that immune-response related pathways tended to be significant, but the number of other detected pathways varied greatly depending on the approach used. Such a variance in outcomes is not unexpected. There is no single omnibus test for trait-pathway associations, as the power of detection is a function of the underlying trait architecture in a given pathway. If some causal pathways mainly influence the trait via a few genes of modest to large effect, while other causal pathways impact trait value through numerous genes of small effect, a method detecting the first class of pathways may have low power for the second class, and vice versa. To address this concern, Gui et al. recommended using several different methods, ideally choosing from self- versus competitive-hypothesis methods, and methods that used information from just a few SNPs in each gene (e.g., Min-p) and methods that incorporate the effects over all of the SNPs (e.g., GRASS).

Describing the Structure of Networks (Systems Biology)

To close our discussion on the ongoing merger of quantitative genetics (QG) and functional genomies, we note that phenotypes are the output of highly complex systems involving a vast number of molecular intermediates that interact with each other and with their environment. The lowest level of organization of such systems are the effects of individual genes on a specified trait (GWAS). At the next level, these genes interact with each other in simple pathways (moving from some starting input to a final product, e.g., a metabolic pathway $A \rightarrow B \rightarrow C \rightarrow D$). Pathway-based GWAS offers a very crude glimpses into these interactions by asking if a set of GWAS hits is enriched for targets in a particular pathway (or set of pathways). We can also examine regulatory (as opposed to a metabolic) pathways, and probe some of their features by searching for regQTLs. Individual pathways themselves are embedded into much more complex structures (**networks**), that, at some level, likely encompass every molecular feature within a cell. The newly emerging field of **systems biology**—an eclectic blend of concepts from physics, graph theory, cybernetics, biochemistry, and molecular and cellular biology—attempts to model the organization, evolution, and functional implications of these structures. Brief introductions to the field, mainly focusing on the topology of biological networks, and its implications, are given by Barabási and Oltvai (2004), Vidal et al. (2011), and Hu et al. (2016), while Civelek and Luisis (2014) examines trait variation from a systems biology viewpoint.

As this chapter has highlighted, the machinery of quantitative genetics applies to any object that shows variation, and this includes network structures. As briefly introduced in Appendix 2, we can describe the static structure of a network with a matrix, $\mathbf{M}$, which describes how objects in a pathway (nodes, such as particular molecular features) are connected by edges (Figures 21.5, A2.9, and A2.10). Edges could be directional (a arrow indicating that one node influences another), or undirected (shown by a line such as in a protein-protein interaction map, indicating which proteins contact each other in a cell). The collection of nodes and their corresponding edges at any point in time is called the topology of a network, and can obviously change over time. Vidal et al. (2011) has called this structure the **interactome**, and its study **edgetics**.

The next level of resolution beyond topology would be the strength of the edges a network, and the highest level of resolution would be the dynamics over the network (how the edges change over time, both in terms of connectivity and strength). Very formally, one has a matrix $\mathbf{M}(t)$ describing the network at some time t, and a vector $\mathbf{m}(t)$ whose elements are the concentrations of various molecular features (again, at time t), and we wish to map $[\mathbf{M}(t), \mathbf{m}(t)] \rightarrow [\mathbf{M}(t + \delta t), \mathbf{m}(t + \delta t)]$. The full solution is an almost unthinkably complex problem, but one can chip way it from two different directions. The first is top-down: Do some basic features arise from such systems, largely independent of their component pieces?

The answer is yes, and biological robustness (relative insensitivity to perturbations) might be one such **emergent propriety** of the topology of biological networks (e.g., Barabási and Oltvai 2004; Kitano 2004, 2007; Levy and Siegal 2008; Masel and Siegal 2009; Appendix 2). The second direction is bottom-up, in that components of $\mathbf{M}$ (such as the probability at a given node is connected to another specified one, and the strength of this connection) are quantitative traits, and the machinery of QG can be used to probe some of their features, such as the amount of pleiotropy and the nature of any underlying regQTLs. Finally, a growing number of modelers are examining the implications for network structures for generating QG features, such as epistasis and pleiotropy, e.g., Omholt et al. (2000), Ayroles and Zeng (2008), Kliebenstein (2009), and Hu et al. (2011). Balancing the energy and enthusiasm towards applying QG methods to pathway and network variation, Flint and Ideker (2019) caution that "the difficulties in integrating network and genetic data are under appreciated," and run the risk of false positive finds often seen in the early days of small-sample GWAS. It is clear that GWAS offers cautionary tales to bear in mind moving forward, the main one being that large sample sizes are critical, and yet most of the current data on network variation is based on very small samples.

WHAT DOES GENOMICS TELL US ABOUT TRAIT ARCHITECTURE?

Finally, we come to the two central questions in quantitative genetics: (i) what is nature of molecular processes that translate genetic into phenotypic variation (the **genotype-phenotype map**, **GP**), and (ii) how has evolution shaped both the nature of standing variation and the systems underlying the GP map. We start by reviewing the debate on whether trait variation is mainly the result of common alleles of small effect or rare alleles of large effect. This naturally leads to the **missing heritability** concern that was expressed in the early days of GWAS, wherein genome-wide significant SNPs accounted for only a small fraction of the heritability estimated using more classical approaches (e.g., Chapters 21–25, 31, and 32). When then briefly examine the architectures of a few classic traits before finishing with models that try to synthesize many of the results in above and from Chapters 17–20 into a framework for how quantitative variation is generated. Our treatment of many of the technical population-genetic issues that arise here will be rather brief. Our second volume both reviews the very rich theory of population genetics (WL Chapters 2-5 and 7-10) and discusses population-genetic models of selection on quantitative genetics in detail (WL Chapters 24–28), and so we refer to the reader to this material for a much deeper discussion.

Common versus Rare Alleles

An ongoing debate in quantitative genetics is whether the bulk of genetic variance is due to a few alleles of large effect or many alleles of small effect (WL Chapter 28). This argument goes back to the rediscovery of Mendel at the start of the 1900s, and the ensuing debate on whether evolution was driven by mutation creating new alleles of large effect (the Mendelian view) or by selection acting on small effect alleles over a large number of genes (the biometrician viewpoint). As mentioned in Chapter 1, Fisher formed the field of quantitative genetics by considering Mendelian segregation over a large number of factors. While this framing merged key concepts from the Mendelians (genetics) and biometricians (modern statistics), it did not really answer the question about which effect-size class was evolutionary more important.

As a result, this debate continues to resurface in many forms (WL Chapters 24–28). In the context of human disease, geneticists quickly separated rare diseases (population frequencies of generally less than one in a thousand) from "common" diseases (those occurring at higher frequencies), such as heart disease, diabetes, obesity, and various cancers. Many rare diseases were found to be caused by rare alleles of large effect, often (but not always) by disrupting protein structure (Botstein and Risch 2003). While it was slowly realized that common diseases likely had an important polygenic component (e.g., Carter 1969), there was divergent opinion on their underlying genetic architecture. The **common**

disease/common variant (**CDCV**) hypothesis (Lander 1996; Cargill et al. 1999; Chakravarti 1999) posited that "common" alleles (frequencies $\geq$ 1%–5%) of small effect underlie most of the disease cases. Conversely, the **common disease/rare variant** hypothesis (**CDRV**)—also called the **heterogeneity hypothesis** (Frayling et al. 1998; Bodmer 1999), **rare variant hypothesis** (Bodmer and Bonilla 2008), or **multiple rare variant hypothesis** (Orozco et al. 2010)—posits that rare alleles underlie the disease. Under the rare variant model, disease cases represent a massively heterogeneous collection of rare alleles, each of large effect. It is worth stressing that this debate is much deeper than a simple academic exercise, as its resolution has profound implications on the best strategies for isolating causal factors and represents two very different scenarios for the age, and evolution, of causal alleles. Finally, we stress that rare vs. common is, of course, a false dichotomy, as both sources contribute variation to complex traits and diseases. The real question is what fraction of the variation is accounted for by a given marker-frequency class (either tagged SNPs, or, more ideally, causal sites). Overviews of many of the surrounding issues are discussed by Bodmer and Bonilla (2008), Gorlov et al. (2008), and Schork et al. (2009).

Practical implications of this debate concern both mapping and functional characterization. Considering the latter first, large-effect alleles often result from disruptions in causal gene products (altered protein or functional RNA sequences) or major disruptions in their regulation. As such, they may offer rather direct signals of causal features. In contrast, smaller effect alleles may represent much more subtle perturbations, such as small changes in regulatory features, and may be rather uninformative about the biology underlying a trait. Indeed, as we have seen, it is challenging to associate a significant SNP effect with a particular gene. Thus, even if most variation is from common alleles, the most important signals for a biologist may reside in rare alleles (Momozawa and Mizukami 2012; Chakravarti and Turner 2016; Faraone 2017; Ferraro et al. 2020; Hyman 2020).

As regards to mapping, the key implication of the these two competing hypotheses follows from Example 20.1: the correlation, r^2, between linked alleles rapidly diminishes as their allele frequencies diverge. Hence, the typical marker scored on a SNP chip (usually with a minor allele frequency $\geq$ 0.05) does not tag rare causal alleles, but easily tags common causal alleles. Because rare alleles are hard to tag on SNP chips, they are often imputed (Chapter 20). However, imputation accuracy is a function of the underlying LD, and the quality (and especially size) of the reference population (at least ten copies of the allele must be present; Chapter 20). Hence, the imputation quality for rare alleles is often very poor (e.g., Example 21.15). Methods to accommodate rare alleles (Chapter 20) usually proceed by aggregation of variants (e.g., an excess of rare alleles at candidate genes in cases). If the CDCV model is correct, a high-powered GWAS should tag a very significant fraction of the genetic variance of a trait. If the CDRV model is correct, GWAS would only capture a small fraction of the trait variance, and rare alleles methods must be used.

Whether causal alleles are mainly rare or mainly common also has profound evolutionary implications, as the frequency of an allele is informative about its history (WL Chapters 2, 8–10). Generally speaking, a *common allele is an old allele*, while a *rare allele is typically young*. This is certainly the case under pure drift (WL Equation 2.12), and also true under many forms of selection (WL Chapter 5). Ancient alleles are thus likely no worse than very weakly deleterious, while more deleterious alleles tend to be much younger, and much rarer. While the odd allele or two can quickly rise to moderate frequency by positive selection (such as the sickle cell mutation under a malarial environment), these tend to be exceptions to common alleles being nearly neutral. Hence, the common vs. rare allele debate is also a statement on the nature of selection on a random causal allele for a trait (Examples 21.12 and 21.13).

Another interesting difference between loci harboring common versus rare alleles was noted in several population genetic studies (Hartl and Campbell 1982; Slaktin and Rannala 1997; Reich and Lander 2001). These considered the degree of **allelic identity** at a causal locus. This is the probability that two randomly drawn disease alleles from a locus are IBD, the reciprocal of which has been called the effective number of alleles. For a population

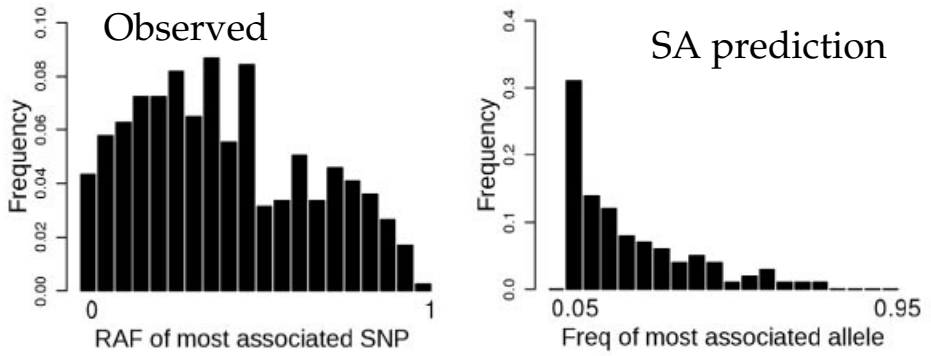

Figure 21.6 **Left:** The observed distribution of frequencies of lead SNPs tagging risk alleles (RAF) for 17 common human diseases. **Right:** This distribution would be highly skewed toward zero (becoming very L-shaped) if synthetic associations between a tag SNP and a collection of rare variants is a common GWAS feature. (After Wray et al. 2011).

expansion following a bottleneck (such as modern humans), a locus whose total frequency of disease alleles is very small shows considerable heterogeneity (very low identity between risk alleles and thus a high effective number of such alleles), while a locus where risk alleles are more common shows much less heterogeneity. For example, Pritchard and Cox (2002) noted that a study of 424 Hemophilia B families found that most cases were caused by distinct alleles: there were 167 distinct Hemophilia B mutations, the most frequent of these only accounted for 5% of all mutations. Conversely the major ΔF508 allele of cystic fibousus accounts for roughly 2/3 of all European CF-contain chromosomes. While there is a predominant CF allele, rare alleles also occur, with roughly 270 different mutations seen among 27,000 CF-harboring chromosomes. This heterogeneity difference between loci with common versus rare alleles is a *transient* feature, reflecting past demography, but it is a signal that can persist for hundreds of thousands of years until the numerous initial copies of the common allele are slowly replaced by new mutations (Reich and Lander 2001). As noted by McClellan and King (2010), the history of human bottlenecks leads to the seemly paradoxical observation that "most human *variation* is ancient and shared, but most *alleles* are recent and rare". Namely, common (and older) alleles likely survived migration bottlenecks and are therefore widely dispersed across human populations. Conversely, most alleles are rare variants, often representing variants not present before the migrations out of Africa, and thus are more regional. If most causal variants are common, they are likely widely shared across human populations, but if they are rare, they may be very region-specific.

A hybrid model in the rare versus common debate was offered by Dickson et al. (2010). Their **synthetic association model** (**SA**) postulated that a chromosomal segment tagged by a common SNP might, by chance, have accumulated a number of rare variants, and, in total, these could generate a detectable SNP marker effect. Given the very low expected r^2 between a common and rare allele (Example 20.1), the resulting rare variants effects would have to be rather substantial to generate a detectable maker effect. Orozco et al. (2010) noted that such large-effect variants, if they exist, could be detected within pedigrees (as opposed to a population sample) by standard linkage methods (Chapter 19). They suggested that the lack of such detectable associations implies that synthetic associations are not a major source of GWAS variance. An ever more direct line of evidence against SA was given by Wray et al. (2011). Their simulations found that the frequency distribution for tagged SNP showing the largest marker effect would be highly skewed toward zero if SA were common (Figure 21.6). This skew arises because a haplotype containing a SA of rare variants harbor several "common" SNPs, with the largest signal being assigned to the least common of these (i.e., the common SNP with the lowest population frequency). This is simply a manifestation of the fact that the largest r^2 (and hence largest signal) occurs for the SNP with the least

marker-causal SNP frequency mismatch (Example 20.1). As shown in Figure 21.6, a survey of GWAS results for 17 common human diseases instead showed a distributed shifted toward intermediate, rather than low-frequency, SNPs. Finally, simulations by Thornton et al. (2013) showed that tagged markers tend to be in strong LD with only a single deleterious mutation of large effect.

Example 21.12 As mentioned, the rare versus common debate is really about the nature of the **allelic spectrum** for a given trait/disease: the joint distribution of the effect size (a) and frequency (x) for causative alleles. The feature that connects a and x is the strength of selection (s) on a given allele, which generates a population distribution, $\phi(x|s)$, of allele frequencies as a function of the nature of selection. A largely unresolved issues is how the effect size a of a new mutation impacts the nature of selection (s) it experiences. Selection could be generated by a *direct* impact of the focal trait on fitness (such as the difference between early versus late onset alleles) and/or by *indirect* pleiotropic effects on other fitness components. A common assumption (e.g., WL Chapters 25 and 28) is that larger effect mutations are expected to have more deleterious pleiotropic effects, resulting in stronger selection against them, generating an inverse correlation between effect size and allele frequency. This is indeed a fairly common observation across humans, maize, and yeast (e.g., Park et al. 2011; Simmons et al. 2014; Wallace et al. 2014a; Zeng et al. 2018; Bloom et al. 2019; Glassberg et al. 2019; Schoech et al. 2019), but one must correct for Beavis effects that overestimate $|a|$ values for rare allele (e.g., Iles 2008). This inverse relationship has important implications for the impact of rare variants.

To see this, consider the situation where the underlying causal alleles are entirely neutral. In this setting, the effect size should be independent of allele frequency. The simplest model for $\phi(x)$ under neutrality is the Watterson distribution (WL Equation 2.34b), which states that the population frequency of sites with minor allele frequency x is proportion is $[x(1-x)]^{-1}$. The resulting additive variance contributed by a site with frequency x is thus expected to be

$$\sigma_A^2(x) \cdot \phi(x) \propto 2a^2 x(1-x) \cdot [x(1-x)]^{-1} = \text{constant} \tag{21.10a}$$

The resulting fraction of the total additive variance for a trait under this model from alleles of frequency $x \leq p$ is thus p, implying that rare alleles ($x \leq 0.01$) only account for one percent of the total genetic variation. For rare alleles to have a much greater impact on the total variance, a^2 must increase as x decreases, and/or *more* rare alleles are present than predicted under the Watterson model. Even for strictly neutral alleles, the latter is true in humans, as the Watterson assumption is a long term stable population size, while populations passing through bottlenecks and subsequent expansion display an excess of rare alleles (WL Chapter 2). Further, selection also inflates the number of rare alleles relative to Watterson. Are these factors sufficient to create a prominent role for rare alleles? At least in humans, models suggest that this is unlikely.

This was nicely illustrated by Zeng et al. (2018) and Schoech et al. (2019), who used a two-component mixture model for the additive effect a of an allele by assuming (for SNP j),

$$p(a_j \,|\, x) = \delta_0 \cdot \pi_0 + N(0, [2x_j(1-x_j)]^S \sigma_a^2) \cdot (1-\pi_0) \tag{21.10b}$$

where $(1-\pi_0)$ is the **polygenicity**, the fraction of all SNPs that impact a trait, the delta function δ_0 denotes a point mass at zero ($a = 0$), and S is a selection parameter. A value of $S < 0$ implies that average a^2 values increase as x decreases (corresponding to negative selection against alleles), while $S = 0$ corresponds to a neutral assumption of no correlation between a^2 and x. MCMC (Appendix 8) can be used to estimate the model parameters S, σ_a^2, and π_0, an approach Zeng et al. called **BayesS**. Note that by rearranging Equation 2.3b,

$$E[a^2 \,|\, x, a^2 > 0] = \sigma^2(a \,|\, a^2 > 0) + (E[a \,|\, a^2 > 0])^2 = [\,2x_j(1-x_j)]^S \sigma_a^2 + 0^2 \tag{21.10c}$$

showing that the variation associated with SNPs with an MAF of x is

$$\big([\,2x_j(1-x_j)]^S \sigma_a^2\big) \cdot 2x_j(1-x_j) = [\,2x_j(1-x_j)]^{1+S} \sigma_a^2 \tag{21.10d}$$

Both Zeng et al. and Schoech et al. considered over two dozen, largely non-overlapping, traits/diseases from the UK Biobank. Zeng et al. found that all but one of their traits had a negative estimate of S (ranging from -0.609 to 0.012), 24 of which were significantly negative, with a median S of -0.37. The polygenicity $(1-\pi_0)$ had a median value of 5.4% and ranged from 0.6% to 14.0%. Schoech et al. obtained very similar results for S. Substituting these S values into various population genetic models for $\phi(x|s)$ showed that no more than 10% of the variance could be due to rare alleles ($x \leq 0.01$).

A more recent study by Zeng et al. (2021) used a summary-statistic version (**SBayesS**) of BayesS, and examined a much larger set of traits, finding almost all had significantly negative estimates of S (with a median value of -0.6). They estimated that, on average across traits, around 1% of the human genome are mutation targets ($1-\pi_0 \simeq 0.01$) for that trait, with a mean selection coefficient for segregating variants of -0.0007. Relative to other traits, common diseases showed a smaller mutational target size but stronger selection (π_0 smaller, $|S|$ larger). More granularity can be applied to values of S by examining its value over different functional classes of variants (i.e., different annotations), such as nonsynonymous mutations, methylation sites, etc. For example, Gazal et al. (2018), examining 27 traits (from the UK Biobank), found an average value of $S = -1.10$ for nonsynonymous variants and $S = -0.30$ for other variants.

A final complication in the effect size-MAF relationship is that causal alleles in regions of lower LD (relative to the value expected given their MAF) tend to be both more recent and to also have larger effects (Gazal et al. 2017). Again, this is expected if most trait alleles are under weak negative selection. We examine this, and other, complications further in Chapter 32 in a deeper discussion of estimation, and interpretation, of SNP heritability (the fraction of variation accounted for my SNP effects).

Example 21.13 There is a very detailed, and highly technical, literature from evolutionary genetics on the maintenance of quantitative genetic variation (reviewed in WL Chapter 28), and a largely independent corresponding literature that focuses on polygenic risk variants for human diseases (e.g., Pritchard 2001; Pritchard and Cox 2002; Di Rienzo and Hudson 2005; Peng and Kimmel 2007; Eyre-Walker 2010; Maher et al. 2013; Simons et al. 2014, 2018; O'Connor et al. 2019). While mathematically impressive, most of these models are *very fragile*, in that apparently trivial changes in assumptions can yield quantitatively different predictions. In particular, the choice of modeling how pleiotropy generates correlations between effect size (a) and selection (s) has profound implications. An excellent overview of these issues is given by Johnson and Barton (2005), with a more GWAS-focused review by Sella and Barton (2019), and much more detailed discussion in WL Chapter 28.

One common modeling assumption for maintenance of variation is **stabilizing selection** on our focal trait (fitness declines as one moves away from some optimal value; WL Chapters 5, 16, 29, 30) countered by mutation. If, on a per-locus basis, the strength of selection is weak relative to mutation (the **Gaussian assumption**), the resulting equilibrium distribution of allelic effect sizes is normal, with many small-effect alleles occurring at intermediate frequencies (the common variant model). Conversely, if selection is strong relative to mutation (the **house-of-cards assumption**), we recover the rare variant model, as the resulting distribution is leptokurtic (Chapter 2), with rare alleles of large effects (see WL Table 28.2 for other differences between these models).

As mentioned, when pleiotropy occurs (the variant impacting our focal trait also influences others), the expected outcomes become very model dependent. Simons et al. (2018) developed a model allowing for extensive pleiotropy, wherein a variant impacting a focal trait (with effect a_1) also influences a number of other traits (i.e., generates a vector $(a_1, \cdots, a_n)$ of effects for the n traits it impacts), with the vector of trait values under multivariate stabilizing selection. In this setting, deviation from the optimal value in any direction reduces fitness (WL Chapters 29 and 30), with surfaces of equal fitness satisfying $s = \sum a_i^2/V_s$ (where V_s is the strength of stabilizing selection). Using this assumption, they made a very clever use of Fisher's geometric model for adaptation (WL Chapter 27) to obtain a distribution of effect sizes (a_1) at the focal trait for a mutation with a given selection coefficient, which for a large effective number of traits ($n_e > 10$) is approximately normal, with $a_1 \,|\, s \sim N(0, sV_s/n_e)$. Applying their model to height and BMI data, they estimated that the fraction of variation

arising from alleles of detectable effect sizes is much greater for height (50% of the total variation in height should be detectable using current GWAS sizes) than for BMI (only 15% of the variance should be detectable). Further, they estimated that the total mutation rate for height variants was around five times that for BMI variants.

Distribution of Allelic Effect Sizes

While theoretical models (Examples 21.12 and 21.13) make predictions about the distribution of allelic effect sizes, what do the data say? A number of approaches have been used to estimate the effect-size distribution, starting with either a single exponential or gamma distribution (Otto and Jones 2000; Hayes and Goddard 2001; Weller et al. 2005), and moving to mixture distributions, typically based on weighted sums of of normals (Chapter 16). As detailed in Chapters 31 and 32, the use of mixtures traces back to attempts by animal breeders to use marker data to predict breeding value (Meuwissen et al. 2001). The basic data used are typically (but not always) GWAS summary statistics, such as the slope β of the gene-dosage regression (Equation 20.1a) or its logistic regression counterpart (Equation 20.3b), which is essentially an estimate of the allele effect a. One then fits the observed values of β with a mixture model. For example, the **BayesR** method of fits a point mass at zero (δ_0) and then three normals with increasingly smaller variances (Erbe et al. 2012), corresponding to allelic classes with increasingly smaller effects,

$$p(a_j) = \pi_0\delta_0 + \pi_1 N(0, \sigma_g^2 \cdot 10^{-2}) + \pi_2 N(0, \sigma_g^2 \cdot 10^{-3}) + \pi_3 N(0, \sigma_g^2 \cdot 10^{-4}) \qquad (21.11a)$$

Under this model, a fraction π_0 of all SNPs have no effect, so that $\pi = (1 - \pi_0)$ is the polygenicity. The π_i and the variance parameter σ_g^2 can be estimated by either ML (Appendix 4) or MCMC (Appendix 8). Holland et al. (2020) coined the term **discoverability** for the effect size variance (essentially, the expected power of detecting variants drawn from that size class), with effects being drawn from the last normal in Equation 21.11a having the lowest discoverability. Note that Equation 21.11a, unlike the models given by Equation 21.10, does not condition the effect size on allele frequency. This is because it effectively assumes the Watterson distribution, so that the expected contribution to the variance from alleles with MAF x is independent of x (Equation 21.10a). We return to this assumption in Chapters 31 and 32.

Moser et al. (2015) applied BayesR to seven of the diseases studied in the WTCCC (Chapter 20): bipolar disorder (BD), coronary artery disease (CAD), Crohn's (CD), hypertension (HT), type 1 and 2 diabetes (T1D, T2D), and rheumatoid arthritis (RA). With the exception of T1D, the other six diseases had an estimate of over 60% of their total variance due to alleles drawn from the smallest-effect size (π_4, corresponding to a variance of $\sigma_g^2 \cdot 10^{-4}$), with four of these diseases having over 75% due to this class: CD (94%), HT (88%), CAD (84%) and T2D (78%). The estimates for T1D were roughly 25% of its variance due to alleles drawn from this category, while almost 72% of the variance was attributed to the largest effect group (π_2, variance of $\sigma_g^2 \cdot 10^{-2}$).

Similar mixture models have been examined by Thompson et al. (2015), Zhang et al. (2018), Holland et al. (2020), and O'Connor (2021). All concluded that traits are massively polygenic (normally with at least tens of thousands of estimated causal sites), and that GWAS sample sizes in the hundreds of thousands to millions are required to capture the bulk of the heritability. For example, Zhang et al. used a model with point mass at zero, and then either one or two normals (respectively, their **M2** and **M3** models),

$$a_j \sim \pi_0\delta_0 + \pi_1 N(0, \sigma_1^2) + \pi_2 N(0, \sigma_2^2) \qquad (21.11b)$$

with the M2 model not including the last term. M3 generally performed better, and was used to estimate the effect-size distribution for 32 traits. They found that most of the variance was due to small-effect alleles, with the required GWAS sample size to capture 80% of

the heritability ranging from a few hundred thousand to several million. O'Connor (2021) noted that the range of effect size between the 10th and 90th percentiles was around a 100-fold difference, and around 600-fold between the 5th and 95th percentiles. Based on these estimates, genetic architectures are massively polygenic, with a vast range of effect sizes.

Example 21.14 O'Connor et al. (2019) noted that the measure $\pi = (1 - \pi_0)$ for polygenicity is a bit misleading, and a more general metric would be how even effects are distributed over causal loci (essentially, an effective number of causal sites). They noted that schizophrenia is underpinned by thousands of small-effect common variants, while Niemi et al. (2018) found that rare severe neurodevelopmental disorders, which are largely expected to be monogenic (cause by rare, highly deleterious mutations), still have around 8% of their variance explained by common alleles of very small effect. Polygenicity as measured by the fraction (π) of SNPs with an impact on the trait would be roughly the same in both these settings, and yet clearly their architecture are very different (polygenetic for schizophrenia; Mendelian, with variable penetrance, for neurodevelopmental disorders). O'Connor suggested that a better polygenicity measure is the *evenness of effect sizes* over the causal sites, which would be high with schizophrenia and very low (dominated by a few loci) with neurodevelopmental disorders.

The contribution of allelic effect to the variance scales as β^2, and the variance of these variance effects scales as β^4. If M SNPs (trimmed for independence) are scored, the expected number that are causal is $M\pi$, where we will now refer to π as the **polygenicity fraction**, while **O'Connor metric** for the effective number of causal SNPs is

$$M_e = M\,\frac{3}{\kappa_4}, \qquad \text{where} \quad \kappa_4 = \frac{E[\beta^4]}{(E[\beta^2])^2} \tag{21.12}$$

where κ_4 is the scaled kurtosis of the β and equals three when β follows a normal distribution (Chapter 2). We denote Equation 21.12 as the **polygenicity kurtosis**. Note that only a fraction π of the β are different from zero, and if these effects follow a normal distribution, then $3M/\kappa_4 \simeq \pi M$, the same number of causal SNPs predicted using the polygenicity fraction. O'Connor et al. developed an extension of LD regression (Chapters 20 and 32), **stratified LD fourth moment regression** (**S-LD4M**), to estimate M_e.

Applying their estimator to 33 traits (with an average GWAS sample size of 360,000), they found that M_e based on just considering common SNPs (MAF $> 5\%$) had effective number of causal sites ranging from 500 to 30,000 (with a median value around 3000). When low frequency SNPs (MAF from 0.5% to 5%) were used, the estimated M_e values were about 25% of the common M_e values. For example, height had around 3600 effective causal sites for common alleles and 800 effective sites for low-frequency alleles. Hence, common alleles had a more even distribution of effect sizes that low-frequency alleles.

Genetic Architectures and "Missing" Heritability

The decade following the first GWAS was period of both considerable excitement and a good deal of consternation. Despite billions having been spent globally on extensive GWAS studies and the corresponding development of genomic resources, most detected hits were of small effect, largely in noncoding regions, and accounted for only a tiny fraction of the trait heritability. This angst culminated with the highly cited paper of Manolio et al. (2009) on the problem of **missing heritability** not detected by the GWAS studies of the time. For example, human height has a highly repeatable heritability estimate (from resemblance between relatives) of around 0.7–0.8, and despite over 40 significant GWAS hits for height (in 2009), these accounted for only about 5% of the phenotypic variance (a little over 7% of the expected heritability). As Manolio et al. summarized, this was the standard observation over a number of diseases and traits, raising the question of what generated this gap between GWAS-based and relative-based estimates of heritability (the former often called the **SNP heritability**; Chapter 32). More formally, the gap was between

the additive variance estimated from SNP effects, $\sum a_i^2 p_i(1-p_i)$ (the sum being taken over all significant SNPs, which can be modified for LD), and value of σ_A^2 estimated from standard relative-based designs (Chapters 22–31).

Proposed explanations for this gap fell into three, not necessarily exclusive, categories. First, this was simply an issue of power with the GWAS designs at the time. Second, there were genetic features not fully captured by a GWAS, such as the effects of CNV, epigenetics, and perhaps other, yet to be discovered, phenomena. The final explanation was that the GWAS results were indeed correct, with the gap arising because relative-based estimates of h^2 were systematically (and dramatically) inflated. A number of different viewpoints were offered on these various resolutions (e.g., Eichler et al. 2010), generating a considerable number of papers, some thoughtful, some speculative, and some rather ignorant of basic features of quantitative genetics.

As we have hinted throughout this chapter, it now appears that the correct resolution is lack of power. Before developing this point, we first discuss overestimation of h^2 by relative-based methods, which has some grain of truth. As detailed in Chapter 7, estimation of the additive variance follows by partitioning the phenotypic covariance for given a pair of relatives in various sources of genetic variation (Equation 7.12) plus for any shared environmental effects. When relatives can only share (at most) single IBD alleles, the shared genetic variance is dominated by σ_A^2, with potentially much smaller contributions for additive epistatic sources (e.g., $\sigma_{AA}^2, \sigma_{AAA}^2$, etc.). For relatives that can share two alleles IBD (e.g., full sibs, monozygotic twins), dominance (and dominance epistatic terms) can also enter into the genetic correlation. Zuk et al. (2012) correctly noted that if these nonadditive terms were sufficiently large, relative-based estimates of h^2 can be significantly inflated. However, these nonadditive genetic components (i.e., those other than σ_A^2) are generally expected to contribute very little to the resemblance between relatives for two reason. First, the coefficients on nonadditive variance terms are much less than the coefficient on the additive variance (Equation 7.12). Second, and perhaps more fundamentally, in most segregating natural populations, the nonadditive variance components themselves are expected to be much smaller than the additive variance. This is simply a consequence of most genetic variation loading onto the additive component when the minor allele is uncommon (Hill et al. 2008; Hill 2010; Mäki-Tanila and Hill 2014). As the Watterson distribution shows, this is expected to be the case in natural populations. For example, Zhu et al. (2015) found that SNP-based estimates of dominance variance were small over a set of 80 human traits, and did not contribute to much of the missing heritability. Similar results were obtained by Hivert et al. (2021) using 70 traits and a sample size of over 250,000 (an average SNP heritability for additive effects of 0.208 and an average value of 0.001 for dominance). The exception to additive variation overpowering other terms is in segregating populations formed by line crosses, where segregating alleles frequencies (by construction) are 0.5 (Examples 21.16 and 21.17), resulting in more of any underlying nonadditivity mapping into nonadditive *variance* components.

Relative-based estimates of the additive variance can also be inflated by shared environmental effects among close relatives. Two different approaches can be used to estimate additive variation while minimizing any impact from common environments: use extended pedigrees so that shared environmental effects are likely more diffuse than within a single family (Chapters 31 and 32) or exploit the variation in IBD sharing among the same set of relatives caused by Mendelian segregation (realized versus pedigree kinship; Example 8.2, Chapters 8 and 32). A pair of sibs that, by chance, shares more IBD alleles should be more similar than another pair that, also by chance, shares a smaller fraction of IBD alleles, with a regression of their squared difference on amount of sharing estimating the additive variance. This approach was developed for sib pairs by Visscher et al. (2006) and extended to general sets of relatives by Young et al. (2018). By contrasting IBD differences between sibs (or more general relatives) in the same family (kinship), shared environmental effects are controlled. Both of these approaches have been applied to human traits. Zaitlen et al. (2013), in an analysis of 23 traits in an large Icelandic sample (38,000) with deep pedigrees,

found find that h^2 estimates based on close relatives (parent-offspring, sibs) were somewhat inflated relative to estimates based on more distant relatives (and hence less likely to share environmental features). For most traits, the inflation was modest (about 15% for height), so that while it does slightly close the gab between GWAS and relative estimates, a very substantial difference remains. Similar results (relative-based estimates of σ_A^2 were slightly to modestly inflated) were seem when estimates of additive variance based on within family segregation were compared to more traditional estimates (Young et al. 2018; Young 2019; Kemper et al. 2021).

The second explanation for missing heritability—namely, features not captured by a GWAS—is both speculative and increasingly less viable as ever-larger GWAS have closed much of the heritability gap (Example 21.15). However, there is still debate on what is causing the lack of power: common alleles with increasingly smaller effects, or increasingly rarer alleles of large effect. As GWAS sample sizes increases, ever smaller additive variances, $r^2 2p(1-p)a^2$, associated with a marker can be detected, where r^2 is the LD with the tagged marker and p the causal allele frequency. For a rare allele, this is $\sim r^2 2a^2 p$ when the marker is scored directly and $\sim r_{imp}^2(r^2\, 2a^2 p)$ when the marker is first imputed (where r_{imp}^2 is the correlation between the actual marker allele and its imputed value). For a common allele, the additive variance tagged by a marker approaches $a^2/4$. This arises as the MAF approaches 0.5, in which case r^2 likely approaches one as most common SNPs are well tagged. For a rare causal allele r^2 (with a common marker SNP) is still expected to be small, being a decreasing function of the causal allele frequency (Example 20.1). It is this expected small value of r^2 that motivated the synthetic association model, as only a tiny fraction of the true variance of large-effect, but rare, alleles would be captured by common SNPs (underestimating their contribution by $1/r^2$). This would be a 10-fold effect for $r^2 = 0.1$ (which would still be a relatively high correlation LD for a rare allele). However, as we have detailed, the data do not support this model (e.g., Figure 21.6).

What impact might whole genome sequencing (WGS) have on the search for missing heritability? The answer depends on whether the heritability gap is due to common alleles of very small effect or very rare alleles of large effect. The impact of WGS is to improve the accuracy of tagging, namely increasing r^2 to one. Common causal alleles are either already captured directly, or tagged with very high accuracy by other common alleles in high LD, so that WGS is unlikely to increase tagging efficiency (Caballero et al. 2015). In this case, only increased sample size (rather than scoring more markers) can generate the increase in power needed to detect common alleles of very small effect.

Rare alleles are a potentially different matter. If rare alleles are accurately captured with high efficiency by imputation from the scored SNPs ($r_{imp}^2 \simeq 1$), then WGS, by itself, does not really improve power. However, if very rare alleles are poorly imputed, for example by residing in regions of low LD, then WGS results in a significant increase in power. This was indeed seen for height (Wainschtein et al. 2022; Example 21.15). One caution is that the use of very rare alleles introduces the potential for subtle biases. In particular, as Chapter 20 stressed, corrections for population structure are critical, especially when searching for very small allelic effects in a very large GWAS. Most such corrections are based on common (and thus older) alleles, with the population structure for rare (and thus younger) alleles being potentially rather different. For example, individuals sharing a very rare allele are likely to have shared a recent relative, and thus the potential of sharing a recent common environment (Young 2019).

Human height is one of the best studies traits, and offers considerable insight (and history) into the missing heritability debate. As Example 21.15 highlights, the bulk of the initial missing heritability was due to common alleles of small effect, but very recent whole-genome sequence data finds that a substantial proportion is also due to very rare alleles of large effect in low-LD with common SNPs.

Example 21.15 Human height is truly the ultimate quantitative trait. Galton's work on the

resemblance in height between parents and their adult offspring (Galton 1886; Figure 1.1) was instrumental to Fisher's 1918 founding of quantitative genetics. Height has been the subject of a number of ever-larger GWAS projects, and the history of these studies nicely tracks the journey from the beginning of the GWAS era, to navigation through the rough seas of missing heritability, and into the whole genome sequencing (WGS) era.

While the most widely quoted figure for the heritability of height is 0.8, Yang et al. (2015) suggested that this is a slightly inflated estimate due to shared environmental effects among relatives (typically parent-offspring or sibs). They noted that Hemani et al. (2013) obtained an estimate of $h^2 \simeq 0.7$ based on IBD variance between pairs of sibs. This matches the estimate of 0.7 by Zaitlen et al. (2013) using extended pedigrees. A more recent study by Kemper et al. (2021) also obtained an IBD-sharing-based estimate of around 0.7 using full sibs, but a larger estimate (0.8) using a more general set of relatives. Hence, in the following discussion we will use the 0.7 value as a *lower bound* for the h^2 that must be explained by markers.

Gudbjartsson et al. (2008) examined 40,000 individuals (mostly of European descent) and discovered 27 genomic regions containing one or more variants associated with height. The estimated effect sizes ranged from 0.3 to 0.6 cm and together explained around 4% of the phenotypic variance in height (6% of h^2, assuming a value of 0.7). By two years later, Lango Allen et al. (2010) had detected almost 200 loci achieving genome-wide significance in a sample of 180,000, which accounted for around 14% of the heritability. One obvious source of missing heritability was the exclusion of markers with nontrivial effects, but which had effect sizes too small to meet the stringent genome-wide significance threshold. Lango Allen examined this issue by considering SNPs that were of nominal, but not genome-wide, significance. When included, the fraction of explained heritability rose to 23%. Four years later, Wood et al. (2014) had found 700 variants with genome-wide significance in a sample of 253,000, accounting for 20% of the heritability. Removing the restriction to SNPs of genome-wide significance, the most strongly associated 2000, 3700, and 9500 SNPs accounted for 26%, 30%, and 36% of the heritability. Yengo et al. (2018) examined a meta-analysis with close to 700,000 individuals, finding 3300 independent marker effects that accounted for roughly 35% of the heritability. Extrapolating from estimates of the effect-size distribution, Zhang et al. (2018) predicted that a sample of over a million would be required to account for 80% of the heritability. How good was this prediction? Yengo et al. (2022) extended their meta-analysis to over 5.4 million individuals, largely European, but with other ancestries as well. They found that 12,000 independent common SNPs (located in 7200 distinct genomic regions, ranging from 70 to 700kb, covering around 20% of the genome) accounted for 40% of the phenotypic variance ($\simeq$ 60% of h^2), falling far short of the expected value projected from estimates of the effect size distribution. Over 30% of these regions were more than 50 kb away from any known genes.

The clear trend is that the expected number of detected sites increases with sample size, with the roughly 135-fold sample size increase from 2008 to 2022 resulting in a roughly 120-fold increase in genome-wide significant markers, a 270-fold increase in the number of genomic regions, and a ten-fold increase in the amount of heritability assigned to markers. In the words of Yengo et al. (2018), "increasing GWAS sample sizes continues to deliver." However, Yengo et al. (2022) suggested that GWAS saturation for common alleles may have been reached, as they predict most newly discovered variants will be restricted to the 7200 genomic regions found in their 5.4 million GWAS meta-analysis. Holland et al. (2020) estimated the number of causal sites as roughly 95,000 (based on estimates of π), with a similar estimate by Boyle et al. (2017a). Because these studies used common SNPs as markers, these effects are, in large part, due to common causal alleles, and thus of relatively modest effects. As noted in Example 21.14, O'Connor et al. (2019) suggested that a better estimate of the polygenicity is the effective number of sites (the more equal the effect size distribution, the larger the effective number). They estimated that height had 3600 effective causal sites for common alleles (MAF > 0.05) and 800 effective sites for low-frequency alleles (MAF of 0.005 to 0.05). They offered no estimate for very rare alleles (MAF < 0.005).

An alternative to assigning effect sizes to individual markers (fixed effects modeling) is to treat allelic effects as random and estimate the total amount of variation they explain. Using common markers, Yang et al. (2010b, 2011b, 2015) applied this approach to samples of 4000 and 44,000, with the marker variance accounting for 62% and 80% of the heritability. Hence, the variance component approach is more powerful for detecting missing heritability, as it includes information from *all* markers, not just those declared significant by some criteria.

However, note that the 12,000 variants of Yengo et al. (2022) accounted for over 85% of this common-SNP based estimate of h^2. An especially interesting observation was made by Yang et al. (2011b), who noted that the length of a chromosome was very highly correlated with the amount of variation explained by markers on that chromosome (a similar observation was seen for schizophrenia by Lee et al. 2012). Hence, height-associated markers are, to a first approximation, randomly and uniformly, scattered throughout the genome. However, the more recent analysis of Yengo et al. (2022) found that, while widely scattered, only about 20% of the genome contained regions haboring height variants. Loh et al. (2015b) similarly inferred that over 70% of 1-MB genomic regions in humans harbor at least one causal site for schizophrenia risk.

The variance component approach can also be used with WGS data. Wainschtein et al. (2022) estimated the SNP-associated variance in a sample of 25,500 fully sequenced individuals, considering all SNP alleles present in at least five copies in the sample (a MAF of 0.0001, $\sim$ 900,000 SNPs). Following Yang et al. (2015), they used a variance component model (VC) and estimated a SNP-based heritability of 0.48 (68% if h^2 = 0.7). They then used this same set of SNPs to impute rare alleles, with the heritability estimate using this enlarged set (common plus imputed SNPs) ranging from 0.50 to 0.56 (depending on the imputation model). Thus, including imputed alleles resulted in a slight improved of the amount of explained variation. They then use WGS for directly score rare alleles, and the resulting heritability estimate improved significantly to 0.7. Most of the improvement came from very rare variants ($0.0001 <$ MAF < 0.001) that were in low LD with common SNPs, and hence with very low imputation accuracy. Thus, while common variants of small effects recover about 2/3 of the relative-based heritability, very rare variants account for the majority of the remaining difference.

Example 21.16 A very interesting, and somewhat different, perspective on missing heritability is offered from the sibship analysis of segregating yeast crosses (which serves as a good representative for any line-cross QTL experiment). Example 21.1 discussed the genetic architecture of yeast eQTLs, while physiological traits (colony growth size under different environmental features) have also been examined (e.g., Ehrenreich et al. 2010; Bloom et al. 2013, 2019) and we focus on the latter. Bloom et al. (2013) examined 46 growth-related traits in just over 1000 haploid segregant lines from a cross between a laboratory and a vineyard strain. Taking this collection of lines (sibs) as our population, the additive variation within this sibship can be estimated using the variation in IBD sharing among sibs (Chapters 8 and 32). Bloom et al. used this approach to estimate narrow-sense heritability values. Broad-sense heritability can be estimated from the between-line variance (Chapter 24), with the difference between the two heritabilities being a measure of the impact of any nonadditive variance (these would only be additive epistastic terms, because the lines are haploid). They then mapped QTLs and examined the fraction of the narrow sense heritability they recovered, which ranged over traits from 72% to 100%, with a median value of 88%. By considering ever-larger subsets of the lines, they showed that this fraction increased with the number of lines (the analog to accounting for more of the missing heritability by increasing a GWAS sample size). They also found that nonadditive variance was significant, with the difference between narrow and broad sense heritability ranging from essentially zero to roughly 50%, depending on the trait. Two-locus epistasis (e.g., $A \times A$) explained only a tiny fraction of this difference.

By construction, there are no rare alleles in the collection of segregating lines, as all have expected frequency 1/2. Bloom et al. (2019) expanded this study by considering multiple crosses, namely 14,000 segregants from all pair-wise crosses between 16 diverse lines (so that the rarest allele frequency is expected to be around 6% in the resulting synthetic population). Again, detected QTLs accounted for the vast major of the estimated narrow-sense heritability (median value of 68%). However, Bloom et al. noted that rare alleles (those alleles segregating in the lines, but found at frequencies < 0.01 in a 1000-isolate reference collection) accounted for a disproportionate amount of the additive variance. Such alleles were about 28% of all variants, but their median total contribution the additive variance was 52%. Note that this does not necessarily provide support for rare alleles having a large impact in a standard GWAS, as their frequencies were artificially high due to the population construction, but does indicate that rare alleles tend to have much larger effects.

Example 21.17 While much of our focus on massive GWAS studies, rightly so, has been on those from human genetics, there is also an impressive list of results from maize (Buckler et al. 2009; Brown et al. 2011; Kump et al. 2011; Poland et al. 2011; Tian et al. 2011; Wallace et al. 2014a, 2014b; Xiao et al. 2016, 2017). Wild maize lines have LD that typically decays in under 2kb, offering much higher mapping resolution than in humans. This feature, coupled with the NAM line collection (Chapter 18), makes maize a powerful system for fine mapping. The NAM lines are a set of 200 RILs from each cross of 25 maize lines (globally sampled) to a common parent (B73), capturing the variation present over these 26 diverse lines (Figure 18.12). The resulting set of 5000 inbred lines can be highly replicated, offering a considerable increase in power (estimating the phenotypic value of a genotype by the average over a series of clones versus from a single observation for an outbred individual). A typical NAM GWAS design scored nearly a million plants in four different growing locations over two years (8 environments). Traits examined include flowering time, leaf architecture, pathogen resistance, and inflorescence features (male tassels and female ears). Ignoring inflorescence traits, most traits were dominated by additive alleles of small effect with little detectable epistasis, with most ($> 80\%$) of the additive variance captured by SNPs. G x E was observed, but the variance it accounted for was much less than the variance from additive effects. Allelic heterogeneity was common, with alleles replicated over some, but not all, families, and different variants segregating at the same genomic location over the collection of families (often with opposite effects with respect on the B73 reference allele). Detected QTL were sometime enriched for known candidate genes, but most QTL do not appear to be associated with any known candidates (loci at which major-effect alleles have been detected). The genetic architecture of inflorescence traits was bit different from the other traits in that they tended to have much larger effects, with ear effects being larger than tassel effects. Pleiotropic loci were detected that controlled both ear and tassel elongation, again with the ear effects being larger. Brown et al. (2011) suggested that this architectural difference for ears resulted from recent selection on ear morphology during domestication, favoring fixation of large-effect alleles (WL Chapter 27).

Beyond the Infinitesimal Model

The emerging view of quantitative trait variation is that it is massively polygenic, with the average variance attributable to any causal site being very small, and with a range of allelic-effect sizes spanning several orders of magnitude. This pattern is similar to, but subtly different from, Fisher's **infinitesimal model**, in which a very large number of loci, each of very small effect, underlies trait variation (WL Chapter 24). While *allelic effect sizes* do not follow the infinitesimal model, the behavior of the *variance components* associated with the underlying sites is much closer to Fisher's characterization. The small range in the per-site variances (relative to the much larger range of underlying allelic-effect sizes) is a result of the inverse correlation between effect size and allele frequency. This association, best considered as more of a trend than an absolute rule, appears to be at least partly driven by larger-effect alleles generally being more deleterious than smaller-effect ones.

While we may have some mild clarity on the evolutionary underpinning of trait variation, two apparently perplexing observations pose deep questions for its underlying molecular basis. The first is the massive number of causal sites. For example, estimates of the number of sites impacting height are up to three times the total number of protein coding genes in the human genome (Example 21.15). Second, almost all variation appears to map to noncoding regions. How can we account for these observations? Clearly, a central issue in the modeling of quantitative variation is the role of gene regulation.

The early days of molecular genetics saw a transition from a focus on simple prokaryotic gene switches (such as the *lac* operon in *E. coli* or the *cro-c1* switch in phage λ) to more elaborate models for eukaryotes where gene regulation was set within a far wider net of players (e.g., Britten and Davidson 1969). As high throughput functional genomics tools

became available, it seemed that most features in a cell were connected to one another. Cellular networks are typically structured as **domains** (or **modules**) of highly interconnected players, separated from other such domains by a few tenuous, but none the less present, connections. More formally, many biological networks appear to be **small world** (Appendix 2), meaning that one can connect any two nodes (or more generally, local hubs organizing subnetworks) via a very small number of steps. Further, as we have seen, there are layers of regulatory elements that can act in much more complex, and far subtler, ways than those envisaged in the early days of simple gene switches with just a few discrete components and sites of action. Thus, it became apparent that trait variation is largely determined by the actions of very complex regulatory networks (e.g., Chakravarti and Turner 2016). One interesting complication to this network view of quantitative variation was noted by He (2017): the structure of biological networks (being small world and scale free; Appendix 2) can generate considerable homeostasis. Hence, networks are both sufficiently robust to perturbations, but also loose enough to allow a certain amount of variation to leak through.

It is against this backdrop that Pritchard and colleagues (Boyle et al. 2017a, 2017b; Liu et al. 2019) proposed their **omnigenic model** (omni = all). Their tenet that "the connected-network aspect of the omnigenic model is a parsimonious model that can potentially explain the major observations" was simply a codification of the widespread belief that one must understand networks in order to understand quantitative variation. Their key innovation was to consider the network generating trait variation as consisting of a limited number of **core genes**, modified (in *trans*) by a much larger set of **peripheral genes** that have very slight individual, but very strong collective, regulatory impacts on the core genes. Under this model, *any* gene that shows regulatory variation in a tissue that impacts the focal trait can potentially have a small, but nontrivial, effect (hence the term "omni").

While initially a bit nebulous as to what exactly corresponded to a core gene, a more formal definition based on the concept of mediation was proposed by Boyle et al. (2017b) and Liu et al. (2019). Core genes directly impact the trait (for example, by coding for QTT), while peripheral genes impact the regulation of these traits via mediation. Hence, conditional on core genotypes and expression levels, a trait is not further impacted by the genotypes or transcripts at any peripheral gene. However, as remarked by Liu et al., "it is important to note that our definition of core genes is a simplification of a more complex reality," who further noted that the real operational key to the omnigenic model is that "trait heritability is mainly driven by peripheral genes that *trans*-regulate core genes." Under this model, most large-effect (and therefore rare) variants are from the core genes, being either strong *cis* acting sites (such as eSNPS and sSNPS) or structural variants, while the more numerous (and of smaller effect) common allele sites almost entirely associated with the peripheral gene set. This view was slightly modified by Liu et al. (2019) who noted that a master regulator peripheral gene (co-regulating a number of the core genes) could have a large effect via the summation of a number of small effects over each impacted gene. A more detailed look at the impact of core genes was offered by Sinnott-Armstrong et al. (2021), who examined the genetic architectures of three traits (urate, IGF-1, and testosterone) whose major pathway components are fairly well defined. While they found that the lead SNPs were highly concentrated in core genes, only 10-20% of SNP-based h^2 was due to variants in core pathway components. Estimates of the total number of causal sites ranged between 4000 (testosterone) and 12,000 (urate).

Boyle et al. (2017a) hypothesized that the omnigenic model offered a useful search strategy to better characterize quantitative variation, by focusing on detecting variants in the core genes. They suggested whole exome sequencing in an attempt to detect rare, but functionally informative, mutations in these genes. Such variants, for the most part, would identify at least some of the core genes, which may provide useful biological insight. However, Wray et al. (2018) noted that a key component in GWAS-based functional characterization is always sample size, and that the every-increasing sample sizes of GWAS can both detect rare alleles while also furthering our understanding of sites with small effects. They suggested a better use of resources is to simply increase GWAS sample sizes in general, rather than

amplifying a specific class of sites (e.g., whole exome sequencing).

As discussion about the omnigenic model highlights, the next generation of quantitative genetics methodology will heavy focus on biological networks. Some of the machinery offered in this chapter will no doubt be massively extended in the near future to provide biological, computational, and analytic tools to continue exploration of these structures.