Section 1 General principles of genetics and genomics

Chapter 1

Linkage and associations

Elizabeth J. Rossin and Benjamin M. Neale

Introduction

Human genetics is one of the most promising approaches to identifying the cellular underpinnings of human diseases and traits. For diseases whose etiology is largely unknown, identifying genes that contribute risk can lead to novel biological insights and potentially reveal proteins and pathways to target with therapeutics. Historically, the search for such genetic variation that influences phenotype has been particularly successful in rare genetic disorders, termed Mendelian disease, that are caused by severe mutations in DNA: classic examples of such diseases include hemochromatosis, cystic fibrosis and phenylketonuria [1]. For these diseases, DNA changes in particular genes lead to deficient or altered protein that in turn results in a cascade of physiological outcomes, ultimately culminating in the medical sequelae that define the disease. Not only have these findings helped elucidate the biological pathways important to these phenotypes, but also understanding the damaged cellular processes has been proven to be relevant to patients' medical treatment. A primary goal of human genetics is to understand disease biology and ultimately aid in the identification of novel therapeutic design.

The application of genetics to severe rare diseases that follow clear inheritance patterns in families has led to the successful identification of the root cause in many instances. These Mendelian diseases are almost completely caused by genetic factors, which explains the success of genetics to unequivocally determine the cause. In contrast, complex traits are characterized by the combination of many genetic and environmental factors that together create the phenotype. An additional consequence of this complex trait architecture is that the familial clustering of the trait does not follow a clear and predictable inheritance pattern. For most complex phenotypes, we do not understand the bulk of the underlying pathophysiology, in spite of the fact that many of these traits are clearly heritable. Since the nineteenth century, scientists and physicians have studied twins and families for complex phenotypes and identified clear evidence of heritability. The fact that traits tend to run in families and that more genetically similar family members tend to be more phenotypically similar provides empirical support of the genetic hypothesis. Consequently, the identification of genetic variants is possible and provides the opportunity to gain insight into the biological processes relevant to human disease. Twin and family studies in sleep phenotypes have revealed significant heritability; the earliest observation of sleep phenotypes being heritable was made in 1937 when Geyer reported higher sleep profile concordance in monozygotic twins than dizogotic twins [2].

As with many traits, the majority of sleep disorders and sleep-related traits are complex phenotypes. However, there are some examples of familial diseases that present with disordered sleep as either a primary or secondary finding. Phenotypes in both these categories include diseases such as restless leg syndrome (RLS) and narcolepsy–cataplexy as well as quantitative traits in normal individuals including duration and quality of sleep. A number of instances of sleep disorders segregating in a Mendelian fashion within large families have been documented, but there are also well-established studies of heritability of sleep and sleep disorders as complex traits as discussed later in this chapter [3–9].

Identifying genes for heritable Mendelian and complex traits alike requires genetic mapping, i.e. the identification and localization of genes that underlie heritable phenotypes. Genetic mapping is accomplished by correlating DNA variation with phenotype.

The Genetic Basis of Sleep and Sleep Disorders, ed. Paul Shaw, Mehdi Tafti and Michael Thorpy. Published by Cambridge University Press.

In some instances, the DNA variant being tested will in fact be the causal variant for the phenotype. In other instances, the DNA variant tested will simply be correlated with the truly causal variant. When two genetic variants are correlated, this correlation is referred to as linkage disequilibrium. The two primary analytic techniques for genetic mapping are linkage and association. In linkage mapping, segments of the genome are tracked in families to determine whether exactly the same region of DNA is shared by members of the family that share phenotypic status. Historically, linkage has been extremely successful at the identification of Mendelian disease genes but has had limited success at the identification of risk loci for complex traits. In contrast, association aims to correlate DNA variants with phenotype in the population, as variation that increases the chances of disease should be enriched in a case sample.

In this chapter, we will discuss the methodological considerations surrounding linkage and association studies as well as results of both approaches as they relate to sleep and sleep disorders. The clear heritability of sleep-related phenotypes has spawned a number of efforts to identify regions in the genome that are suspect for contributing to disease or phenotype. Consequently, a number of linkage and association studies have been carried out to determine the genetic factors that underlie these complex phenotypes. Here, we discuss these methods and the current state of results.

Linkage

The term linkage refers to the phenomenon whereby continuous stretches of DNA are inherited together during meiosis unless separated by recombination. Recombination refers to the process of chromosomal cross-over in which parts of chromosomes break and rejoin when homologous chromosomes align during meiosis (Figure 1.1). The further apart two loci, the more likely they will be separated by recombination during the lining up of homologous chromosomes and eventually end up in different daughter cells. Thus, in linkage the aim is to roughly decipher the location of a disease-causing gene relative to a nearby sequence by tracking the concordance between genetic markers, whose genomic positions are already known, and phenotype. The earliest attempts at linkage mapping were carried out by Alfred Sturtevant in the laboratory of Thomas Morgan in the early 1900s in *Drosophila*, when he realized that he could map the linear order of genes by tracking patterns of correlations between genotype and phenotype in fly crosses, with the assumption that meiotic cross-overs would lead to association only between markers physically near to the phenotype-causing mutation [10]. Linkage in families became feasible around 1980 when Botstein and colleagues proposed the idea of using restriction fragment length polymorphisms (a type of variant that disrupts a restriction enzyme cut site and is therefore easy to assay) throughout the genome to systematically map human genes associated with disease [11]. This breakthrough in methodology led to the mapping of the Huntington's gene on chromosome 4 in 1983 [12] followed by the systemic documentation of dense genome-wide polymorphic sites and the subsequent mapping of now over 2,000 Mendelian diseases [13].

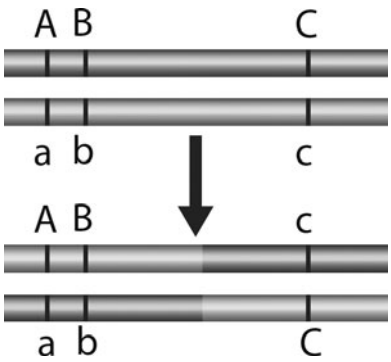

Figure 1.1 Linkage. A depiction of recombination during meiosis is shown. Three loci are depicted, each with two alleles (A/a, B/b, C/c). Due to proximity, a random recombination event will most likely separate the C/c locus from the other two. If counted over a number of meioses, one would observe that A is likely to be present with B on the same chromosome (and likewise for a and b), but that concordance with one of the C/c alleles is more random. Therefore, one would conclude that the C/c locus is far away and that the A/a and B/b loci are close to one another.

The approach to linkage mapping involves assaying genetic markers throughout the genome within families where multiple members are affected by the disease of interest. Earlier linkage studies were characterized by larger pedigrees with subsequent work extending to other study designs such as affected sibling pairs. The ideal genetic markers are ones that are easily assayed, ones that are sufficiently polymorphic across individuals to ensure a high frequency of heterozygosity and ones that are frequent throughout the genome so that a dense map can be achieved. Most early linkage studies used 300–400 microsatellites, which are polymorphisms with variable-number tandem repeats (VNTRs) that are

distributed throughout the genome at a density of 1 per 5–10 cM [14]. More recently, linkage analysis has relied on single nucleotide polymorphisms (SNPs), which are usually bi-allelic (i.e. lower rate of heterozygosity) but are much denser throughout the genome and are much more inexpensive to assay. For any category of variant, the first approach to linkage is to test each marker for linkage by comparing the odds of its being near in the genome to the disease-causing mutation versus the odds of its being independent of the disease-causing mutation (that is, it is far enough away that assortment with the disease-causing mutation becomes independent).

A variety of approaches to linkage analysis exist, the most classic of which is known as parametric or model-based linkage analysis. Here, we assume that the trait in question is determined by a single locus and that familial resemblance is due only to this single locus according to the presumed inheritance pattern, whose parameters involve assumptions on the mode of inheritance as well as the penetrance of the disease-causing allele. The statistic typically used for this test is known as a LOD (logarithm of the odds) score, which is calculated as:

$$LOD = \log_{10} \frac{L(data|\theta, f_{DD}, f_{Dd}, f_{dd})}{L(data|\theta = .5, f_{DD}, f_{Dd}, f_{dd})}$$
$$= \log_{10} \frac{(1-\theta)^{NR}\theta^{R}}{.5^{NR+R}}$$

where L denotes the likelihood, NR refers to the number of non-recombinants in the family (the number of affected individuals with the allele), R refers to the number of recombinants (the number of affected individuals without the allele plus the number of unaffecteds with it), f refers to the probability that an individual of a particular genotype is affected (which will vary depending on model assumption) and θ refers to the presumed probability of the marker and disease-causing allele ending up recombined, where $0 < \theta < .5$ (50% being the maximum probability consistent with independent assortment). The designation of the risk allele is typically achieved using the grand-parental generation of a pedigree, and counting subsequent meioses within a pedigree as R or NR requires that the meiosis be *informative*, a description that means we can determine the parental origin of an offspring's alleles. Figure 1.2 shows an example pedigree and discusses the outcome under a model of full penetrance versus a model of incomplete penetrance.

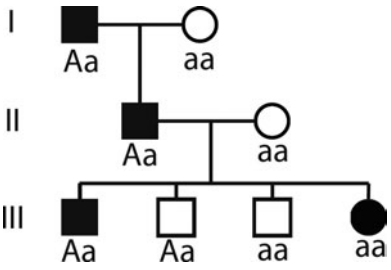

Figure 1.2 Parametric linkage analysis. (**A**) Under a dominant model with full penetrance, we use generation I to phase, which yields A as the risk allele. Under this model, III-1: NR; III-2: R; III-3: NR; III-4: R. (**B**) Under a model of incomplete penetrance, the assessment becomes: III-1: NR; III-2: NR; III-3: NR; III-4: R. NR: non-recombinant. R: recombinant.

Conventionally, a LOD score of +3 or greater denotes sufficient evidence for linkage at the tested marker at the presumed recombination frequency; therefore, LOD is calculated over a range of thetas and the maximum score achieved, if over 3, is the final estimate of the true theta, which serves as a proxy for the distance between the marker and the disease allele. If LOD never rises above 3, one assumes that there is not sufficient evidence to make a conclusion about linkage (usually due to insufficient number of informative meioses in the pedigree), and if the LOD goes below −2, conventionally we presume that there is enough evidence to deem the two markers definitively unlinked at the corresponding theta. Finally, LOD scores can be combined over multiple unrelated pedigrees to boost power to detect linkage, under the assumption that the same locus is causal in each family.

The analysis can include variations on the chosen model. For example, penetrance may be age-dependent, such as in Huntington's disease. Alternatively, there may be sex-specific penetrances in the case of a disease that affects people differentially based on sex. Each factor of the proposed genetic model can therefore contribute to the final designation of recombination status; however, the investigator must specify parameters of the model and loss in power is correlated to the degree to which the chosen model is inappropriate.

Alternative methodologies for linkage analysis can be used, such as non-parametric linkage analysis and multipoint linkage analysis. Non-parametric linkage analysis (also known as model-free) does not assume a specific genetic model for disease. One such approach, known as the affected sib pair method, tests for excess sharing of marker alleles identical by

descent (IBD) in affected sib pairs [15]. IBD here means that exactly the same segment of DNA is carried by two members of a pedigree. Multipoint linkage analysis tests aim to determine the IBD states for all pairs of individuals across a pedigree by leveraging information from multiple markers. With the identification of these IBD states, a more formal test of excess IBD sharing based on sharing disease phenotype can be conducted. These approaches are described in detail elsewhere [16].

A number of disease and study characteristics that aid in successful linkage mapping included highly penetrant causal genetic variants, relatively little environmental influence on the phenotype, large families and minimal locus heterogeneity. Linkage has therefore been very successful in mapping Mendelian diseases (although some loci, if of high enough effect, can be mapped via linkage in genetically complex diseases).

The majority of sleep-related phenotypes that have been successfully mapped via linkage involve familial sleep disorders. Two categories of disorders are described here: primary disorders of sleep, including narcolepsy–cataplexy as well as familial advanced sleep phase syndrome, and disorders with sleep disturbances, including RLS. Universally, success met by investigators using linkage usually involved large families with multiple affected members.

Linkage results in narcolepsy–cataplexy

Narcolepsy is a disorder characterized by excessive daytime sleepiness and abnormal rapid eye movement (REM) manifestations including sleep paralysis, hypnagogic hallucinations and sleep-onset REM periods [17]. The strict definition of narcolepsy is narcolepsy–cataplexy, which refers to individuals whose narcoleptic symptoms include cataplexy, a sudden and transient loss of muscle tone. Familial forms of narcolepsy that follow a clear inheritance pattern are very rare. For the most part, the risk of narcolepsy to relatives of an affected individual is low (1–2%), albeit higher than the average population risk (.02–.18%) [18]. Furthermore, the concordance rate of monozygotic twins is estimated at 25–31% [18], suggesting influence of the environment and non-Mendelian inheritance. Nonetheless, in 2004 Dauvilliers *et al.* identified a large French family with narcolepsy–cataplexy tracking in an autosomal dominant fashion. They successfully mapped a susceptibility locus to chromosome 21q (LOD = 4.00) [16], a region containing over 20 genes. However, further linkage attempts in narcolepsy have not been successful. Likely because of the heterogeneous nature of its genetic architecture, narcolepsy has seen more success with association testing which will be discussed in further detail later in this chapter.

Linkage results in familial advanced sleep phase syndrome (FASPS)

FASPS is an autosomal dominant circadian rhythm disorder whereby the sleep–wake cycle is shifted 4 h earlier [19,20]. The initial study that showed it to be inherited in an autosomal dominant fashion was a linkage study on a large family with over 20 affected individuals [19]. One linkage peak was identified on chromosome 2q (LOD = 5.25). Within the linkage region, the gene *PER2* was found to contain a frameshift mutation in the binding site for CKIδ, which is a kinase that phosphorylates PER2. Following this study, Xu *et al.* described in a family of five affected individuals a mutation in *CKIδ* leading to the same phenotype [19]. Although the latter study did not use linkage but rather candidate gene sequencing, they provided strong evidence for the importance of this mutation by showing perfect segregation with disease and showing its absence in 250 controls.

Linkage results in restless leg syndrome

RLS is a disorder characterized by parasethesias described as an irresistible urge to move one's legs [21–24]. These urges often occur at rest and cause sleep disturbance, leading to chronic sleep deprivation. RLS is fairly common, with the prevalence estimated to be between 1.2 and 15% depending on the population [25]. The mode of inheritance is debated in the literature, with some families showing autosomal recessive and autosomal dominant inheritance patterns and other families exhibiting more complex inheritance pattern with environmental influence [25]. Nonetheless, linkage studies have been successful throughout the last decade. The first locus to be documented was on 12q in a French-Canadian sample under a recessive model (maximum LOD score 3.42) [21] and was then confirmed in other families [26,27]. This finding was then followed by the identification of four additional linkage peaks at 14q13–21 in an Italian family [28], 9p24–22 [29] and 2q [30] (see Table 1.1 for details). Although some

Table 1.1 Human susceptibility loci for sleep and sleep disorders.

Trait	Estimated heritability	Loci via linkage	Loci via association (candidate gene[c] and OR[d])		Publications
Restless leg syndrome	60%	12q, 14q21, 9p24–22, 2q	*MEIS1* *BTBD9* *MAP2K5/LBXCOR1* *PTPRD* *TOX3*[b] *2p14*[b]	1.68 1.47 1.41 1.29 1.35 1.23	Desautels *et al.* (2001) [21]; Bonati *et al.* (2003) [28]; Chen *et al.* (2004) [29]; Levchenko *et al.* (2004) [26]; Desautels *et al.* (2005) [26]; Pichler *et al.* (2006) [30]; Winkelmann *et al.* (2007) [31]; Stefansson *et al.* (2007) [32]; Winkelmann *et al.* (2011) [33]
Narcolepsy		21q	*HLADQA1/DQB1*[a] *TCRα* *P2RY11*	1.79 1.54 1.28	Honda *et al.* (1983) [34]; Hallmayer *et al.* (2009) [35]
Familial advanced sleep phase syndrome (FASPS)		PER2	*CKIδ*[a]		Toh *et al.* (2001) [19]; Xu *et al.* (2005) [20]
Sleepiness (Epworth Sleepiness Scale)	0.29	–	*PDE4D*	NR	Gottleib *et al.* (2007) [36]

[a] Association identified through candidate gene analysis.
[b] Association not yet replicated.
[c] Although loci are often named by the closest gene to the lead SNP, one cannot assume that the named gene is causal until definitive proof is provided.
[d] Estimated ORs are generally concordant across studies. Here, we arbitrarily report the odds ratio from the study with the strongest association.
NR, not reported in the paper.

families can be explained by a single locus, these four loci found through linkage do not explain all familial cases of RLS. Later in the chapter, we discuss the use of association testing in RLS and success therein.

Complex phenotypes

Complex phenotypes are influenced by multiple genetic and non-genetic factors. As a result, these phenotypes cluster in families but do not follow any clear mode of inheritance. Unlike the rare, highly penetrant mutations of Mendelian disease, the contributing genetic factors in complex traits are presumed to individually impart only a small risk for disease; the more risk factors an individual has, the more their risk for disease onset. Furthermore, environmental influence plays a large role in many complex phenotypes. These factors make linkage analysis ill-suited for discovering risk alleles, because any one allele will not segregate cleanly with disease.

Complex phenotypes are divided into two classes: continuous and categorical. A continuous trait is one that does not have a discrete scale (i.e. a measurement can take any value), such as height in meters, body mass index (BMI) in kg/m^2 or duration of sleep in hours. Continuous traits are often studied in the general population to identify genes and pathways that play a role in dictating variation; however, selected samples such as extremes of the distribution are also used to boost power. A categorical trait is one that is qualitative and that falls into non-overlapping groups, such as a diagnosis of RLS or insomnia, each of which are yes/no. These groups may be ordered (e.g. low, medium, and high) or unordered (e.g. blue, green, and yellow). The most typical categorical traits studies are those of disease with affected and unaffected as categories. Although these diseases are usually studied using a case-control model, dichotomous traits are not unlike continuous traits in that they can be assumed to result from complex inheritance involving many genes, and the designation of one group over another is assumed to be based on an underlying liability distribution to which a threshold is applied (see Figure 1.3). Although not discussed in detail in this chapter, this idea is termed the Liability Threshold Model, with liability being one's

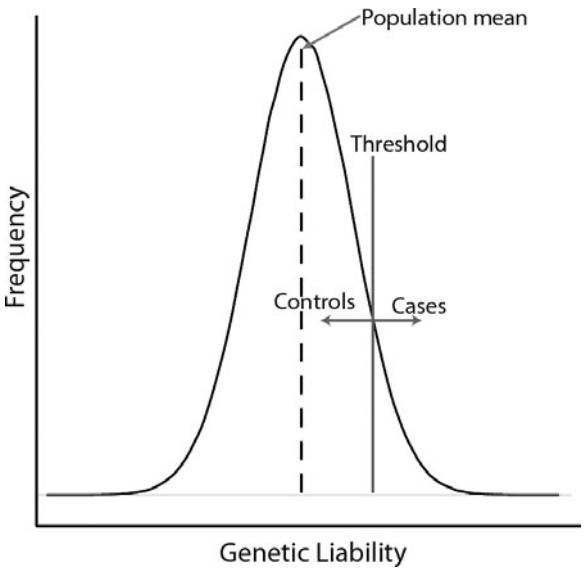

Figure 1.3 Liability Threshold Model. The liability threshold model suggests that for dichotomous traits influenced by many genetic factors each of small effect, an underlying distribution exists that depicts the distribution of liability toward a disease (predisposition) across a population. A threshold is applied to this distribution, above which individuals exhibit the trait (cases) and below which individuals do not (controls). The location of the threshold is determined by the disease prevalence – that is, the area under the curve to the right of the threshold should be equal to the population prevalence of the disease.

predisposition or vulnerability to a disease or phenotype [37]. In practice, this distribution is not observed (i.e. latent) but assumed.

The study of complex traits involves analyzing variation of phenotype within a population of individuals, assumed to be a product of genetics and environment. While linkage analysis focused on specific crosses, complex trait analysis considers variation of a trait within a population and the degree to which genetic variation contributes to the phenotypic variation. Inherently, therefore, the study of complex traits involves the study of populations, rather than families. For example, when studying duration of sleep, one would first observe the natural variation in sleep duration in a population and then try to estimate the degree to which that variation is due to genetics.

The degree to which genetic factors contribute to phenotypic variance is termed heritability. Heritability is the proportion of phenotypic variance that is due to inherited factors influencing the trait. Typically these calculations are made by comparing close relatives. The most frequent approach taken to estimate heritability is twin comparison: fraternal twins (dizygotic, or DZ) share on average half the amount of DNA in contrast to identical twins (monozygotic, or MZ) that share all of their DNA. Heritability can thus be approximated by twice the difference between the phenotypic correlation of MZ twins and DZ twins. The other main approach historically taken is the comparison of parent–offspring pairs. In this case, heritability is the square of the correlation coefficient between mid-parent and offspring phenotypic scores. Understanding the heritability of a trait is a critical first step in setting expectations for results of genetic endeavors. Methods for the estimation of heritability have been developed for multivariate traits as well as more complex family structures [38], but such methods are beyond the scope of this chapter.

Methodology in studying complex phenotypes

The large-scale success of linkage analysis for monogenic diseases naturally encouraged investigators to apply the same methodology to complex traits. However, it quickly became clear that this approach was underperforming in more common complex phenotypes despite strong heritability. As discussed, these traits are unlike Mendelian phenotypes in that they are highly polygenic, influenced by environment, and contributed to by genetic variants of individually very low effect. These factors make linkage analysis much less powerful in identifying the molecular genetic basis of these traits.

Association analysis, on the other hand, is an approach that tests for differences in allele frequencies that correlate with phenotype. The core test is to compare the allele frequencies between cases and controls or test for mean differences in a continuous trait conditional on genotype. Compared to linkage analysis, this approach is better powered to detect such associations of weak effect as it is a test of means, rather than variances, and that large cohorts of unrelated individuals can be tested jointly, rather than focusing simply on large affected families.

The first attempts at association testing were based on candidate gene studies where investigators compared variants between cases and controls within a single gene of interest, such as the mapping of the human leukocyte antigen locus to autoimmune disease [39] and the association between variants at APOE and Alzheimers disease [40]. In sleep, many candidate circadian genes thought to be involved in

controlling sleep through studies of *Drosophila* have been studied in humans [9]. These include *CLOCK*, *PER1*, *PER2*, *PER3*, *TIMELESS*, *CKIδ*, and *CKIε* [19,20,41–44]. Candidate gene association studies have been met with variable success. The criticism of this approach is that results were often not replicable, however, as often times nominal significance was counted as significant ($p < 0.05$) without correcting for the number of traits tested and because population substructure cannot be easily accounted for, which we now know to be a major confounder of association studies ("population stratification," discussed in this chapter) [45]. In the early 2000s, investigators sought a more unbiased survey of the entire genome. This approach is known as a genome-wide association study (GWAS), now the gold standard methodology for identifying genetic associations to complex traits.

Much of the focus for GWAS in complex traits over the past decade has been on common variation. Common SNPs are defined as those whose minor allele frequency is > 1%. In the European population, there are 10 million sites in the genome at which individuals' genotypes vary [46]. These sites comprise about 90% of an individual's heterozygous sites throughout their genome [45]. Although genetic variation across the allele frequency spectrum likely contributes to complex traits and disease, theoretical arguments grounded in population genetics predict the genetic architecture of common disease to be at least in part due to common variation (hence the so-called "common-variant common-disease" hypothesis). This argument includes the typical late onset of many common diseases that precludes causal alleles from strong natural selection, causal alleles being neutral in the past and only now having an effect due to recently introduced changes in living situations, recent population expansion allowing detrimental alleles to rise in frequency and phenomena such as heterozygote-advantage [45]. Furthermore, from a practical standpoint, this type of variation is extremely convenient because there is widespread correlation among common variants due to their being relatively old in evolutionary history and recombination happening mostly at hotspots. This means that only a subset of variants needs to be genotyped in a given study to serve as a proxy for nearby DNA, and microarray technology can easily allow for cheap, direct genotyping of these hundreds of thousands (and now a million) SNPs [47]. This chapter will therefore focus on common variation.

The goal of GWAS is to test variants throughout the entire genome for a difference in the number of people carrying the minor (or major) allele between cases and controls or as a function of the trait. A simplified approach to GWAS is described here and involves five steps: sample collection, genotyping, association testing, population stratification, and replication.

Sample collection. The first step in a GWAS is collecting samples, with emphasis on power and appropriate matching of cases and controls. As power to detect association is in part a function of the number of samples, one can roughly predict the approximate number of samples needed to detect associations at different effect sizes (i.e. odds ratios). For example, in a theorized case-control study, to achieve 80% power to detect association at alpha of 5e-8 to an allele of MAF 7% in a disease with 1% prevalence with a relative risk of 1.5 for heterozygotes, we would need 3500 cases and 3500 controls. Typically, the associations we are well powered to catch first are those at relatively common SNPs and of high effect size. Power goes down as effect size and minor allele frequency go down, necessitating larger and larger sample sizes; for this reason, a number of researchers have forged international collaborations to carry out meta-analyses, where cohorts are combined to yield large sample sizes on the order of 10s to 100s of thousands of individuals. A nice tool for power calculations can be found here: http://pngu.mgh.harvard.edu/~purcell/gpc/.

In addition to power considerations, cases and controls need to be well matched on any variable that could confound the analysis. First and foremost, cases and controls should be of the same ethnic background so as to minimize the effects of population stratification (see discussion below). Beyond this, investigators can try and match on any other variable that may confound the analysis – for example, limiting enrollment to individuals of a certain age range. Finally, one should take great care to randomize samples with respect to the timing of their being assayed; separating cases and controls on the time frame in which they were genotyped as well as any platform differences can lead to very large batch effects.

Genotyping. With such large sample sizes, genotyping needs to be technologically easy and

cost-effective. There are many companies that offer cheap, high-throughput genotyping arrays [47]. These technologies have grown from earliest implementations of 100,000 markers to assays with 2.5 to 5 million markers. Technologically, these arrays typically require DNA amplification followed by hybridization to the array with a set of probes that correspond to loci throughout the genome. Allelic discrimination is usually accomplished either through allele-specific primers or through allele-specific probes. Measuring the strength of a platform includes the accuracy (how well it agrees with the known genotype), call-rate (how often it can confidently call a genotype), reproducibility (how concordant the results are across replicates), how well it covers the genome as well as how easily it is multiplexed (i.e. ability to assay more than one sample at a time). Although this chapter mainly focuses on SNP assays, for the past 5 years investigators have been looking beyond SNPs and toward submicroscopic structural variation in the genome known as copy number variation (CNV). These types of variants are typically assayed via array comparative genomic hybridization (aCGH) as well as creative uses of the standard SNP chips to estimate CNV status. For simplicity we will focus mainly on SNP analysis in this chapter, but similar principles of association testing apply when looking at CNVs.

Quality control. Quality control (QC) involves filtering out "bad" data – bad SNPs and bad individuals – that could lead to Type I or Type II errors and is described in detail elsewhere [48]. SNPs are mainly filtered on call-rate/missingness, minor allele frequency, and Hardy–Weinberg equilibrium. Individuals are filtered on gender checks (i.e. stated gender does not agree with genotype), cryptic relatedness and replicates, population outliers (using principal component analysis), and high or low heterozygosity. Ultimately, QC reduces the chance that an association is discovered due to an exogenous effect unrelated to the phenotype being studied, and it cleans up the data to maximize power to discover true associations [49].

Association testing. Once genotypes are collected across samples and QC is completed, each SNP is tested for association to disease. This can be accomplished using a simple chi-squared test or logistic regression if handling a case-control sample or via linear regression if handling a cohort measured on a quantitative trait. Care should be taken to control for any confounding variables in the analysis by adding them as covariates. For example, if studying sleepiness as a quantitative score, age, sex and BMI are typically used as covariates in the analysis as differences in the trait attributable to the covariates can lead to association entirely explained by a the covariate [36]. When looking genome-wide, around 1 million tests are performed in any analysis, and therefore correction for multiple testing is critical. Using the accepted association threshold of $\alpha = 0.05$, the Bonferonni corrected p-value becomes 5×10^{-8}, which represents the gold-standard threshold for genome-wide significance.

Population stratification. Population stratification refers to the presence of any systemic differences in allele frequencies between cases and controls or across individuals according to quantitative trait value that are related to ancestry and not to the phenotype being studied. Two approaches to control for this are genomic control and principal components analysis. λ is assumed to be a constant inflation factor across all loci and is calculated as follows:

$$\lambda = \frac{median(\chi_1^2, \chi_2^2 \cdots \chi_n^2)}{.455}$$

To correct using genomic control, one can divide all association χ^2 values by λ. Genomic control has also proven to be a useful metric for the identification of potential bias in the distribution of test statistics. The other widely used approach is principle components analysis (PCA), implemented in EIGENSTRAT [50]. Here, PCA is applied to genotype data to infer continuous axes of genetic variation and the first axis typically describes population substructure. The principal components attributable to population stratification can then be used as covariates in the association test to remove association due to ancestry. More recently, methods to handle other sources of structure in the data beyond population substructure (i.e. family structure or cryptic relatedness) have been developed [51]. These methods involve the use of mixed models, and software tools are available for fast implementation [52].

Replication. GWAS hits that achieve $p < 5 \times 10^{-8}$ should only be considered true associations when

they are replicated in an independent set of individuals. Although much attention is given to controlling for technical artifacts, covariates and ancestry, unforeseen forces can lead to Type I errors. Therefore, an association that replicates in a set of independent individuals and ideally on a different genotyping platform is considered bona fide. Often, investigators will look to replicate in individuals of different ethnic backgrounds. When replicating, one need only take the top results and genotype them, Bonferonni correcting for the number of variants tested. The final *p*-values reported are typically the combined exploration and replication statistics.

Genome-wide association studies in sleep and sleep-related disorders

GWAS has recently been employed in studying sleep phenotypes. Two general categories have been studied: diseases that manifest with sleep disturbance (discrete traits) and quantitative characteristics of sleep (continuous traits). Discrete phenotypes with sleep sequelae include RLS, narcolepsy and insomnia/hypersomnia, as well as a number of disorders that are known to include disturbances in sleep such as bipolar disorder and attention deficit hyperactivity disorder (ADHD). These are studied using a case-control setup. Quantitative characteristics of sleep include sleep quality, sleep pattern, sleep timing and EEG profiles, which are studied using linear regression.

Sleep diseases as dichotomous traits

Early association results in narcolepsy–cataplexy were discovered through candidate gene association tests. In 1983 Honda *et al.* reported a strong association between HLA-DR2 and narcolepsy in Japanese individuals [34] that was then replicated by many groups in Caucasian individuals and further refined to the DQB1*0602 and HLA DQA1*0102 alleles [18,53]. This is the strongest association found to date for narcolepsy, with over 85% of individuals with narcolepsy–cataplexy and only 12–38% of the general population carrying the minor allele of these SNPs. The large percentage of unaffected individuals carrying the variant suggests that other genetic contributing factors are likely to be involved in narcolepsy–cataplexy.

GWAS in narcolepsy–cataplexy and RLS have been somewhat successful. In 2009, Hallmayer *et al.* reported an additional association to narcolepsy–cataplexy at a locus containing the T-cell receptor alpha gene (TCRα) through a large GWAS with replication genotyping [35]. Subsequently, a large GWAS found and replicated an association at *P2RY11* and showed that the risk allele was associated with decreased expression of *P2RY11* in CD8+ T cells and NK cells [54]. In 2008, Miyagawa *et al.* found a tentative association to an SNP near *CPT1B* in Japanese individuals ($p = 6 \times 10^{-8}$), but more genotyping as well as replication will be required to determine whether this locus is truly associated [55]. Interestingly, narcolepsy is known to involve the loss of ~70,000 hypothalamic neurons producing hypocretin [56]. These genetic findings implicate an autoimmune process that is responsible for the destruction of neurons.

RLS has also benefited from GWAS efforts. Three loci were discovered and replicated in a GWAS by Winkelmann *et al.* in 2007 that were then replicated by others [31–33]. These loci include *MEIS1*, *BTBD9* and *MAP2K5/LBXCOR1*. A fourth locus was discovered in 2008 on chromosome 9p23–24 containing the gene *PTPRD* that was then replicated in an independent cohort [33]. These genes demonstrate a potential role of developmental regulatory factors in RLS that affect spinal cord regulation of sensory perception and locomotor pattern generation, because many of these genes are known to play a role in the developing spinal cord [57]. In general, however, for narcolepsy and RLS these findings only explain a small percentage of the genetic variation.

Other heritable disorders that manifest with sleep disturbances include bipolar disorder and ADHD. The genome-wide association studies of bipolar have been successful, with the most recent meta-analysis identifying a few loci from ODZ4, CACNA1C, and potentially the ITIH1 region [58]. To date, ADHD meta-analysis has yet to identify any significant loci [59].

Diseases such as insomnia and hypersomnia, dissociated REM sleep (such as sleep paralysis and hypnagogic hallucinations) as well as obstructive sleep apnea have been found to cluster in families but have yet to yield any genetic associations [57,60,61]. Phenotypic variability is one possible explanation. For example, many genetically driven risk factors can lead to separate forms of sleep apnea that break the disorder into different categories, such as upper-airway anatomic features, variable lung capacity, and obesity [57]. These factors are likely phenotypes

in and of themselves with distinct risk factors that ultimately lead to sleep apnea but identifying an association with a heterogeneous patient population will require many more people than have been studied to date.

Sleep as a quantitative trait

A large part of the field has focused on sleep as a quantitative trait and found many components to be heritable. Timing (bedtime, waking time, sleep duration), quality (number of times they woke up throughout the night), sleepiness/wakefulness on waking or during the day, and EEG profiles have all been found to be heritable through twin studies, ranging from 15% to 45% [3,5,6,8,9,36,62]. Consequently, a handful of cohorts have been gathered for GWAS to look for alleles throughout the genome that correlate with one of these traits.

To date, only one such study has been successful. Gottlieb *et al.* identified a locus achieving genome-wide significance for association to sleepiness, a quantitative trait measured using an eight-question questionnaire known as the Epworth Sleepiness Scale [36]. This association is on chromosome 5 near the gene *PDE4D*, although it has yet to be replicated. Other results include a tentative association between EEG profiles and *PER3* via candidate gene association testing [43], but this was not verified by other groups [57]. The current lack of results to quantitative sleep traits likely reflects the difficulty of articulating these phenotypes. They are difficult to define and to measure – even EEG, which is perhaps the most objective, yields a very noisy trace that requires significant transformation to be a stable trait to study. With larger sample sizes and better articulated measurements, the striking heritability of these traits may one day yield exciting results as to the genes that may help define their variance in humans.

Conclusion

So far, the genetics of sleep phenotypes remain largely undiscovered. The first promising steps have been taken to understand the biological basis of these traits. Linkage has been forged in many large families to yield a handful of loci for a subset of diseases in sleep. More recently, investigators have begun to organize large cohorts of individuals with which to conduct genome-wide association studies. Although still in the early stages, these studies have shed light on surprising pathways that may be relevant in the control of sleep, including immune-mediated processes and developmental regulatory pathways. In general, however, much more remains to be discovered and most of the observed heritability of sleep disorders and traits is yet unexplained. Based on early success, future efforts in GWAS with larger sample sizes are likely to be fruitful.

Owing to recent technological advances, genome sequencing in medical genetics to discover disease-relevant variants is now a reality [46]. Although their contribution to disease architecture remains unclear, investigators are now beginning to study rare variation (minor allele frequency $< 5\%$) and its role in medical genetics. Despite significant analytic challenges to overcome in studying such variation, fully sequencing exomes and genomes is gradually becoming technologically and economically feasible. The future of genetics of sleep and sleep-related disorders will likely include more rare variation discovered through sequencing.

The tools of linkage, genome-wide association, and eventually sequencing promise to yield new insights into the basis of sleep traits. These are early days for the field of sleep genetics, but leveraging international collaboration to expand sample sizes and rapidly advancing technology to explore genome sequencing will shed light on yet undiscovered genetic factors underlying the heritability of sleep.

References

1. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nat Genet*. 2003;**33**(Suppl):228–37.
2. Geyer H. Uber den Schlaf von Zwillingen. *MGG*. 1937;**73**:524–27.
3. Klei L, Reitz P, Miller M, *et al.* Heritability of morningness–eveningness and self-report sleep measures in a family-based sample of 521 Hutterites. *Chronobiol Int*. 2005;**22**(6):1041–54.
4. Ambrosius U, Lietzenmaier S, Wehrle R, *et al.* Heritability of sleep electroencephalogram. *Biol Psychiatry*. 2008;**64**(4):344–48.
5. De Gennaro L, Marzano C, Fratello F, *et al.* The electroencephalographic fingerprint of sleep is genetically determined: a twin study. *Ann Neurol*. 2008;**64**(4):455–60.
6. Heath AC, Kendler KS, Eaves LJ, *et al.* Evidence for genetic influences on sleep disturbance