[a] *Rough draft version of chapter. Figures to be redrawn.*

4.5 Complex traits: II. The GWAS paradigm

Do you consider yourself a morning person? Your preference of when you go to sleep, and when you wake up, is called ***chronotype*** *and has a heritability of around 25%* [757]. [a]

How could we find the genes that help determine chronotype? How many genes contribute? What tissues do they act in? What are the key molecular pathways?

Being a late-sleeper is correlated with diabetes and schizophrenia. Is late sleeping a ***causal factor*** *for diabetes or schizophrenia? Or perhaps the inverse: do those traits cause people to sleep late? How would we know?*

In the upcoming chapters we'll learn how to tackle all these questions, using a study design known as a ***genome-wide association study (GWAS)****.*

We'll see that chronotype, and most of the other ways that we differ from one another — e.g., by height or weight; cholesterol levels or red blood cell counts; personality or intelligence; or by the diseases that afflict us: diabetes, arthritis, addiction – are so-called ***complex traits****, influenced both by thousands of genetic variants across the genome, and by environmental and random factors* [b] [758].

Figure 4.55: **Endymion of Greek mythology, sleeping.** *According to myth, Zeus gave eternal youth to the moon goddess Selene's lover by putting him into everlasting sleep. (Clearly, Endymion's long-sleeper phenotype was due to environmental factors, not genetics.)* 2nd Century Roman statue of Endymion. Image: Ad Meskens, Wikimedia. [Link]

[b] *Prior to reading this chapter, you may wish to revisit Chapter 4.1 for a light introduction to trait genetics.*

As we saw in Chapter 4.2, there was great progress in the 1980s and 90s at mapping the genes for Mendelian diseases including cystic fibrosis and Huntington's disease using linkage mapping in family pedigrees. But these methods didn't work well for complex traits or diseases, which often "run in families" to some extent, but without clear inheritance patterns.

In parallel, by the 1990s people started to test a different study design, known as **association mapping** [759]. This design is conceptually much simpler than linkage mapping. Suppose we're interested in which SNPs affect a person's height, or whether they are a morning person. We could sequence the genomes of a bunch of **unrelated individuals**, and simply look for SNPs where genotype is correlated with phenotype:

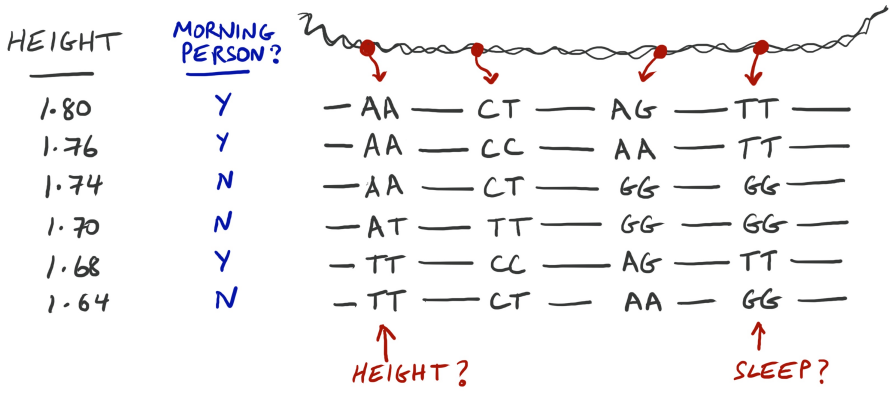


Figure 4.56: **Association mapping. A.** *Genotypes for four SNPs across a small region of the genome in a sample of six unrelated individuals (each row shows the diploid genotypes for one person). The data on the left show their phenotypes for a quantitative trait (height) and a binary trait (is this a morning person?). In this cartoon analysis,* `A` *at the first SNP might be associated with taller height, and* `T` *with being a morning person. Of course in practice we would need much larger sample sizes to be sure.*

For data analysis, it's convenient to turn the three possible genotypes at each SNP into numbers. For each SNP we'll arbitrarily label one allele as 0 and the other allele as 1 [760]. Then, for each SNP, each person's genotype

is the sum of their two alleles: so it can be either 0, 1, or 2. For example:

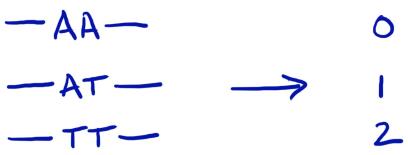


Figure 4.57: **Encoding genotypes as numbers.** *It's convenient to turn the possible genotypes at each SNP into numbers, where the genotype value is the number of alternate alleles carried by each individual:* $\{0, 1, 2\}$ *at each SNP.*

Now, to do a more rigorous test of whether a particular SNP affects height we can set this up as a regression test for each SNP [761] :

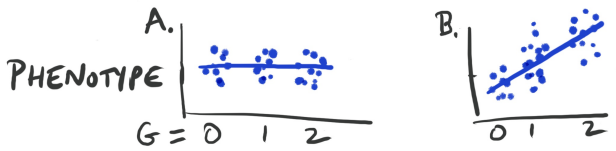


Figure 4.58: **Association mapping for a quantitative trait.** *Each panel shows the relationship between genotype at a particular SNP and phenotype. Each dot is a single individual. Panel* **A** *shows no relationship between genotype and phenotype, but* **B** *suggests that the derived allele increases the trait – we consider this SNP a "hit".*
We test this formally with ***linear regression****.*

Alternatively, we are often interested in **binary traits and diseases**: for example, which SNPs impact risk for diabetes, schizophrenia, or sleep disorders? In this case, we would often collect a sample of individuals who have been diagnosed as having the condition – we refer to these as **cases** – and a sample of **controls** who do not.

Suppose we collect 1000 individuals with a sleep disorder (cases), and 2000 controls. At a SNP that has no impact on the disorder, we expect 1/3 of individuals to be cases, for all three genotypes. But at another SNP where the derived allele increases risk, the fraction of cases increases with genotype:

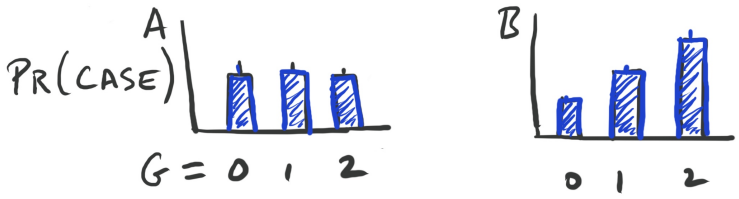


Figure 4.59: **Association mapping for a case-control trait.** *Each panel shows the fraction of individuals in the sample who are cases, within each genotype group.*
A *shows no relationship between SNP and disease risk, but in* **B** *we see that the risk of being a case increases with the number of derived alleles.*
We test this formally with ***logistic regression****.*

This, in a nutshell, is the basic concept of GWAS. After a brief historical interlude we'll dive into some deeper technical details.

A short history of GWAS. These concepts were well-understood by the 1990s, but unfortunately, the early association mapping studies were a complete bust. They typically only looked at a few SNPs in one or two genes and, in retrospect, the sample sizes were far too small. You can see just how dire the situation was in the plot below from a 2002 review article. By that time more than 1600 Mendelian diseases had been linked to genes (shown in pink), compared to less than 10 gene hits for human complex traits (blue):

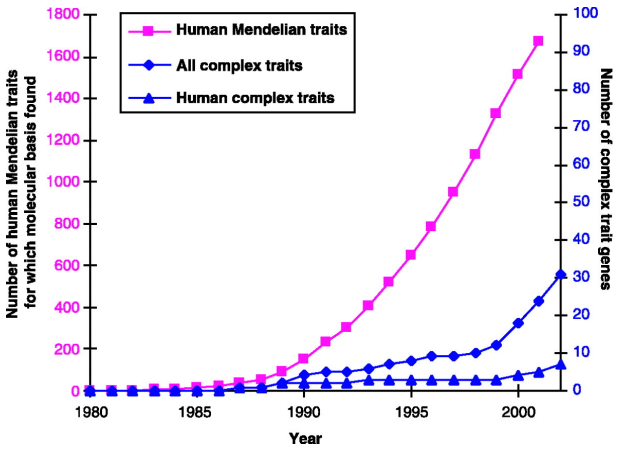


Figure 4.60: **Gene mapping discoveries by year, up to 2002.** *Mendelian traits shown in pink using left-axis labels; complex traits in blue shown with right-axis labels.* *Credit: Figure 1 from Anne Glazier et al (2002) [Link]*

But by then the way forward had already been spelled out – at least in theory. In 1996, a bold paper titled *The future of Genetic Studies of Complex Human Diseases* argued that association studies would be essential for studying complex traits [762]. Neil Risch and Kathleen Merikangas showed that association mapping is far more powerful than linkage mapping for finding alleles with small effects, but argued that it must be carried out at full genome-scale, testing SNPs across all genes at once. The arguments in this paper changed the course of human genetics [c].

[c] *Risch and Merikangas wrote in 1996: "Has the genetic study of complex disorders reached its limits? The persistent lack of replicability ... between various loci and complex diseases might imply that it has. We argue (that)... the future of genetics of complex diseases is likely to require large-scale testing by association analysis."*

The problem was that the Risch and Merikangas proposal was practically science fiction: none of the necessary tools existed to perform genome-wide association studies. In 1996 there was no human genome sequence. Only a handful of SNPs had been found. Both sequencing and SNP genotyping were labor-intensive and expensive.

But this proposal galvanized the human genetics world, and the first GWAS studies would become a reality within just one decade. In this time span, the Human Genome Project was completed (in 2001). In its wake, a major public effort called The International HapMap Project (2005) spearheaded efforts to catalog SNPs, determine their allele frequencies, and to create a genome-wide map of haplotype structure [763]. In parallel, several companies developed new genotyping platforms that were affordable at genome-scale [764].

The first large-scale GWAS was published in 2007 by a large British team known as the WTCCC [765], analyzing 500,000 SNPs in 19,000 research subjects, spread across 7 different diseases. The team identified 24 significant genetic associations in this one study, roughly matching all previous findings across the entire field [d] [766].

[d] *WTCCC leader Peter Donnelly commented in 2007 that "If you think of the genome as a very long road that you are trying to find your way along in the dark, previously we have only been able to turn lights on in a small number of places, but now we can turn on lights in a large number of places–in this case half a million lights"*

This work paved the way for an increasing flurry of GWAS studies, with ever-increasing sample sizes. Over the next decade these led to thousands of new discoveries of variant-trait associations for hundreds of different phenotypes. The plot below charts the fantastic success of GWAS as a discovery engine for complex traits genetics, showing the growth in the number of significant GWAS hits in a major public database, the GWAS Catalog. Much of this was powered by massive increases in sample size:

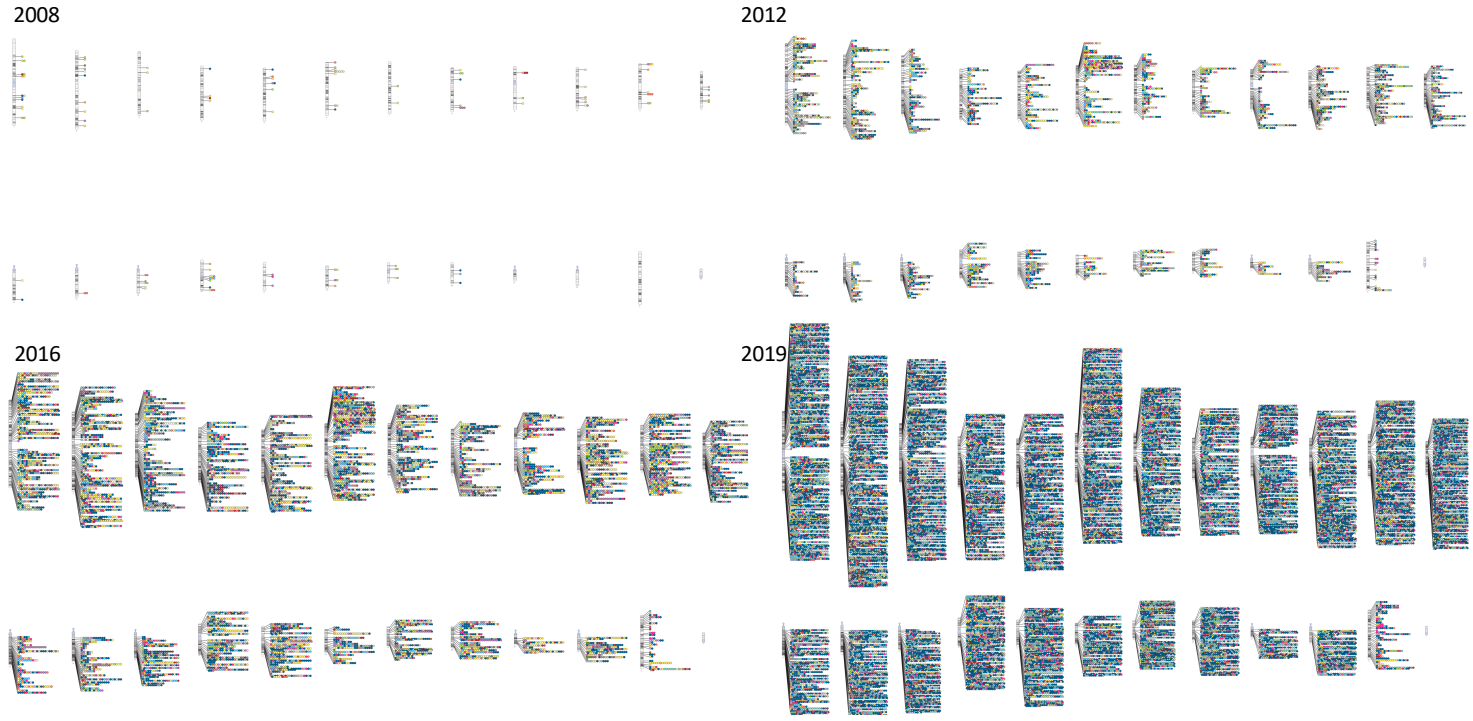


Figure 4.61: Significant GWAS discoveries by year. *Each plot shows a map of distinct genome-wide significant trait-associations across the 24 chromosomes (1...22, X,Y), discovered up to that year.* *Credit: Original figure by Darryl Leja and Teri Manolio; Updated by NHGRI-EBI GWAS Catalog [Link]*

Modern GWAS studies typically involve huge sample sizes – regularly including 10,000s to 100,000s of phenotyped individuals, and increasingly into the millions [767]. Typically, individuals are either genotyped, usually at around 1 million SNPs, or whole genome sequenced. (Sequencing is generally preferred but, prior to the early 2020s, whole genome sequencing was prohibitively expensive for large cohorts.)

Broadly speaking, there are two main kinds of cohorts: **disease-focused cohorts** put a lot of effort into recruiting *cases* (individuals with a specific disease of interest) and then compare them to matched control groups. For example, a landmark 2022 study of schizophrenia included 76,000 individuals with schizophrenia and 243,000 controls [768]. Since schizophrenia is quite rare, this study pooled data from 90 different cohorts to achieve such a large sample size.

Secondly, **biobank cohorts** generally recruit large numbers of individuals without a specific disease focus; instead they generally aim to measure many traits and/or medical conditions for each individual. Biobank cohorts offer huge economies of scale as they sometimes measure thousands of phenotypes in one study; however they are generally underpowered for rare diseases such as schizophrenia.

The most influential of these has been the **UK Biobank** (UKB) which recruited about 500,000 British individuals between the ages of 40–69, starting in 2006. Those individuals have been intensively phenotyped with thousands of data points per person, including physical measure-

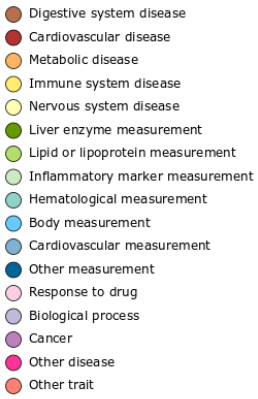


Figure 4.62: **Overview of major categories of GWAS phenotypes,** *as classified by the GWAS catalog.* *Credit: NHGRI-EBI GWAS Catalog [Link]*

ments, questionnaires, blood samples, and imaging. The UKB can access the electronic health records of the research subjects, which allows researchers to track their health prospectively over time. The entire cohort has been whole genome sequenced [e].

The success of the UKB has encouraged development of many other biobanks, including in the US (AllOfUs, Million Veteran Program, Vanderbilt and others), Finland, Estonia, Qatar, Japan, Korea, and China [769]. These vary in terms of their approach to recruitment, what phenotypes they have measured, and how accessible they are to outside researchers.

[e] *A remarkable innovation of the UKB was to make it possible for any qualified researcher worldwide to access this unique resource. In a triumph of open science, this has made the British population the world's most-studied genetic cohort.*

What can we learn from GWAS? GWAS is now a powerful discovery engine for connecting SNPs and genes to traits. What can we do with this? The outputs of GWAS have many current and future applications, including as follows:

- GWAS can establish causal links from genes to diseases, which is very hard to do in any other way since humans are not an experimental organism. This makes GWAS a powerful tool for understanding the **basic biology of diseases**, allowing us to elucidate critical cell types and molecular pathways, and to clarify causal cascades from risk factors to disease;
- Discoveries from GWAS can help us to identify genes that act as major controllers for cell functions in health, disease and aging. As such they are a powerful tool for identifying (and excluding) candidate **drug targets**;
- Polygenic prediction methods allow us to **predict who is at risk** for particular diseases, potentially influencing clinical screening, prevention, and treatment;
- By understanding key molecular pathways to disease we may be able to **stratify patients** who have a specific condition via different molecular pathways, potentially enabling personalized treatment regimens;
- Understanding the genetic architecture of complex traits is an essential step toward **evolutionary models** of how species change over time, and in particular toward understanding human evolution and adaptation.

We'll touch on all these topics as we go along.

GWAS: The nuts and bolts. We're now ready to go over the key technical details of GWAS. Remember, the basic idea is that we collect genotypes and phenotypes for a large sample of individuals (a few thousands in early GWAS up to millions in the largest recent studies). We'll use regression to test whether genotype is correlated with phenotype at each SNP across the genome, typically testing up to $\sim$10 million SNPs.

Let's start with a quantitative trait Y, like height or BMI. For each SNP l, we can fit a regression model like this [f]:

$$\underbrace{Y}_{\text{phenotype}} = \underbrace{\mu}_{\text{intercept}} + \underbrace{G_l \cdot \alpha_l}_{\text{SNP effect}} + \underbrace{C \cdot \delta}_{\text{covariates}} + \underbrace{\epsilon}_{\text{random error}} \tag{4.68}$$

Here Y is the phenotype for individual i; g_l is the genotypes ($= 0, 1, 2$) for individual i at SNP l, and α is the estimated effect size of SNP l. **For each SNP we want to test the null hypothesis that $\alpha_l = 0$.**

We'll also typically include a vector of **covariates**, C, in the regression – these are individual attributes like age, sex, and sampling time/location, as well as genetic ancestry. The covariate values are stored in a matrix C. The regression model fits a vector of effect sizes δ for the covariates (pronounced "delta").

Covariates are factors that are likely to correlate with variation in the phenotype Y. You can think of these as falling into two categories: some of these are likely to be independent of genotype, for example things like age, sex, or measurement batch [770]. By including these as covariates in the analysis we can usually increase power. So for example if we want to model height, we know that average height is different between males and females, so we might code sex as 0/1 to indicate female/male status. The regression model will then fit the corresponding effect size, δ, to capture the average height difference between males and females. We're usually not particularly interested in learning the covariate effect sizes, but *by including these covariates we reduce extraneous variation in Y, and this improves our power to detect the SNP effects we do care about.*

The second category are *covariates that control for population structure*, which can otherwise cause false positives. We'll return to this below.

[f] *We'll use the following **notation for regression coefficients**, though not all of these will appear in every model:*
α: SNP effects on traits
β: SNP effects on genes
γ: gene effects on traits
δ: covariate effects
ϵ: error terms.
In relevant settings this parameterization implies $\alpha = \beta\gamma$ (more on this below!).

Optional Details: Binary traits. The expression above is given for a quantitative trait such as height, but as we saw above, GWAS is often applied to so-called **case-control** or **binary** traits – usually measuring presence/absence of disease. To model this, we'll set the phenotype code for individuals with disease to be $Y = 1$ and $Y = 0$ for individuals without.

This can be analyzed using **logistic regression**, where we replace the left-hand side of the equation with the *logit* function, defined below. In this model, we use the genotype and covariates of each individual to predict the **probability** that they carry the disease [771]:

$$\underbrace{\log \frac{\Pr(Y=1)}{\Pr(Y=0)}}_{\text{logit function}} = \underbrace{\mu}_{\text{intercept}} + \underbrace{g_l \cdot \alpha_l}_{\text{SNP effect}} + \underbrace{C \cdot \delta}_{\text{covariates}} \tag{4.69}$$

Now β_l is measured in units of log odds of disease, $\Pr(Y = 1] / \Pr(Y = 0)$. You can interpret a non-zero effect size α_l as saying that each additional non-reference at allele at SNP l increases the *log odds* that someone has the disease by an amount α_l.

We can fit these models using standard approaches for ordinary least squares regression (for a quantitative trait), or logistic regression (for a binary trait) [772].

For both versions of the analysis (quantitative and binary traits) *we fit the regression using one SNP at a time, and test whether the SNP effect size α_1 is different from zero. We'll summarize this evidence using a p-value – one p-value for every tested SNP across the genome.*

Manhattan Plots. We visualize the GWAS output using a plot known as a Manhattan Plot [g] [773]. Here, the x-axis shows the chromosomes ordered by number, and the y-axis shows the *-log10(p-value)*. Each point is a single SNP. The odd and even numbered chromosomes are plotted in different shades for visual effect.

[g] *The Manhattan Plot is so-named because the patterns of hits evoke skyscrapers towering above a city-scape.*

Most importantly, in a successful GWAS you can see clear towers of signals: **these towers indicate regions of the genome that contain causal variants that impact the phenotype**.

I promised you I'd tell you about the genetics of **chronotype**. Well, here it is. The Manhattan plot below is from a GWAS of 697,000 individuals from UK Biobank and a cohort called 23AndMe; this study identified 351 distinct regions of the genome that impact whether a person is a morning-person [774]:

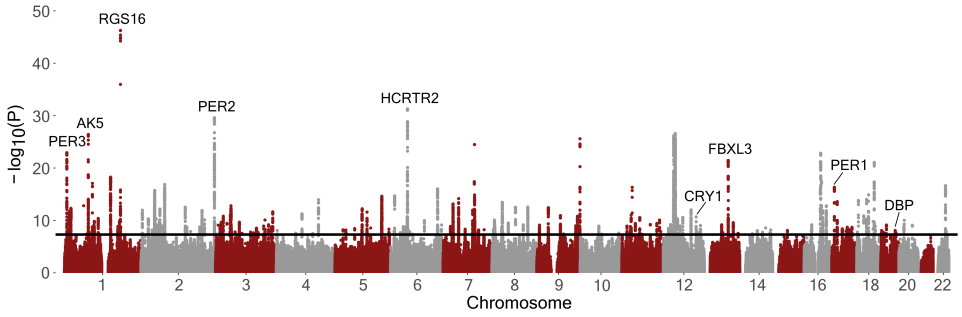


Figure 4.63: **Manhattan Plot for Chronotype.** *Each point represents a single SNP; SNPs are ordered along the x-axis according to chromosome and physical position within each chromosome; the y-axis gives the -log10(p-value) for each SNP, so that **higher values are more significant**. The black line indicates genome-wide significance. Interesting genes that may drive specific hits are printed on the plot.* *Credit: This figure was kindly drawn by Huisheng (Julie) Zhu using data from Samuel Jones et al (2019) [Link] CC BY 4*

You can interpret the y-axis, *-log10(p-value)*, as follows. Suppose that the *p-value* for a SNP is $p = .01$. Then *log10(.01)* is -2), and the negative sign converts this to 2. For a SNP with $p = 1 \times 10^{-10}$, we get $-log10(10^{-10}) = 10$.

The vast majority of SNPs are clumped down in the lower part of the plot (around 99% of the SNPs are below 2). Dots (SNPs) that are higher up on the plot are more significant, and a few SNPs reach up to form skyscrapers of significant signals above the red dotted line that indicates genome-wide significance [h].

[h] *Higher towers indicate more-significant signals; the height of each SNP is proportional to how much trait variation it explains.*

Before going on, we should pause for a moment to admire both the huge scale of modern GWAS (in this case ~7 large football stadiums-worth of research participants) and the impressive success of the findings – many more significant hits in this one study than were found for all complex traits put together through about 2008 [i].

[i] *This Manhattan plot hints at another key point: typical of complex traits, the genetic architecture of chronotype is due to small contributions from hundreds, if not thousands, of common variants spread across the genome. We'll explore this more in the next chapter.*

We'll explain additional key features of this plot next.

The skyscrapers in the Manhattan Plot. The most interesting parts of any Manhattan Plot are the towers of SNPs. What do these represent?

First, each of these represents a genomic region with at least one causal variant that affects the trait. (We'll talk about *how* variants affect traits in a minute.)

Next, *any SNP that is in linkage disequilibrium (LD) with the causal SNP will also have a signal in the Manhattan Plot* [j]. These non-causal SNPs in the tower are known as **tag SNPs** or **LD buddies**. Each tower in the Manhattan Plot represents a bunch of SNPs with varying degrees of r^2 to the causal variant.

Here's an example of what it looks like if we expand the plot to focus on a single skyscraper (shown here for the PER3 locus from the chronotype scan). The format of the plot is similar to a Manhattan Plot, except that it shows just a single megabase region of Chromosome 1, including one major peak of significant SNPs:

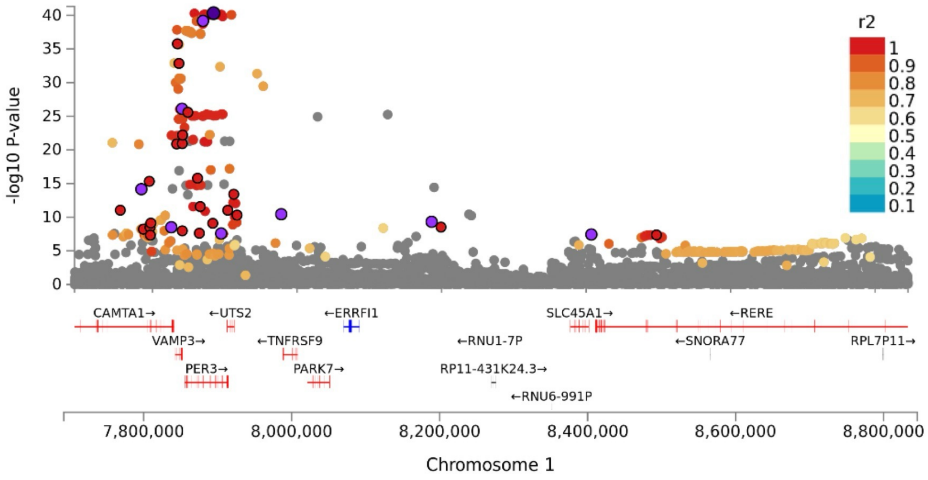


Figure 4.64: **A cluster of genome-wide significant SNPs for chronotype at the PER3 locus.** *The dark purple SNP at top is the lead (most significant) SNP; other significant SNPs are colored by LD with the lead SNP. Genes are indicated at bottom.* *From Figure S2c in Samuel Jones et al (2019) [Link] CC BY 4*

Although there are many significant SNPs in this particular skyscraper of signal, most of these have high r^2 with the lead SNP [775]. Thus, there is likely just one or at most a few causal SNPs in this region, accompanied by many significant (non-causal) tag SNPs that are in LD with the causal variants [k].

This example also hints at another major challenge in GWAS, the **gene-linking** problem. Why did I label this the "**PER3 locus**", and not, for example, after any of the other many genes in the region? For one thing, the signal is more-or-less positioned over the gene PER3, but we'll get back to a more precise answer in a few pages.

The genome-wide significance threshold. When you took your first statistics class, you were likely told that if you do an experiment, and find p is <0.05, that you can reject the null hypothesis. When the null hypothesis is true, this will lead you to (incorrectly!) reject the null hypothesis about 5% of the time; by tradition, this is often considered to be an acceptable error rate.

[j] *Recall from Chapter 2.3 that SNPs that are close together in the genome are typically in LD, meaning that their genotypes are correlated. In most regions of the genome, LD extends over roughly 10–100 KB, though this varies across the genome, and across populations.*

[k] *LD is both a blessing and a curse: It's helpful when we want to detect significant regions of the genome, because there are usually many SNPs pointing us toward each signal. But it's usually unclear which variant(s) are causal in any given region.*

But suppose that you test one million SNPs across the genome. Now, even if there is no true signal in the data at all, you can expect to reject the null at about 50,000 SNPs! This is clearly not useful. Instead, we'll need a much more stringent statistical threshold before we consider any particular SNP as significant in our GWAS.

This issue comes up any time in statistics that we perform many tests at the same time: if we do enough tests we can be sure to reject the null hypothesis by chance, even if the null is true for every single test. This issue is referred to as **multiple testing**. One standard way to deal with multiple testing is to apply something called a **Bonferroni correction**. This says that if we perform n independent tests, then we should adjust the significance threshold to be

$$0.05/n. \tag{4.70}$$

The Bonferroni correction ensures that in an experiment with n tests (where the null is true for all tests) we will have a 5% chance of rejecting the null for one-or-more test across the entire experiment.

Assuming that we test one million SNPs across the genome, the Bonferroni correction suggests we should set a significance threshold of

$$0.05/10^6 = 5 \times 10^{-8} \tag{4.71}$$

This threshold, 5×10^{-8}, is referred to as the **genome-wide significance threshold**. SNPs with p-values smaller than 5×10^{-8} are said to be **genome-wide significant** [1]. The actual number of tested SNPs varies a bit across studies, but people usually stick with the 5×10^{-8} threshold by convention [776] [777].

[1] *You can interpret this as follows: if you did a study that had absolutely no true signal, then you would have only a 5% chance of finding a genome-wide significant SNP anywhere in the genome.*

How much signal is there in my GWAS? One helpful diagnostic for GWAS data is known as a **QQ plot** (standing for quantile-quantile plot). The QQ plot allows to visually assess how much signal there is in the data. Remember that when the null hypothesis is true, the p-value has a uniform distribution between 0 and 1 (by the definition of a p-value). This means, for example, that 1% of p-values should be less than 0.01, and 0.01% should be less than $p = 10^{-4}$.

If we see more small p-values than expected under the null hypothesis, this suggests that the GWAS must be picking up signals of genetic associations. (Or we might worry about confounding effects of population structure, which we'll cover shortly.) The QQ plot is a clever visualization that allows us to easily assess if we have more signal than expected if the null hypothesis was true at all SNPs [m].

[m] *QQ plots are a great tool in general for data visualization when you are doing many tests in one analysis.*

The plot below shows this for the chronotype data. Here, the red line shows the expectation under the null; as you can see, the actual data fit the null distribution well up to ~ 2 on the x-axis; after that the data curve way up above the null line. In other words, around 99% of the data fit the null distribution pretty well, but around 1% of the data show a huge excess of large signals compared to the null expectation:

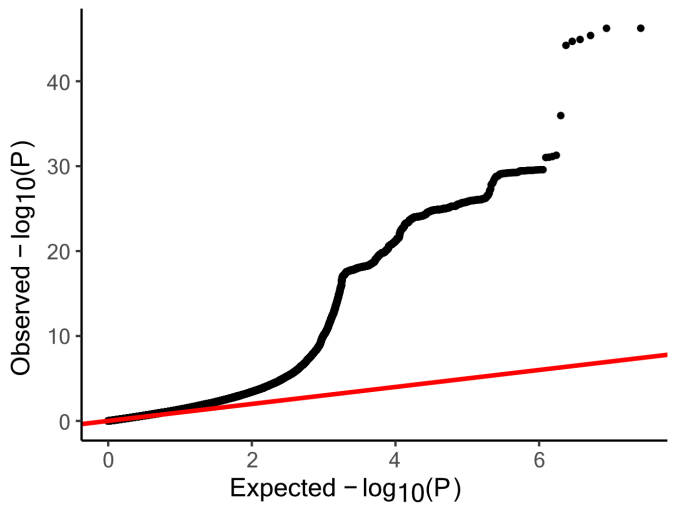


Figure 4.65: **QQ Plot for chronotype GWAS.** *The observed log p-values (in black) snake up above the red x=y line, indicating that there is more signal in the real data than expected under the null hypothesis.*
To draw this plot, we sort the observed p-values, and number them from $i = 1$ (most significant) to n (least significant). p_i is the p-value for the ith most significant SNP. For each SNP, the x-axis shows the expected -log p-values (-log10(i/(n+1)), and the y-axis shows the actual -log p-values (-log10(p_i)), respectively, for the ith SNP.

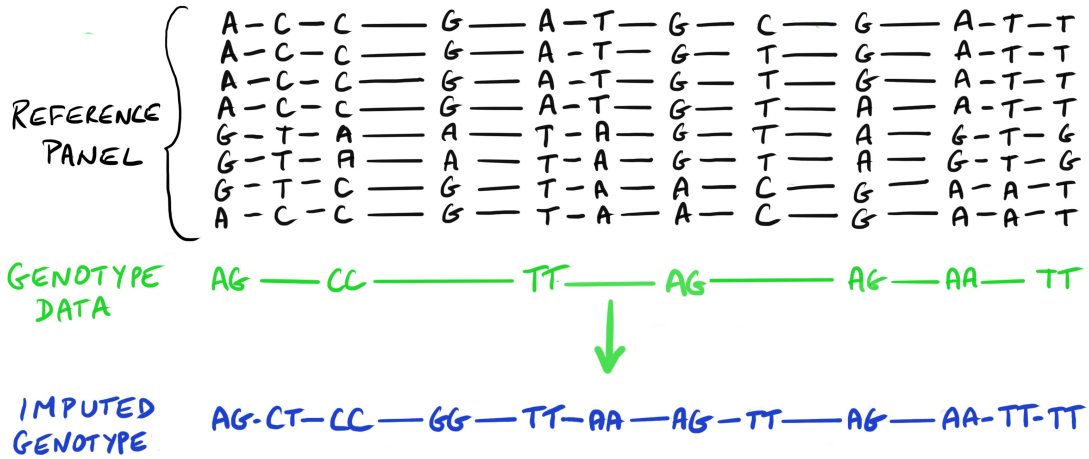
Credit: This figure was kindly drawn by Huisheng (Julie) Zhu using data from Samuel Jones et al (2019) [Link] CC BY 4

The trail of SNPs snaking up above the red x=y line here represents a mixture of many true causal variants, along with their LD buddies. As we'll see in the next chapter, for most complex traits, the bulk of the heritability is due to thousands of SNPs with very weak signals, that come up off the diagonal line here in the QQ-plot but do not achieve genome-wide significance at current sample sizes [778].

Optional details: Sequencing, genotyping, and imputation. There's another technical point that I've glossed over so far. For a big GWAS, it's traditionally been too expensive to perform whole genome sequencing on every individual (this is changing as sequencing costs drop).

Instead, most GWAS use **genotyping arrays** that measure SNP genotypes at a subset of common SNPs across the genome. (For background on sequencing and genotyping see Chapter 1.4.)

Next, we apply a technique called **genotype imputation** that uses LD to impute (predict) the genotype for each individual at unmeasured SNPs by lifting information over from a fully sequenced panel like the 1000 Genomes Project:



Box Figure. Sequence imputation from genotypes. *We start with a reference panel of known genome sequences, and limited SNP data from an individual (or many individuals in a GWAS). Imputation methods com-*

bine these to predict the full sequences in that individual. For more technical details on imputation see Chapter 2.3 [779].

In a typical GWAS study we'll start with genotype data at ~1 million SNPs, and then impute this up to about 10–15 million "common" SNPs (frequencies above 0.1%). Imputation works well for common SNPs but cannot predict very rare variants as these are not well represented in the reference panel. After imputation, we can perform testing on every SNP regardless of whether it was measured directly or imputed; a typical Manhattan Plot would include signals for all tested variants.

We've now covered many of the key technical aspects of GWAS analysis and interpretation; we need to cover one last important technical issue: population structure confounding (and how to fix it):

Population structure confounding. In GWAS we aim to identify genetic variants that influence trait outcomes [780].

But there is an important confounder to worry about (and control for): **population structure**. Allele frequencies often vary across ancestry groups; what if our average phenotype also varies across ancestry groups? When this happens, some SNPs could show **spurious associations** due to the variation in ancestry.

For an example of why this might matter, here's a famous thought experiment by Lander and Schork (1994) [781]:

> *"To give a light-hearted example, suppose that a would-be geneticist set out to study the "trait" of ability to eat with chopsticks in the San Francisco population by performing an association study with the HLA complex. The allele HLA-A1 would turn out to be positively associated with ability to use chopsticks not because immunological determinants play any role in manual dexterity, but simply because the allele HLA-A1 is more common among Asians than Caucasians."*

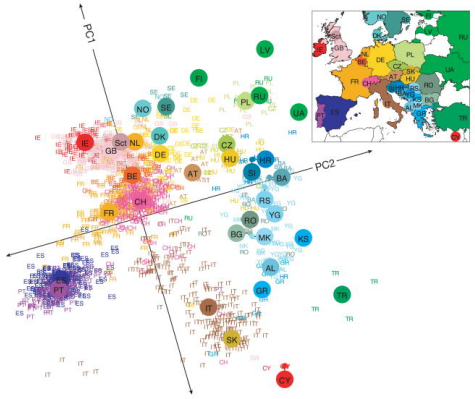


Figure 4.66: **Genetic structure of Europe.** *This shows the PCA projection of the genotype data for 1387 Europeans, colored according to their locations on the inset map.* ***Recall that the PC1 and PC2 show the two leading principal components (axes of variation) of population structure in the sample. For more about this plot and topic see Chapter 3.1.*** *Credit: Figure 1A from John Novembre et al (2008). [Link] Used with permission.*

You can see this illustrated graphically in the cartoons below, showing how uneven sampling of cases and controls from different communities (in **A**) or uneven sampling across continuous structure (in **B**) can result in spurious associations:

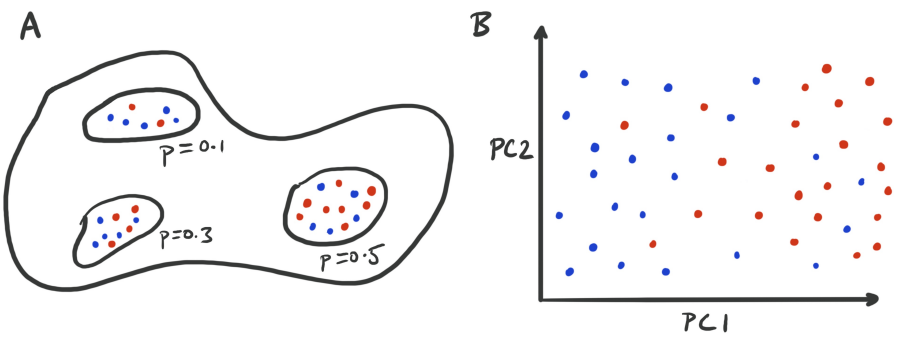


Figure 4.67: **Two models of structure confounding.**
A. *Cases and controls (red and blue) are sampled at different rates in three different cities. Allele frequencies also differ across the 3 cities, suggesting that higher frequency of the derived allele increases the risk of disease.*
B. *Cases and controls are projected into a PCA space (based on genotype data); the cases and controls are distributed differently along PC1 (but the same along PC2). Any SNP that is correlated with PC1 will also appear to be associated with case-control status.*

These types of scenarios can occur whenever environmental factors vary across ancestry groups. In practice this is very common: even nearby communities can differ dramatically in terms of environment, income, opportunities for education and employment, lifestyle and health. Thus, structure confounding is a serious concern in large modern GWAS [n].

To give one important example, the single most-widely studied cohort in human genetics is the **UK Biobank (UKB)**. The UKB is a diverse, broadly-sampled cohort from across Britain. About 80% of the cohort is described as having "white British" ancestry, while the remainder has wide-ranging ancestry including from continental Europe and most other major global populations [782].

To mitigate concerns about ancestry confounding, many analyses of the UKB data focus on the white British subset with the expectation that this group is more homogeneous [783].

But while analyzing just the "white British" subset does eliminate much of the population structure present in UKB, it doesn't get rid of all it. There is still highly significant genetic structure even within Britain [784], as well as strong regional variation in many phenotypes. This is illustrated in the plots below, showing that both genetic PC1 (left), and life expectancy (right) show strong north-south trends across Britain:

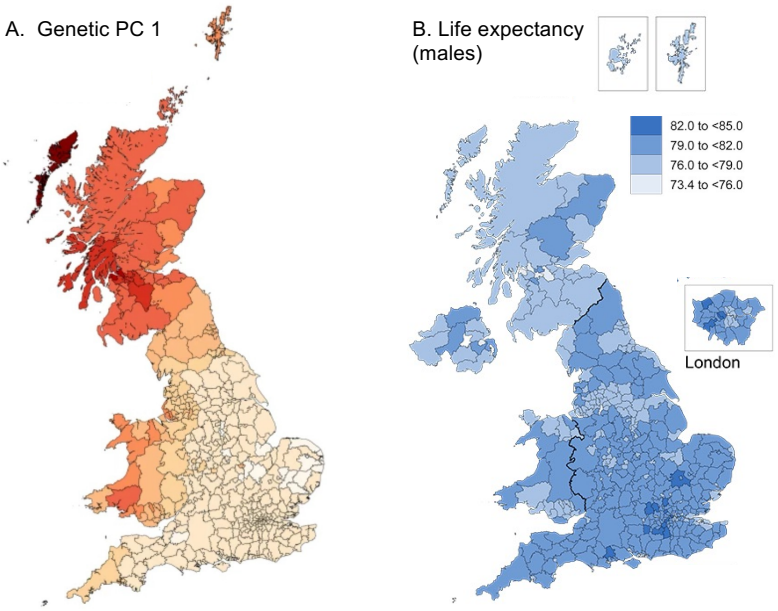


Figure 4.68: **Regional variation in genetic ancestry and phenotypes across Britain.** *Panel A shows regional average values of genetic PC1 across Britain. Panel B plots regional average values for life expectancy, showing that average life expectancy varies by almost 10 years across the UK.*

Credit: A. Modified from Figure 1 of Abdel Abdellaoui et al (2018) [Link], CC-BY-NC-ND 4.0. Data from UKB. B. Modified Figure 3 of [Link]. UK Open Government Licence v3.0. Data 2013-2015.

Given these patterns, **we should expect that if we performed GWAS for longevity in the UK Biobank, we should find signal at all SNPs that vary in allele frequency along PC1 – regardless of whether these SNPs have any causal role in longevity.** The same would be true for many other traits as life expectancy correlates with many facets of health, income, lifestyle, and education. Each of these also varies markedly across Britain [785] [o].

Correcting for population structure. We now have good tools for con-

[n] *Population structure confounding is a concern whenever environmental factors relevant to a phenotype covary with ancestry.*

[o] *While these examples are from Britain, we should expect that structure confounding is likely in any large cohort.*

trolling population structure, the most widely-used of which is **PCA correction**. Application of one of these approaches is an essential step in all GWAS; happily these provide good protection that works well for nearly all applications.

If you wish, you can read about these methods in the box:

Optional technical section on structure correction. We'll cover two main approaches here: (1) PCA-correction; and (2) family-based associations [786]. You can read about a third approach, using **linear mixed models (LMMs)**, in the endnote [787].

PCA correction uses the principal components of the genotype matrix as covariates in the regression. We start by performing PCA on the genotypes of the individuals in our sample to infer the major axes of population structure present in that data set (i.e, the PCs). If any of the PCs are correlated with the phenotype, this is likely to drive spurious associations.

The key idea is to include the leading PCs as covariates in the regression model [788]. If the phenotype is correlated with any of the PCs, these covariate terms can capture that correlation. So if a SNP was correlated with the phenotype only through its correlation with a PC, the PC covariate can now absorb the signal away from that SNP. We refer to this as "controlling for genotype PCs".

To implement this idea, recall that our basic regression model for a quantitative trait looks like this:

$$\underbrace{Y}_{\text{phenotype}} = \underbrace{\mu}_{\text{intercept}} + \underbrace{G_l \cdot \alpha_l}_{\text{SNP effect}} + \underbrace{C \cdot \delta}_{\text{covariates}} + \underbrace{\epsilon}_{\text{random error}} \tag{4.72}$$

Now we simply expand the covariates for individual i to include the first m PCs:

$$\text{covariates} = \underbrace{\text{PC}_1,\ \text{PC}_2,\ ...,\ \text{PC}_m}_{\text{first } m \text{ PCs}}\ ,\ \underbrace{\text{age},\ \text{sex},\ \text{batch}, ...}_{\text{other covariates}} \tag{4.73}$$

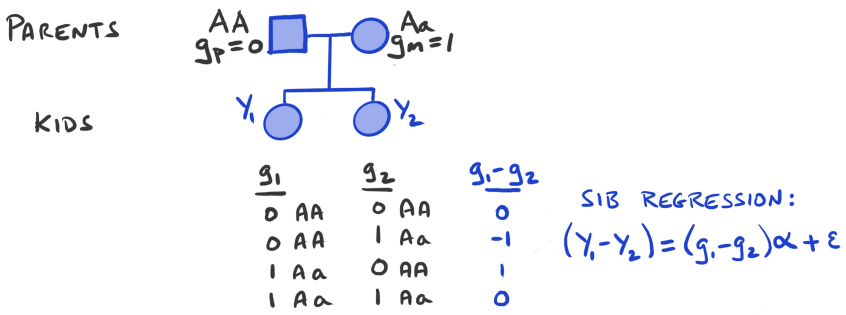
where PC_1 represents the loadings of each individual on the first PC, and so on. Importantly, in this model the δs capture the relationship between each PC and the phenotype – so we can soak up any relationship between phenotype and population structure into the PC covariates.

For example, in the picture above showing longevity across Britain, we can expect the δ for PC_1 to be negative as PC_1 increases going north, while longevity decreases. Additional PCs capture other axes of genetic variation, and any correlation they may have with phenotype. People will often use m=10 or more PCs for a sample like white British [789].

Family-based association tests. An entirely different solution to structure confounding is to perform specialized association tests *within families*. These tests can be extremely robust to confounding, and are often considered a **gold standard** [790] [791].

The clever idea in family-based designs is that regardless of overall population structure, *the transmission of alleles within families depends only on Mendel's rules*.

In the example below, if we condition on the paternal and maternal genotypes being $g_p = 0$ and $g_m = 1$, respectively, then there are four possible combinations of genotypes g_1 and g_2 for the two children [792], each with probability 1/4:



Box Figure. Sib regression controls for structure confounding. *In the example here, given the parental genotypes g_p and g_m, there are four possible combinations of genotypes g_1 and g_2 for the two kids. If there's no causal relationship between the SNP and phenotype Y, then the transmission of alleles is entirely independent of the phenotype difference between the sibs ($Y_1 - Y_2$), regardless of population structure.*

In this kind of test we ask if there's a correlation between genotype and phenotype *within* sibling pairs. For example, suppose you're studying height, and I tell you that $g_1 - g_2 = 1$ (that is, child 1 has one more alternative allele than child 2).

If there's no causal relationship between this SNP (or an LD partner) and height, then the fact that $g_1 - g_2 = 1$ tells you nothing about the relative height of the two sibs. But if I told you that the alternate allele actually *increases* height, now you should guess that child 1 is *taller* than child 2.

To formalize this, a bit of math shows the elegantly simple result [793] that

$$\mathrm{E}(Y_1 - Y_2) = (g_1 - g_2) \cdot \alpha \tag{4.74}$$

where α is the effect size of the SNP on trait Y. The derivation uses the fact that the sibling genotypes are Mendelian random draws from the same parents – but *we don't actually need to know the parents' genotypes*. Given this, we can construct a test of association called **sib regression** that fits a regression model

$$Y_1 - Y_2 = (g_1 - g_2) \cdot \alpha + \epsilon \tag{4.75}$$

across sibling pairs. Since this test depends only on Mendelian transmission within families, it is valid regardless of any structure confounding in the overall sample!

Structure corrections in practice. You might imagine from the above that people would always prefer to use family-based tests of association. However, it's generally much more expensive and logistically challenging to recruit families for a large study. For this reason, most large GWAS use unrelated individuals as a primary analysis, sometimes coupled with follow-up validation in a smaller family cohort. For example, the UK Biobank has ~22,000 sibling pairs in a total sample of 500,000 people.

PCA corrections work well for most applications, and significant hits are generally highly reliable (but with important caveats). Family studies generally confirm significant hits, but show that effect sizes may be subject to a variety of biases. This is especially true for traits with strong population structure or assortative mating, or where there are strong parental influences on phenotype [794]. For example, both height and income/educational attainment are subject to strong assortative mating. Additionally, parents and siblings have strong environmental impacts on behavioral traits like educational attainment.

A closely related issue is that for some traits the environment can be inherited across many generations, most notably for social traits like wealth, culture, and social status (we saw an example of this in the

last chapter [795]). This creates a confounding between recent population structure and genetics that may be difficult to fully resolve, especially for polygenic scores and heritability estimation that include rare variant effects [796].

Third, some analyses have looked at whether polygenic scores differ among populations or over time (in ancient DNA). This has been used to study polygenic adaptation, but it requires very special caution as even the slightest amount of uncorrected confounding can be confused for signal [797]

Well done! You've now made it through many of the key technical aspects of GWAS analysis. We can now turn our attention to what we find in GWAS.

What do we find in GWAS? Let's go learn some biology! So, after all this hard work, we get a Manhattan plot. How can we now use this to learn something about the biology of the trait? We'll address this question over the next few chapters, but we start with some basic insights.

The first analysis is usually to see **what kinds of genes show up as top hits**. Sometimes these map clearly onto known biological processes. For **chronotype** [p], many of the top genes are especially interpretable – this is thanks to decades of research in flies and rodents, culminating in the 2017 Nobel Prize for research on the genes that control circadian rhythms [Link].

The central circuit that controls circadian rhythms in animals is an activation-repression loop involving four main genes with paralogs: BMAL1, CLOCK, CRY and PER (shown at right). The GWAS finds hits at most of these master regulators including BMAL1; CRY1; and all three main PER paralogs: PER1, PER2, and PER3 [798].

But in addition to the expected hits, the GWAS identifies many other master regulators whose roles in circadian rhythms are less well understood. And while it's a nice proof-of-concept to find genes that we expect, the real gold in the Manhattan plot lies in discovering hits in genes that we did not expect, as these point to new biology to be learned.

Among the other hits for chronotype are the genes FBXL3, RGS16, and HCRTR2. *Each gene illustrates a different mode by which a gene might influence circadian pathways*:
• FBXL3 degrades the circadian proteins CRY1 and CRY2 to de-repress CLOCK/BMAL1; a mutation in mice leads to a 27 hour cycle [799];
• RGS16 coordinates the circadian cycling of thousands of cells in a key region of the brain that controls the overall body clock [800];
• HCRTR2 is a receptor for the signaling peptide hypocretin, which regulates sleep-wake cycles in the brain; mutations in HCRTR2 can cause the sleep disorder narcolepsy [801].

But for many other hits in this GWAS it's unclear what role the genes may play in chronotype. **Through this type of analysis, GWAS provides an entrée into diverse aspects of the biology of a trait** [q]. It's usually un-

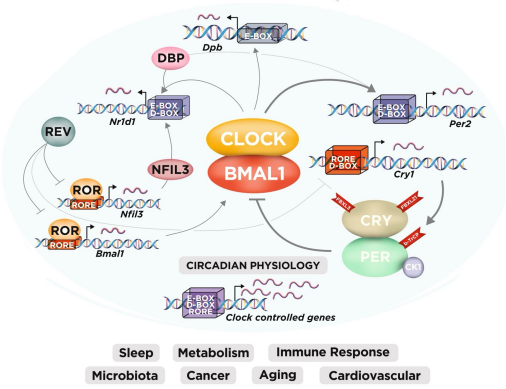


Figure 4.69: **Core gene circuit of circadian rhythm.** *The central circuit is encoded by the CLOCK and BMAL1 transcription factors; these activate CRY and PER which in turn repress CLOCK and BMAL1, causing expression to cycle across the day. This central timekeeping circuit regulates essential downstream processes listed below the main figure.* *Modified Figure 2, Filipa Rijo-Ferreira and Joseph Takahashi [Link]. CC License?*

[q] *In Chapter 4.9 we'll tackle something you may be wondering about: What determines which genes we find (and don't find) in GWAS, and their rankings [Link]?*

clear at first why any given gene shows up in a particular GWAS, and this can prompt painstaking but essential functional work to determine why that gene is relevant to the trait.

Enrichment analyses. Aside from inspecting the top hits individually, we can also look at what types of properties are shared among the significant hits [r].

[r] *Enrichment analyses can point us to the main pathways and cell types that control variation in a trait.*

First, we can test whether the hits are **enriched near genes involved in specific biological processes**. For an example, let's take a look at a GWAS studying the **hemoglobin content of red blood cells** in 450,000 individuals from UK Biobank, performed by Mineto Ota and others from my lab [802]. This GWAS revealed 634 independent genome-wide significant signals.

Hemoglobin is a multi-protein complex expressed by red blood cells that is responsible for oxygen transport to all tissues of the body. So it's no surprise to see GWAS hits overlapping the hemoglobin genes on Chromosome 16. What other processes might be relevant for determining hemoglobin levels?

To address this, Ota first identified the nearest gene to each genome-wide significant hit. He labeled each gene according to its classification in a database called Gene Ontology (GO) which reports gene functions according to a standardized classification scheme [803]. He then asked whether any GO categories were enriched (i.e., appeared more often near significant hits) than expected given the overall genome-wide distribution of GO functions.

As you can see below, most of the hits can be assigned to one of four key functions related to the maturation of red blood cells or synthesis of hemoglobin (the differently colored skyscrapers below) [804]:

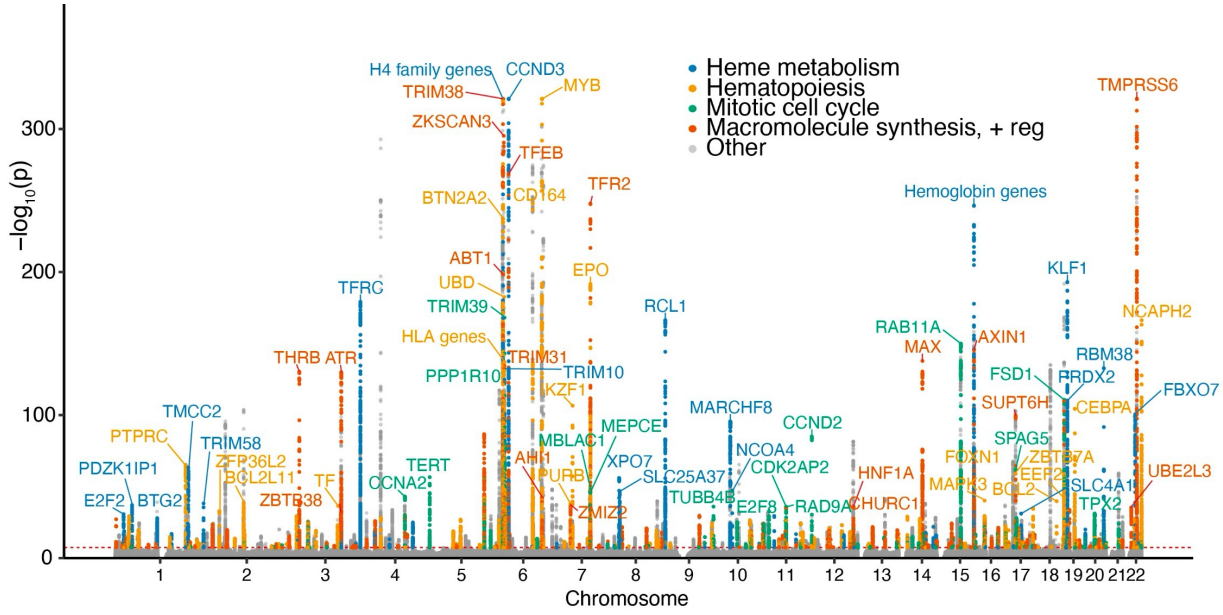


Figure 4.70: Manhattan Plot of hemoglobin levels (MCH). *This study identified 634 independent genome-wide significant signals in a sample of 450,000 individuals from the UK Biobank. The lead hits are labeled with proposed causal genes, and colored according to biological function (see legend in plot).* *Credit: Figure 2A from Mineto Ota et al (2025). [Link] CC-BY 4.0.*

We can also ask **what cell types** are most relevant to variation in a trait. Different cell types and organ systems are responsible for different human traits, and one important use for GWAS is to help us identify critical cell types for specific diseases. For example, we would expect that the genes that show up in a GWAS of hemoglobin should be genes that are active in the red blood cell lineage (spoiler alert: they are [805]).

What about chronotype? The analysis below tests whether the GWAS signal for chronotype is enriched near genes expressed in each cell type:

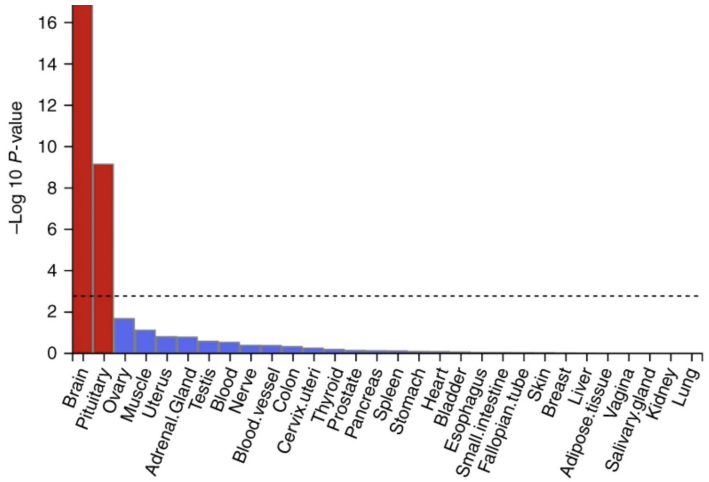


Figure 4.71: **Cell type enrichments for chronotype.** *This analysis tests, for each tissue, whether genes with significant signal in the chronotype GWAS are enriched near genes expressed in that tissue. Brain and pituitary are both highly significant (y-axis), while no other tissue passes Bonferroni testing (dotted line).* *Note: this analysis uses heritability enrichment, estimated by MAGMA, rather than only significant hits. Figure 4a in Samuel Jones et al (2019) [Link] CC BY 4*

This analysis correctly identifies brain and pituitary as the primary regulators of circadian rhythms.

So far, these examples merely serve as a proof-of-principle: you won't get any prizes for showing that hemoglobin is controlled by red blood cells or that sleep is controlled by the brain. But, much more importantly, GWAS made a critical contribution to our understanding of **Alzheimer's Disease (AD)** in the mid 2010s.

AD is a serious neurodegenerative disease and the major cause of dementia in older adults. It's characterized by the accumulation of extracellular proteins called amyloids, eventually leading to severe atrophy of the brain. Despite the importance of the disease, and intense study, for many years it was unclear which cell types in the brain were most important for AD onset and progression.

This changed as early GWAS studies for AD revealed that many of the top GWAS genes show specific up-regulation in a type of brain cell called **microglia**. In contrast, most of the associated genes show lower expression in other brain cell types including neurons:

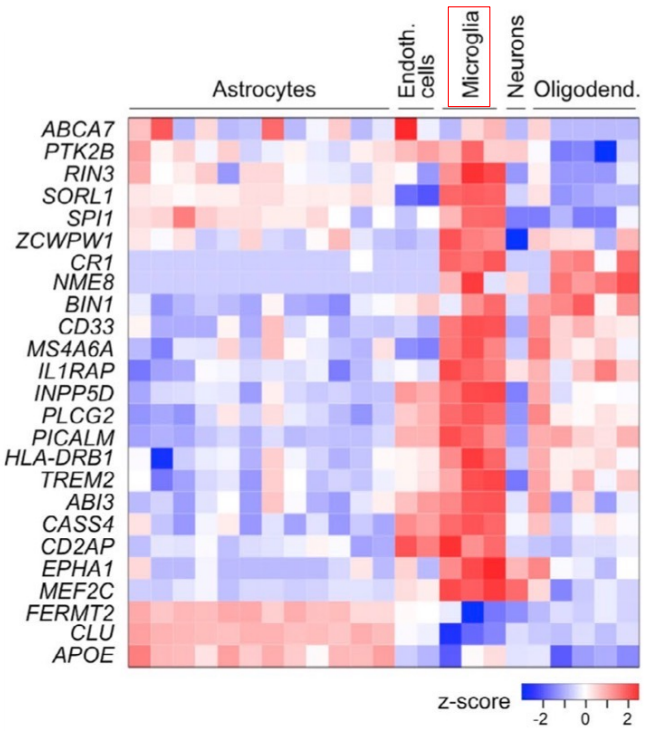


Figure 4.72: **Many Alzheimer's GWAS genes are highly expressed (red) in microglia.** *The rows show different genes identified in GWAS for AD; the columns show gene expression for relevant cell types isolated from human central nervous system. Note that most of the hits are colored red in microglia, indicating that these genes are up-regulated in that cell type. Meanwhile, a smaller number of genes show specific up-regulation in astrocytes.* *Figure 1A of David Hansen et al (2018) [Link] CC BY-NC-SA 4.0*

The finding that many of these Alzheimer's-associated genes are up-regulated in microglia suggests that they may have specific important functions in this cell type, and hints that dysfunction of microglia may be a major causal factor in AD. The role of microglia is now widely accepted; microglia act as the major form of immune defense in the brain and are responsible for cleaning up extracellular plaques and other detritus that can lead to AD [806].

As we go on, we'll see many more examples of applications of GWAS. But before this, it's helpful to dig a bit deeper into the types of variants that drive GWAS signals.

How do SNPs affect complex traits? Roughly speaking, genomes encode two kinds of information: genes encode protein sequences, and regulatory regions tell each cell how much of each protein to produce, depending on cell type and context. SNPs can alter either type of encoded information, sometimes with highly deleterious consequences [s].

As we have seen, protein coding changes are major drivers for monogenic diseases and cancer. But around 90% of GWAS signals are driven by regulatory effects. These change the expression (or splicing) of nearby genes.

To think about how this might happen, remember that regulatory information is encoded in DNA sequences called promoters and enhancers. Proteins called **transcription factors (TFs)** bind these sequences to drive expression of nearby genes (see Panel A, below). This usually happens by physical looping of the DNA to bring the transcription factors into contact with the promoter.

[s] *For background on these topics see Chapters 1.2 and 1.3. For more on protein-coding changes in monogenic diseases and cancer see Chapters 4.2 and 4.3. And we'll do a deeper dive into regulatory genomics in Chapter 4.7.*

But in Panel B, the alternate allele at a SNP disrupts the binding site of the TF, blocking it binding there. This results in reduced expression of the nearby gene:

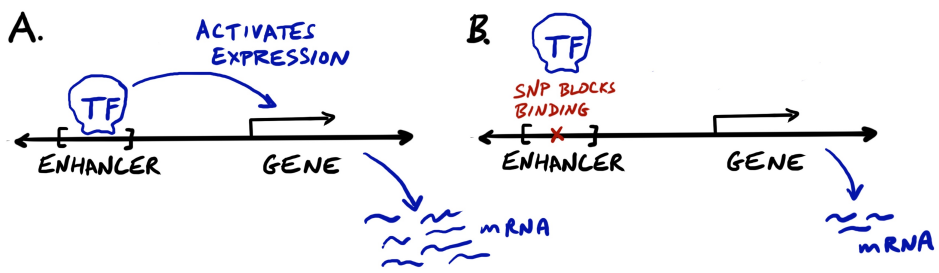


Figure 4.73: **Basic model for a regulatory variant.** *In* **A**, *the TF binds a DNA sequence motif in an enhancer, and activates expression of a gene. In* **B** *an alternative allele eliminates the TF binding motif. The TF fails to bind, lowering gene expression.*
A SNP that changes expression is known as an ***eQTL (expression quantitative trait locus)****; you'll read more about these in Chapter 4.7.*

Why does this matter? Cells are highly precise machines, and any change in gene expression can affect the functions of the cell in deleterious ways that, in turn, can affect traits. For example, remember that the CRY and PER genes are part of an essential circuit regulating circadian rhythms. You can imagine that even small changes in expression of these genes might impact the output of this circuit. We'll also show an important example for hemoglobin, shortly.

We can sketch a simple quantitative model:

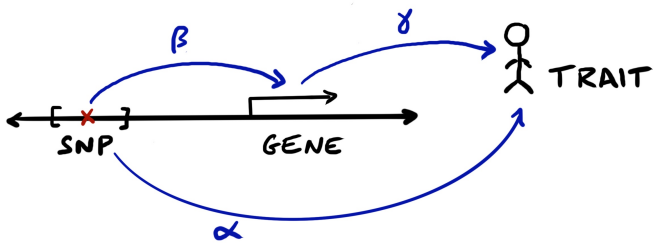


Figure 4.74: **The causal pathway from SNP→Gene→Trait.** *In GWAS we estimate α, which is the SNP→Trait effect size.*

As above, α measures the effect size of the SNP on a quantitative trait: for example, the alternate allele at this SNP might increase hemoglobin levels by $\alpha = 0.1$ standard deviations (SDs).

Now, β measures the effect of that SNP on gene expression in the relevant cell type: for example, the alternate allele might increase expression by $\beta = 5\%$.

Lastly, γ measures the effect of a unit change of expression of this gene on the trait: for example, increasing expression of this gene by 100% might increase hemoglobin by $\gamma = 2.0$ SDs. For understanding the biology of a trait γ is the thing we really care about. We define the **importance of a gene for a trait as** γ^2. And **the sign of** γ **tells us the direction of effect**: whether an increase in expression will increase, or decrease the trait [t].

[t] *We'll return to these topics in Chapters 4.8 and 4.9.*

This suggests a simple relationship:

$$\alpha = \beta \times \gamma. \tag{4.76}$$

We can develop a better model by allowing the gene effect γ to depend on β. This is a **gene dose response curve (GDRC)** ("gene dose" refers to

the expression level of a gene) [807]. For each gene we assume some relationship between expression and the expected trait value:

$$\alpha = \gamma(\beta). \tag{4.77}$$

where γ is a continuous, usually monotonic, function of β. Here's how this might look:

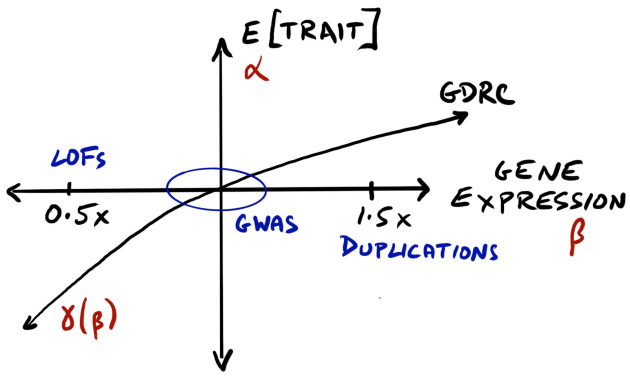


Figure 4.75: **The GDRC imagines a relationship between expression of a relevant gene and the expected phenotype.** *GWAS variants (eQTLs) typically have much small effects in both axes. More significant lesions to the genome (which are usually very rare) can have larger effect sizes: heterozygous Loss of Function (LoF) mutations reduce expression by 50%, while a heterozygous duplication would often increase it by 50%.*

As you see above, when we do GWAS we usually find variants with small effects on expression, and small effects on the trait. One challenge with these is that we don't usually know β (effect of SNP on gene). This means that standard GWAS can tell us that a gene is relevant for a trait, but it doesn't give us a precise ranking of gene importances, or the direction of effect [u] [808].

In the next section we'll talk about how rare large-effect mutations – especially LoFs – can help with this problem, but usually with much lower power.

[u] *When we want to learn about trait biology we usually care about γ (which we don't measure directly in GWAS). But for genetic prediction (Chapter 4.6) we care about SNP effect sizes, α.*

Optional Detour: Fine Mapping and Gene Linking. GWAS gives us significant SNPs, but to learn about biology we usually care about genes. *What gene is responsible for driving a hit and in what cell type?* This is very often a first step into understanding mechanisms.

But to get to these questions, we need to tackle two twin problems that GWAS analysts spend a lot of time on:

• Fine mapping: What is/are the causal variant(s)?

• Gene linking: What is the relevant gene? You can get a sense for both of these challenges in the PER3 example above (Figure 4.64).

Fine Mapping. Our first problem is that, due to LD, many SNPs may have similar significance levels. This means that it is difficult to know which SNP is causal (or indeed whether there may be more than one causal SNP). *Fine mapping* refers to the goal of identifying causal SNPs.

So far, we have been assessing GWAS signals using so-called frequentist approaches (i.e., with p-values). But for fine mapping it's more convenient to use Bayesian methods. It's beyond our scope to describe Bayesian methods in detail, but you can think of these as being useful for quantifying evidence for different models: for example, if there are 100 SNPs under a GWAS peak, what is the probability for each

SNP being the causal driver of the signal?

This is quite a complex area and if you need to do fine mapping you should read more in these reviews and methods papers [809]. For now I'll give you some simple rules of thumb for thinking about data:

- **How do p-values relate to the probability that a particular SNP is causal?** We measure the relative evidence for each SNP being causal using something called a Bayes Factor (BF). For a single SNP, a BF measures how much more likely are the GWAS data for a model where this SNP is causal versus a model with no causal SNPs.

  Importantly, the ratio of BFs for two SNPs equals the ratio of the probabilities that each of them is causal [810]. So for example, if the BF on SNP A is 10× higher than on SNP B, then we can interpret this as saying that the probability that SNP A is causal is 10× higher than the probability that B is causal (assuming exactly one causal variant in the region).

  A very useful rule of thumb is that – under simplifying assumptions – *the ratio of BFs between SNPs A and B equals the ratio of p-values* **[check this; add endnote]** [811]. This is extremely helpful when we look at a Manhattan Plot. If for SNP A, -log10(p)=40; for SNP B, -log10(p)=39; and for SNP C, -log10(p)=38, then A is 10× more likely than B to be causal, and 100× more likely than C. So if there's only one causal SNP it will usually be fairly close to the peak of the Manhattan Plot.

- **How can we know if there is more than one causal variant?** Early work assumed that causal variants were rare, so that each peak would likely contain only a single causal variant. We now know that they often contain multiple causal variants. Modern computational methods can tackle this using the SNP p-values and the matrix of LD values among all SNPs to identify distinct causal variants [812].

  We can also get a gut sense of this by examining the LD (r^2, the squared correlation) between each SNP and the most significant SNP in a region. *When there is only a single causal SNP, the -log10(p) on any given SNP is approximately r^2 times the largest -log10(p)* [813]. So for example if the lead SNP in a region has a -log10(p)=50, then a SNP with $r^2 = 0.6$ to the lead SNP should have a -log10(p) = 30. If it's much higher than 30, this suggests there is likely a second causal variant.

In addition to the GWAS data, we can also use functional information to point to likely causal SNPs. For example, we might ask: Do any of the candidate SNPs lies in a region of active chromatin or a predicted transcription factor binding site in a relevant cell type? Are any of the lead SNPs also eQTLs? Do any of the lead SNPs show signals in reporter assays [814]?

Variant→gene linking. What is the relevant gene? Remember that only about 10% of GWAS hits affect protein-coding regions. For the rest, we need to figure out which nearby gene is most likely to be causal.

For noncoding variants, we would usually think in terms of the causal SNP changing cis-regulation (or changing splicing) of a nearby gene. Causal variants usually act over distances up to a few tens of kilobases, though in rare cases across distances up to a megabase. If these concepts are unfamiliar to you, you can learn more in Chapter 4.7; at that point we'll also cover the famous story of how long-range regulation at the FTO locus was discovered.

People tackle the gene-linking problem with a combination of methods and heuristics; this is not a completely solved problem and is a highly active research area [815]:

Variant-first approaches: After fine-mapping the most likely causal variant, we can use a variety of methods to identify which gene it may be regulating:

- **nearest gene:** Most cis-regulation is short-range, and in the absence of other information your best

bet is the nearest gene. Even in advanced prediction methods, the nearest-gene feature is often the single most important feature;

- **functional genomics queries:** More advanced methods aim to link variants to genes by functional evidence: e.g., that the SNP is an eQTL for a nearby gene, or that chromatin looping data show that it physically interacts with particular nearby promoter;
- **targeted functional experiments:** Newer approaches use targeted experiments to assess whether a SNP impacts expression of nearby genes. One approach supposes that the SNP may sit within an enhancer, and uses CRISP silencing to repress the region around the enhancer. One can then test whether this affects expression. More specific (but currently more technically challenging) is to precisely edit the nucleotides of a candidate SNP to compare the effects of both alleles on the same genome background [816]. These experimental methods are conceptually appealing but they do require assumptions about the relevant cell type and conditions.

Gene function-based approaches:

- **nearby functional candidates:** An orthogonal, powerful approach is to use information about the functions of nearby genes. In Figure 4.64 above (the fine-mapping example for chronotype), one gene directly under the peak is PER3, which forms a critical part of the chronotype circuit. This seems unlikely to be a coincidence, suggesting that PER3 is highly likely the causal gene here. As an alternative to this vibes-forward approach, recent work focuses on quantitative methods to identify features that are shared among genes within GWAS peaks (e.g., gene functions and expression patterns across tissues) to pinpoint the most likely causal genes [817].

Burden tests for rare protein-coding variants. Even though most GWAS signals are caused by noncoding variants, we can still learn a lot from rare protein-coding variants. Compared to the common noncoding SNPs we often detect in GWAS, these often have very *large-effect sizes*, and they also tell us about the *direction of effects of genes on traits* [v] [818].

[v] ***Natural selection generally holds large-effect variants at low frequencies.*** *This principle has important implications for GWAS, and we'll study it in detail in Chapter 4.9.*

This is illustrated in a famous 2006 study of **LDL cholesterol levels**, led by Jonathan Cohen and Helen Hobbs. High levels of LDL, known as "the bad cholesterol", are a major risk for cardiovascular disease, heart attacks and strokes. Prior work had identified the gene **PCSK9** as a regulator of LDL levels but the precise function was unclear.

Gene sequencing by Cohen and Hobbs found two PCSK9 nonsense mutations segregating in the African American population at appreciable frequencies. Both mutations were predicted to render the PCSK9 protein nonfunctional (i.e., these are Loss-of-Function (LoF) mutations). They then genotyped these two mutations in a larger cohort, finding that 2.6% of their sample were heterozygotes for an LoF. Among these carriers, *they found a 28% reduction in mean LDL, and an 88% reduction in CHD*:

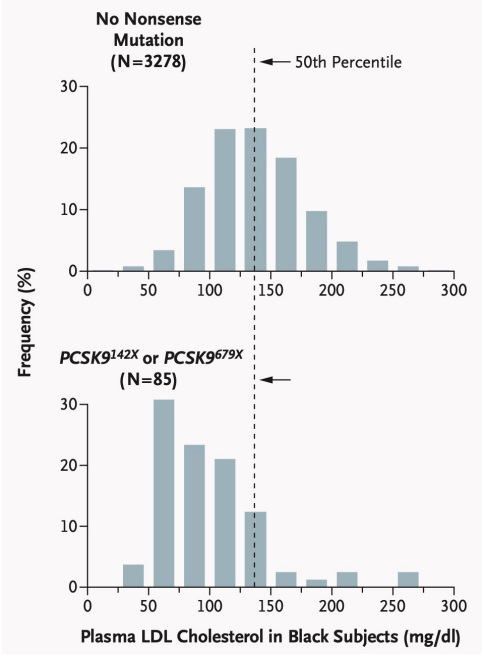


Figure 4.76: **LDL levels in persons with or without PCSK9 nonsense mutations.** *Notice that carriers of PCSK9 mutations have dramatically lower LDL levels.* *Credit: Figure 1 Jonathan Cohen et al (2006) [Link] Request Permission*

We expect that heterozygous carriers of LoF mutations should express the gene at about 50% of normal levels. Therefore, *this work showed that reduced expression of PCSK9 has a causal influence on lowering LDL levels* [819].

This seminal discovery led to the development of **a major class of drugs called PCSK9 inhibitors**. Inspired by the naturally occurring loss-of-function mutations, PCSK9 inhibitors lower PCSK9 protein levels, and were approved by the FDA in 2015 to treat high LDL cholesterol [820].

The PCSK9 study illustrates three important ways in which rare variant studies differ from conventional GWAS:
• The authors achieved power by jointly analyzing multiple rare variants predicted to have similar functional effects – i.e., do carriers of *any* LoF mutation differ significantly from non-carriers;
• The direction of effect of PCSK9 on LDL levels was immediately clear: reductions in PCSK9 expression decrease LDL;
• The magnitude of effect of LoFs on the trait was very large compared to the effects of common variants found by standard GWAS.

LoF burden tests. We can generalize the PCSK9 analysis to any gene with segregating LoFs. In short, the analysis tests whether the mean phenotype in individuals that carry an LoF mutation differs from individuals who do not. This is known as a **burden test**.

For example, I showed you the common-variant GWAS analysis of hemoglobin (Figure 4.70), which identified 634 independent signals. We can also analyze hemoglobin using genome-wide burden tests in UK Biobank [821]. The results are shown as a *volcano plot*, where each dot is a gene, and the plot shows effect size, on the x-axis, against significance, on the y-axis:

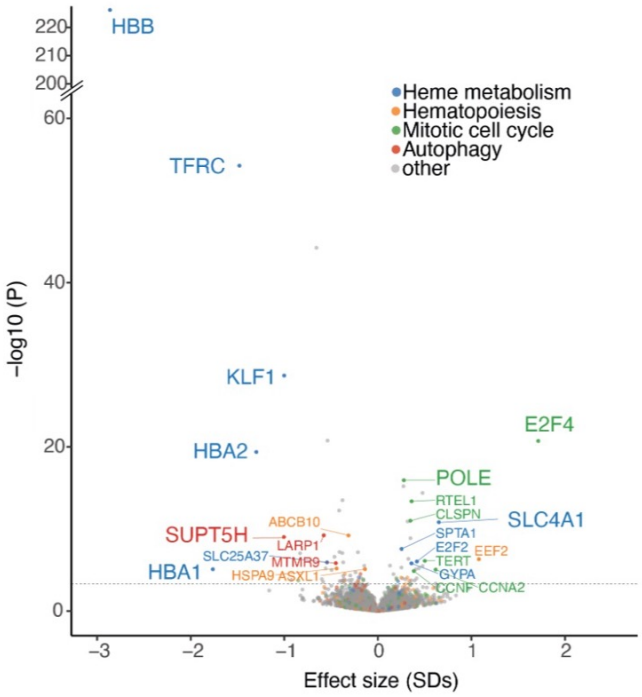


Figure 4.77: **LoF burden tests for adult hemoglobin levels (MCH).** *The x-axis shows the estimated effect size of LoFs; the y-axis shows significance of each gene. Genes are colored by functional category. The dotted line indicates 10% FDR.* *Credit: Modified from Figure 2b Mineto Ota et al (2025) [Link] CC-BY-4*

Compared to normal GWAS, we can interpret the raw results more directly. For example, LoF mutations in the core hemoglobin genes HBA1, HBA2, and HBB all reduce hemoglobin levels by about 1-3 standard deviations each, while mutations in the erythroid transcription factor E2F4 increase hemoglobin levels by nearly 2 SDs. In contrast, conventional GWAS is harder to interpret, but finds many more signals. Despite the differences, both analyses highlight similar pathways (compare to Figure 4.70).

We close this chapter with one of the most inspiring applications of GWAS.

Figure 4.78: **A representation of the pain of sickle cell crisis**, *by the artist Hertz Nazaire who died from complications of sickle cell disease in 2020 at age 48.* *For a moving personal account by Hertz Nazaire, see [Link].*

Case study: BCL11A and sickle cell treatment. Sickle cell disease is a devastating disease that affects individuals who inherit two defective copies of the HBB gene which encodes a component of the hemoglobin molecule. (You may recall from Chapter 2.6 that HBB mutations occur at relatively high frequencies in central and western Africa due to balancing selection, as heterozygous carriers are protected against malaria.)

Red blood cells in affected individuals are liable to collapse into a deformed shape known as a "sickle". The sickled cells lack the elasticity of normal red blood cells and can block blood vessels, causing recurring episodes of intense pain and tissue damage known as *sickle cell crisis*. Meanwhile degradation of sickled cells in the spleen leads to severe anemia, as well as poisoning from the release of heme molecules into the bloodstream.

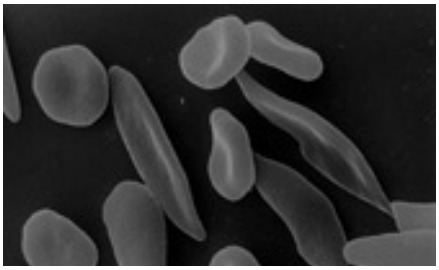


Figure 4.79: **Micrograph of sickled and normal red blood cells.** *NIDDK [Link] Public Domain.*

Crucially, the symptoms of sickle cell disease don't usually start until around 6 months of age. This is because infants produce an alternate form of hemoglobin called fetal hemoglobin, HbF. The HbF molecule is encoded, in part, by the genes HBG1 and HBG2, instead of the adult gene

HBB (mutated in sickle cell) [822]. After birth, fetal hemoglobin is gradually replaced by adult hemoglobin, leading to symptoms.

But research in the 1990s showed that some patients continue to express low levels of fetal hemoglobin, and that this is highly protective against death from sickle cell disease. In the plot below, the upper dotted line shows the survival of patients with high HbF expression (top 25%) compared to the rest (solid line) [823]:

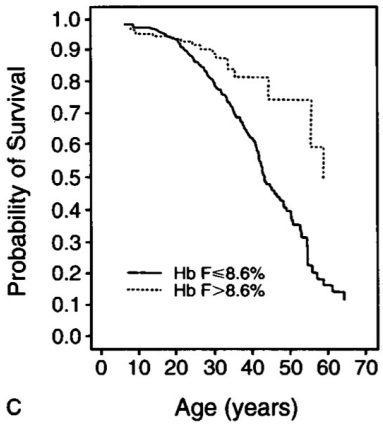


Figure 4.80: **Survival for sickle cell patients as a function of naturally circulating HbF (1994).** *The upper line shows survival curves for patients with HbF expression in the top 25%, compared to the rest.* *Credit: Figure 1c from Orah Platt et al (1994) [Link]*

Would it somehow be possible to maintain fetal hemoglobin expression throughout life? What genes could do this?

With this logic in mind, there was great interest in performing GWAS for fetal hemoglobin levels. The first studies of this in 2007 and 2008, identified several hits, including one in a gene called BCL11A [824]. You can see an updated Manhattan Plot (from 2025) here:

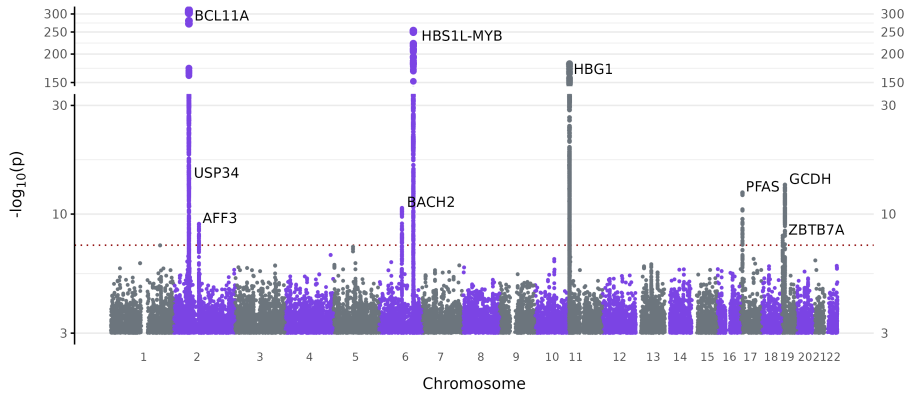


Figure 4.81: **Manhattan Plot for fetal hemoglobin (HbF).** *The top hit, BCL11A, led to a cure for sickle cell disease.* *Note the nonlinear scaling on the y-axis. SNPs with $p > 10^{-3}$ not shown. Top signals include BCL11A; MYB, a master regulator in blood cell development; and at the HBG1 gene itself. Credit: Figure kindly provided by Xiaoheng Cheng and Vijay Sankaran; redrawn from Figure 1B of Chun-Jie Guo et al (2025) [Link] CC BY 4*

Follow-up studies narrowed in on the intriguing signal in the intron of BCL11A [825]. BCL11A is a transcription factor that plays a role as a master regulator in blood and brain development. In a 2008 paper, Vijay Sankaran and colleagues showed that the SNP that is the top hit in the GWAS signal lies in a BCL11A enhancer, and results in a 3-fold difference in gene expression of BCL11A in erythroid progenitors (the cells that develop into red blood cells). Next, they showed that BCL11A is essential for regulating the switch to adult hemoglobin:

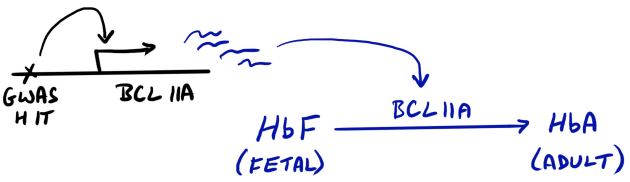


Figure 4.82: **BCL11A is required for the switch from fetal to adult hemoglobin.** *The GWAS hit lies in an erythroid-specific enhancer. The alternate allele reduces BCL11A expression, reducing the transition from fetal-to-adult hemoglobin.*

This was wildly exciting: would it be possible to turn off BCL11A, to stop the switch from fetal hemoglobin? BCL11A is essential in many tissues, so we couldn't want to eliminate it everywhere, but the GWAS signal pointed to a specific enhancer that was active in precisely the correct cells. Could one inactivate that specific enhancer only?

During the following years, there was a great deal of work on how to target BCL11A in sickle cell patients [826]. One approach extracts the patient's own blood-forming stem cells from bone marrow, and uses CRISPR-Cas9 to delete the critical BCL11A enhancer. After chemotherapy to eliminate the existing stem cells, the edited stem cells are reimplanted back into the patient. From this point on, the patient should express mainly fetal hemoglobin.

This approach has worked just astonishingly well – close to an absolute cure for most patients:

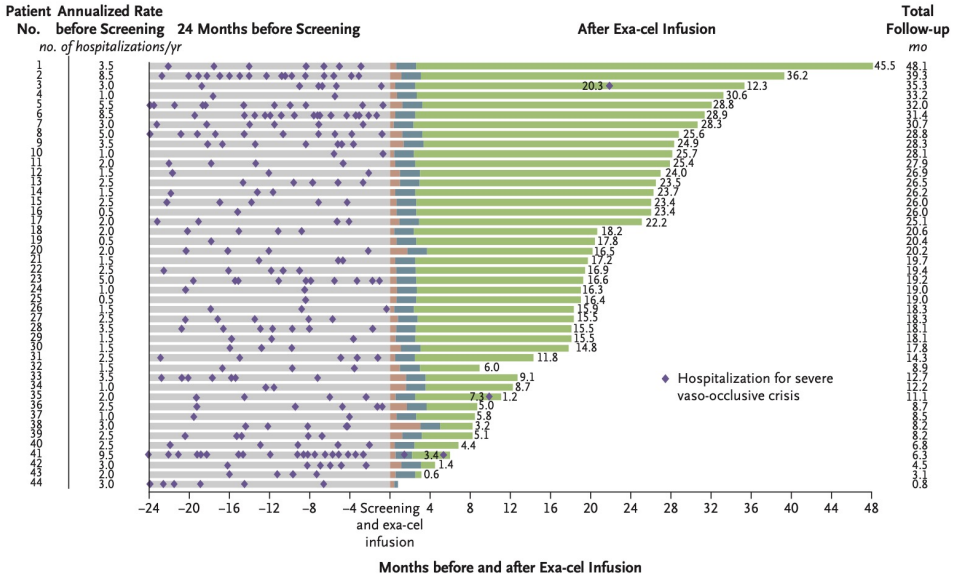


Figure 4.83: **Clinical trial results of BCL11A enhancer editing.** *Each row is a different patient. Purple diamonds indicate hospitalization for sickle cell crises before treatment (gray) and after treatment (green). Hospitalizations were eliminated for most patients post-treatment.*

Credit: Figure 1b Haydar Frangoul et al (2024) [Link]

In 2023, based on clinical trial data, two BCL11A-based therapies were approved by the FDA [827].

This inspiring story shows the power of GWAS to identify genes that reveal new mechanisms of disease (and regulation of disease), that can serve as therapeutic targets. Unfortunately, the story is not complete yet, as the therapies are still extravagantly expensive (both have list prices in excess of $2M in 2026), so there is still active work on more affordable alternatives.

Well done! You've officially reached the end of the beginning. You're an expert in the basics of GWAS! In the next two chapters we'll dig deeper into two key topics: polygenic models, and functional genomics of complex traits.

Notes and References.

[757]Recent estimates of common SNP heritability (which probably run slightly low as they don't include low-frequency variants) are around 12-21% , while twin heritability estimates (which generally run high) are around 50%. Jones et al 2019 [Link], Twin refs from Kalmbach 2016 [Link]

[758]For this chapter I am grateful for thoughtful advice from Hakhamanesh Mostafavi, Molly Przeworski, xxxx.....

[759]For histories of GWAS see: maybe Visscher papers, Clausnitzer review, Uffelman 2021

[760]By convention we often set the minor or derived allele as 1 (or sometimes the non-reference, with respect to the human reference genome). If we reverse the labeling we change the sign of the regression effect α, but not the p-value or $|\alpha|$.

[761]For height we would typically also include sex as a covariate in the regression since this has a strong effect on height.

[762]Risch and Merikangas paper. (Neil was one of my teachers when I was a graduate student at this time.)

[763]HapMap Wikipedia page: [Link]; flagship 2005 paper [Link]

[764]The most widely-used products were developed by Affymetrix and Illumina.

[765]Wellcome Trust Case Control Consortium. WTCCC followed several smaller but influential studies including: REFS

[766]WTCCC paper; Donnelly quote from [Link].

[767]At the time of writing, the largest GWAS ever conducted analyzed height by pooling 281 different cohorts to achieve a total sample size of 5.4 million individuals! They were able to collect this astonishing sample size because almost any medical study collects height, and usually weight, years of education and other data, as covariates. Yengo REF.

[768]Trubetskoy 2022

[769]Zhou 2022 [Link]. Another major cohort is the commercial 23andMe cohort, which is absolutely huge ($>$ 10 million), but relatively inaccessible to academic researchers. At the time of writing 23andMe is restructuring as a non-profit and the future of this cohort is currently unclear.

[770]One other common use of covariates is to regress out correlated traits of less interest. For example, measures of lung capacity correlate with height. But if we're studying lung function we probably don't care about SNPs that affect height. So we could include height as a covariate.

[771]The logit function is a standard approach that allows us to model binary data using a continuous variable. The logit function can convert any real number into a probability (that an individual is a case). We then assume that the actual case status of the individual is Bernoulli-distributed given this probability. For more on the logit function see [Link].

[772]One popular software package for this is called Regenie. The Methods section of that paper gives more technical details about GWAS tests: REF [Link]

[773]The term "Manhattan Plot" first appears in the GWAS literature in several papers in 2008. Prior to that it was likely in verbal use, though it's unclear who first used the term. The term appears to have been used very occasionally in other fields prior to this: for example see a 1994 book on "Hot and dense nuclear matter" (p560 [Link]).

[774]Jones et al 2019 REF.

[775]The original paper seems to have colored significant SNPs with $r^2 < 0.5$ to the lead SNP in purple, and considers these independent hits. I'm unconvinced that these are all true hits in the absence of formal analysis, which is why I don't discuss this in the main text. There are also some high-scoring SNPs colored gray; I can't see an explanation of these.

[776]I would say that people use the $5e-8$ threshold as a default for three main reasons: (i) as a matter of convention; (ii) because of LD between SNPs, the effective number of tests is less than the number of SNPs – even with whole genome sequencing, there are probably only around 1M independent tests at common SNPs; (iii) experience suggests that for well-powered GWAS studies, tests that pass this threshold do tend to replicate. That said, some authors have called for more stringent cutoffs for modern GWAS that use very high numbers of low frequency variants as these have lower LD with common SNPs. REFS.

[777]If you're familiar with multiple testing, you may also wonder whether we could use false discovery rates (FDR) here instead of Bonferroni. But in practice, FDR methods are tricky to use in GWAS because LD breaks the standard assumptions of FDR. This is because a single causal variant can be responsible for many significant SNPs. Thus, a standard FDR approach greatly over-estimates the amount of true signal. However, FDRs are very useful in related settings where we can define independent tests, such as gene-level burden tests which we describe below.

[778]The QQ plot is a helpful visual summary of how much signal is in a GWAS, but I should caution that it isn't entirely

quantitative because many of the SNPs with small p-values are non-causal SNPs in LD with a nearby causal variant.

[779]See also the Haplotype Reference Consortium: [Link]

[780] **A short history of methods for population structure confounding.** Structure confounding first became a major concern with the rise of candidate gene studies in the 1990s. At that time it was generally believed that association studies of unrelated individuals were unreliable. As an alternative, there was a great deal of interest in family-based association studies. The most famous of these was the **TDT** (transmission disequilibrium test), which used data from parent-offspring trios, where the offspring were ascertained for having a disease of interest (Spielman et al, 1993). Roughly speaking, the concept was to test whether the allele frequencies in the offspring differed from the allele frequencies of the *untransmitted* alleles carried by the parents but not passed down to the affected children. While the TDT provides absolute protection against structure confounding, it is usually far more difficult to recruit families, and practically impossible for late-onset diseases. SPIELMAN

Starting with a series of papers in 1999-2000 there was new interest in resuscitating association studies with unrelated individuals. Pritchard and Rosenberg (1999), and Devlin and Roeder (1999) introduced the idea that the signal of population structure should be spread across many/all variants, while true associations should be concentrated on a sparse set of causal variants (and LD partners). Devlin and Roeder suggested a concept called **genomic control** that estimated the amount of structure confounding and then applied a downward adjustment on all the test-statistics. Genomic control provided the first broadly applicable adjustment for structure confounding, but it is statistically inefficient because it applies the same correction to all SNPs regardless of whether they are correlated with the structure or not. PRITCHARD/ ROSENBERG, DEVLIN/ROEDER

The next major class of models aimed to estimate the relevant structure and correct for it. STRAT (Pritchard et al 2000) started by applying STRUCTURE to the case-control samples. EIGENSTRAT (Price et al 2006) proposed the use of PCA for this problem; this is now the most widely-used approach in human genetics. At around the same time, TASSEL (Yu et al 2006) introduced the use of linear mixed models (LMMs) with a relatedness matrix. LMMs are widely used in agricultural genetics and to some extent in human genetics. They are also a foundational technique for estimating SNP heritability. PRITCHARD, PRICE, YU.

[781]Lander and Schork reference

[782]Bycroft

[783]Bycroft et al identified a "white British" subset of 410,000 individuals in UKB who have "very similar ancestral backgrounds based on results of the PCA" using "a combination of self-reported ethnic background and genetic information". They suggested these individuals can serve as "a set of individuals with relatively homogeneous ancestry to reduce the risk of confounding due to differences in ancestral background" [Link]

This strategy makes analyses more robust, but has also been criticized for excluding non-European ancestries from human genetics analysis.

[784]REFs on British structure

[785]Cook et al 2020 [Link], Abdellaoui 2019 doi: 10.1038/s41562-019-0757-5; Simon Myers NG paper

[786]My writing in this section has benefitted from an excellent blog post by Iain Mathieson [Link]

[787]**The linear mixed model (LMM) approach** to correcting population structure uses a so-called **random effects** model to control for varying degrees of relatedness in the sample. This approach was originally developed for use in maize breeding experiments, where there may be different levels of relatedness in the same study, including both strong population structure and complex familial relationships. Subsequent work has focused on algorithmic speedups to make this computationally practical for biobank scale data. CITE: Yu et al (2006). Kang [Link], Zhou and Stephens 2012, others.

If two individuals are related (either in a population structure sense, or through familial relationships) then we expect them to have more similar genotypes than two random individuals from the population. In a statistical sense, this means that they have a positive genetic covariance. Similarly, we also expect those two individuals to have more similar phenotypes than two random individuals – i.e., that they have a positive phenotypic covariance.

The LMM approach connects these two ideas by assuming that the phenotypic covariance of any two individuals is simply proportional to their genome-wide genetic covariance. The model is "random" in the sense that if two individuals are closely related we don't know in advance if they will have high or low phenotype values, but we expect them to be high or low together.

Specifically, in the LMM we add a new term $Z \cdot u$:

$$Y = \mu + G_l \alpha_l + C\delta + \underbrace{Z \cdot u}_{\text{random effects}} + \epsilon \tag{4.78}$$

where Z is an $n \times n$ identity matrix (where n is the number of individuals) and u is an $n \times 1$ vector of random effects. [In experimental or agricultural applications Z is an $n \times m$ design matrix assuming m distinct lines or genotypes. In hu-

mans we can assume that all genotypes are distinct and simply use an identity matrix for Z.]

We model as u as coming from a multivariate normal distribution of the form

$$u \sim \text{MVN}(0, \lambda K) \tag{4.79}$$

where λ is a scaling factor and K is an $n \times n$ covariance matrix of SNP genotypes. We'll provide more details on this when we get to heritability estimation in the next chapter.

As a rule of thumb, I would say that PCA correction is most appropriate when we expect that the structure confounding is mainly driven by environmental effects that correlate with population structure, which is our biggest concern for human genetics. The LMM correction is very helpful for handling genome-wide polygenic effects as well as data with complex familial relationships; LMMs are widely used in agriculture, but probably somewhat less in human genetics.

Although I've presented these two models in different ways, they are mathematically similar as the PCs are eigenvalues of the genotype covariance matrix. But they make different assumptions about the relationship between genotype and phenotype. The PCA approach gives the model total flexibility to fit each of the lead PCs according to its relationship with the phenotype. For example, if the phenotype is strongly correlated with PC_3 only, the model will happily estimate a large δ for PC_3. In contrast, the LMM expects that the phenotypic variance along each PC is proportional to its eigenvalue. So in this sense, the LMM is less flexible than the PCA model, but it allows us to account for *all* levels of relatedness at once – from major structure to pairs of close relatives, which the PCA, using only lead PCs, cannot.

The PCA model is probably most appropriate if we don't expect any particular parametric relationship between environmental confounders and structure, while the LMM corresponds to our expectations under a polygenic background model.

[788]This approach was first proposed in a now-classic EIGENSTRAT paper by Alkes Price et al (2006).

[789]One practical issue is how many PCs to include. There's not very clear theoretical guidance on how to choose m; for many studies people will use 20 or 30 PCs, with the idea that this is usually more than enough and there's no real cost to using more covariates than necessary in the very large samples that are used for GWAS.

[790]The classic version of family-based tests is Spielman et al's TDT. A practical summary of family-based tests is given in the Plink software manual [Link]. The QFAM test that I describe here was originated by Fulker et al, and Abecasis et al. The early papers on this topic can be hard to follow; the clearest summary of the methods that I know of is by Guan et al 2025. REFS: Spielman et al. Fulker et al (1999, AJHG) and Abecasis et al (2000, AJHG), Guan/Young 2025

[791]Family-based tests can also tease apart complex phenomena including indirect parental effects and assortative mating (which cause bias in conventional GWAS), depending on the sampling design.

[792]I'm abusing the notation for g here by using the subscripts to index family members for a single SNP, instead of the g_l notation above where l indexes SNP. Of course we could expand this out to something like $g_{p,l}$ but with some loss of readability.

[793]The following is based on the Methods section of Guan et al (2025), with minor changes in notation. In Family GWAS (FGWAS) we model the phenotype of a child i as

$$Y_i = \alpha g_i + \alpha^*(g_p + g_m) + \epsilon_i \tag{4.80}$$

where α is the **direct genetic effect** (DGE) of the child's genotype g_i on their phenotype Y_i. Here, g_p and g_m are the paternal and maternal genotypes, and α^* is known as the non-transmitted coefficient (NTC).

The NTC estimates how much of the child's phenotype can be predicted from the alleles that are in the parents *but not transmitted to the child*. In a standard genetic model we would not expect the untransmitted alleles to affect the phenotype of the child, but we can if there is structure confounding: in that case, the untransmitted alleles are also informative about ancestry and hence phenotype. NTCs also capture so-called **indirect genetic effects** from relatives (for example, imagine alleles that cause parents to introduce more books into the house), and **assortative mating**.

The expression above assumes that we have access to parental genotypes, but it's often much easier to collect sibling pairs, especially for adult traits. In this case, we might subtract the expression for Child 2 from the expression for Child 1, giving us:

$$Y_1 - Y_2 = \alpha(g_1 - g_2) + \epsilon_1 - \epsilon_2 \tag{4.81}$$

which is equivalent to the expression in the main text, after combining the error terms. Guan et al write that "Estimates of " α ... "are free from confounding due to nonrandom mating and most gene-environment correlation, the exception being .... (indirect genetic effects) from siblings". Under random mating, the effect size in conventional GWAS is equal to the sum of the direct and indirect genetic effects: $\alpha + \alpha^*$.

[794]Mostafavi 2019, Kong/Young, others

[795]Arbel paper

[796]Mathieson and McVean 2012, Gusev blog, Wainschtain paper, extended figure 1

[797]Berg, Sohail papers; Barton commentary; more recent better examples, eg Mathieson

[798]Note that BMAL1, aka ARTNTL, is not significant in the version of the summary statistics that we plotted above, but Jones et al report it as significant in the full meta-analysis which is not available.

[799]FBXL3 binds CRY1 and CRY2 to facilitate ubiquitination and degradation. [Link]

[800]RGS16 is a repressor of GPCR signaling and helps to regulate many physiological processes; among these functions it coordinates the circadian cycling of thousands of cellular clocks in the suprachiasmatic nucleus center (SCN) via regulation of PER1 [Link]

[801]https://pubmed.ncbi.nlm.nih.gov/15301991/

[802]Ota et al 2025. REF

[803]GO [Link]. I'm simplifying the details a bit in the main text. Ota's analysis was supplemented with pathways from MSigDB [Link], which includes the heme synthesis pathway.

[804]The key functions are heme metabolism (i.e., genes involved in synthesis of the hemoglobin molecule); hematopoiesis (genes involved in development of blood cells); mitotic cell cycle (these genes affect the speed of maturation of red blood cells); and macromolecule synthesis (here the function name is a bit obtuse, but many of these genes are involved in a process called autophagy that is essential for red blood cell maturation).

[805]S-LDSC analysis of these data correctly pinpoints erythrocyte lineage precursors for hemoglobin.

[806]REFS

[807]Milind et al, and other references

[808]For discussion of gene importance, see Spence et al 2025. There's another important dimension that I have not yet introduced, which is cell type: you can potentially think of a different γ for every cell type, and that's really what we would like to learn.

[809]Stephens/Balding review; Wakefield ABM; FINEMAP, SuSiE, and PAINTOR papers.

[810]This is assuming equal prior probabilities, which is a common starting assumption. Alternatively we may sometimes give different priors according to functional evidence in which case we need to weight this also by the ratio of the priors.

[811]

[812]cite SuSiE again

[813]Suppose that SNP a is causal and SNB b is in LD with strength r^2 to a. Let $\mathrm{E}(z_a^2)$ be the expected value of the GWAS test statistic on SNP a (where z is distributed as a standard normal). Then the expected test statistic on SNP b is $r^2\mathrm{E}(z_A^2)$, assuming $r^2\mathrm{E}(z_A^2) >> 1$ (see e.g., Bulik Sullivan 2014). Then we note also that $-\log_{10}(\text{pvalue}) \approx 0.22 \cdot z^2$ (Spence et al 2025 Supp note Equation 14) from which we see that the $-\log_{10}$(pvalue) of noncausal SNPs scales with r^2 to the causal SNP. REFS: Bulik-Sullivan, Spence.

[814]cite eg Sabeti and Montgomery lab papers

[815]Review by Costanzo et al 2025 [Link]; Key papers include: Engreitz ABC papers; Gazal 2022 [Link]; Weiner 2022 [Link] Weeks 2023 [Link]; Schipper 2025 [Link]

[816]eg Gasperini 2019, Morris 2023

[817]Weeks 2023

[818]Select some suitable references here, eg Marouli Nature, Spence Nature, hs/shet papers.

[819]A gene called LDLR (LDL Receptor) is responsible for extracting LDL from the blood stream, and thus acts as a negative regulator of LDL levels. PCSK9, in turn, inhibits LDLR. Thus, lowering PCSK9 expression increases LDL Receptor activity.

[820]A class of drugs known as statins are widely used for lowering LDL levels but have toxicity issues at high doses. PCSK9 inhibitors are now an important treatment for patients with stubbornly high LDL. REFS: Raedler 2016 [Link]; Hobbs 2024 [Link]

[821]Backman for burden tests.

[822]Adult hemoglobin consists of two α-globin chains (encoded by HBA1 and HBA2) and two β-globin chains (encoded

by HBB). Fetal hemoglobin consists of two α-globin chains and two γ-globin chains, encoded by HBG1 and HBG2. After birth, a development switch turns off expression of HBG1 and HBG2 genes in favor of HBB.

[823]Platt 1994 [Link]

[824]Thein 2007 [Link]; Uda 2008 [Link]; Lettre 2008 [Link]

[825]Sankaran 2008 Science; Xu 2011 [Link] Sankaran 2013 [Link]

[826]Esrick 2020; Frangoul 2020, Frangoul 2024; all nejm

[827]link to FDA

Notes and References.

[757]Recent estimates of common SNP heritability (which probably run slightly low as they don't include low-frequency variants) are around 12-21% , while twin heritability estimates (which generally run high) are around 50%. Jones et al 2019 [Link], Twin refs from Kalmbach 2016 [Link]

[758]For this chapter I am grateful for thoughtful advice from Hakhamanesh Mostafavi, Molly Przeworski, xxxx.....

[759]For histories of GWAS see: maybe Visscher papers, Clausnitzer review, Uffelman 2021

[760]By convention we often set the minor or derived allele as 1 (or sometimes the non-reference, with respect to the human reference genome). If we reverse the labeling we change the sign of the regression effect α, but not the p-value or $|\alpha|$.

[761]For height we would typically also include sex as a covariate in the regression since this has a strong effect on height.

[762]Risch and Merikangas paper. (Neil was one of my teachers when I was a graduate student at this time.)

[763]HapMap Wikipedia page: [Link]; flagship 2005 paper [Link]

[764]The most widely-used products were developed by Affymetrix and Illumina.

[765]Wellcome Trust Case Control Consortium. WTCCC followed several smaller but influential studies including: REFS

[766]WTCCC paper; Donnelly quote from [Link].

[767]At the time of writing, the largest GWAS ever conducted analyzed height by pooling 281 different cohorts to achieve a total sample size of 5.4 million individuals! They were able to collect this astonishing sample size because almost any medical study collects height, and usually weight, years of education and other data, as covariates. Yengo REF.

[768]Trubetskoy 2022

[769]Zhou 2022 [Link]. Another major cohort is the commercial 23andMe cohort, which is absolutely huge ($>$ 10 million), but relatively inaccessible to academic researchers. At the time of writing 23andMe is restructuring as a non-profit and the future of this cohort is currently unclear.

[770]One other common use of covariates is to regress out correlated traits of less interest. For example, measures of lung capacity correlate with height. But if we're studying lung function we probably don't care about SNPs that affect height. So we could include height as a covariate.

[771]The logit function is a standard approach that allows us to model binary data using a continuous variable. The logit function can convert any real number into a probability (that an individual is a case). We then assume that the actual case status of the individual is Bernoulli-distributed given this probability. For more on the logit function see [Link].

[772]One popular software package for this is called Regenie. The Methods section of that paper gives more technical details about GWAS tests: REF [Link]

[773]The term "Manhattan Plot" first appears in the GWAS literature in several papers in 2008. Prior to that it was likely in verbal use, though it's unclear who first used the term. The term appears to have been used very occasionally in other fields prior to this: for example see a 1994 book on "Hot and dense nuclear matter" (p560 [Link]).

[774]Jones et al 2019 REF.

[775]The original paper seems to have colored significant SNPs with $r^2 < 0.5$ to the lead SNP in purple, and considers these independent hits. I'm unconvinced that these are all true hits in the absence of formal analysis, which is why I don't discuss this in the main text. There are also some high-scoring SNPs colored gray; I can't see an explanation of these.

[776]I would say that people use the $5e-8$ threshold as a default for three main reasons: (i) as a matter of convention; (ii) because of LD between SNPs, the effective number of tests is less than the number of SNPs – even with whole genome sequencing, there are probably only around 1M independent tests at common SNPs; (iii) experience suggests that for well-powered GWAS studies, tests that pass this threshold do tend to replicate. That said, some authors have called for more stringent cutoffs for modern GWAS that use very high numbers of low frequency variants as these have lower LD with common SNPs. REFS.

[777]If you're familiar with multiple testing, you may also wonder whether we could use false discovery rates (FDR) here instead of Bonferroni. But in practice, FDR methods are tricky to use in GWAS because LD breaks the standard assumptions of FDR. This is because a single causal variant can be responsible for many significant SNPs. Thus, a standard FDR

approach greatly over-estimates the amount of true signal. However, FDRs are very useful in related settings where we can define independent tests, such as gene-level burden tests which we describe below.

[778]The QQ plot is a helpful visual summary of how much signal is in a GWAS, but I should caution that it isn't entirely quantitative because many of the SNPs with small p-values are non-causal SNPs in LD with a nearby causal variant.

[779]See also the Haplotype Reference Consortium: [Link]

[780] **A short history of methods for population structure confounding.** Structure confounding first became a major concern with the rise of candidate gene studies in the 1990s. At that time it was generally believed that association studies of unrelated individuals were unreliable. As an alternative, there was a great deal of interest in family-based association studies. The most famous of these was the **TDT** (transmission disequilibrium test), which used data from parent-offspring trios, where the offspring were ascertained for having a disease of interest (Spielman et al, 1993). Roughly speaking, the concept was to test whether the allele frequencies in the offspring differed from the allele frequencies of the *untransmitted* alleles carried by the parents but not passed down to the affected children. While the TDT provides absolute protection against structure confounding, it is usually far more difficult to recruit families, and practically impossible for late-onset diseases. SPIELMAN

Starting with a series of papers in 1999-2000 there was new interest in resuscitating association studies with unrelated individuals. Pritchard and Rosenberg (1999), and Devlin and Roeder (1999) introduced the idea that the signal of population structure should be spread across many/all variants, while true associations should be concentrated on a sparse set of causal variants (and LD partners). Devlin and Roeder suggested a concept called **genomic control** that estimated the amount of structure confounding and then applied a downward adjustment on all the test-statistics. Genomic control provided the first broadly applicable adjustment for structure confounding, but it is statistically inefficient because it applies the same correction to all SNPs regardless of whether they are correlated with the structure or not. PRITCHARD/ ROSENBERG, DEVLIN/ROEDER

The next major class of models aimed to estimate the relevant structure and correct for it. STRAT (Pritchard et al 2000) started by applying STRUCTURE to the case-control samples. EIGENSTRAT (Price et al 2006) proposed the use of PCA for this problem; this is now the most widely-used approach in human genetics. At around the same time, TASSEL (Yu et al 2006) introduced the use of linear mixed models (LMMs) with a relatedness matrix. LMMs are widely used in agricultural genetics and to some extent in human genetics. They are also a foundational technique for estimating SNP heritability. PRITCHARD, PRICE, YU.

[781]Lander and Schork reference

[782]Bycroft

[783]Bycroft et al identified a "white British" subset of 410,000 individuals in UKB who have "very similar ancestral backgrounds based on results of the PCA" using "a combination of self-reported ethnic background and genetic information". They suggested these individuals can serve as "a set of individuals with relatively homogeneous ancestry to reduce the risk of confounding due to differences in ancestral background" [Link]

This strategy makes analyses more robust, but has also been criticized for excluding non-European ancestries from human genetics analysis.

[784]REFs on British structure

[785]Cook et al 2020 [Link], Abdellaoui 2019 doi: 10.1038/s41562-019-0757-5; Simon Myers NG paper

[786]My writing in this section has benefitted from an excellent blog post by Iain Mathieson [Link]

[787]**The linear mixed model (LMM) approach** to correcting population structure uses a so-called **random effects** model to control for varying degrees of relatedness in the sample. This approach was originally developed for use in maize breeding experiments, where there may be different levels of relatedness in the same study, including both strong population structure and complex familial relationships. Subsequent work has focused on algorithmic speedups to make this computationally practical for biobank scale data. CITE: Yu et al (2006). Kang [Link], Zhou and Stephens 2012, others.

If two individuals are related (either in a population structure sense, or through familial relationships) then we expect them to have more similar genotypes than two random individuals from the population. In a statistical sense, this means that they have a positive genetic covariance. Similarly, we also expect those two individuals to have more similar phenotypes than two random individuals – i.e., that they have a positive phenotypic covariance.

The LMM approach connects these two ideas by assuming that the phenotypic covariance of any two individuals is simply proportional to their genome-wide genetic covariance. The model is "random" in the sense that if two individuals are closely related we don't know in advance if they will have high or low phenotype values, but we expect them

to be high or low together.

Specifically, in the LMM we add a new term $Z \cdot u$:

$$Y = \mu + G_l\alpha_l + C\delta + \underbrace{Z \cdot u}_{\text{random effects}} + \epsilon \tag{4.82}$$

where Z is an $n \times n$ identity matrix (where n is the number of individuals) and u is an $n \times 1$ vector of random effects. [In experimental or agricultural applications Z is an $n \times m$ design matrix assuming m distinct lines or genotypes. In humans we can assume that all genotypes are distinct and simply use an identity matrix for Z.]

We model as u as coming from a multivariate normal distribution of the form

$$u \sim \text{MVN}(0, \lambda K) \tag{4.83}$$

where λ is a scaling factor and K is an $n \times n$ covariance matrix of SNP genotypes. We'll provide more details on this when we get to heritability estimation in the next chapter.

As a rule of thumb, I would say that PCA correction is most appropriate when we expect that the structure confounding is mainly driven by environmental effects that correlate with population structure, which is our biggest concern for human genetics. The LMM correction is very helpful for handling genome-wide polygenic effects as well as data with complex familial relationships; LMMs are widely used in agriculture, but probably somewhat less in human genetics.

Although I've presented these two models in different ways, they are mathematically similar as the PCs are eigenvalues of the genotype covariance matrix. But they make different assumptions about the relationship between genotype and phenotype. The PCA approach gives the model total flexibility to fit each of the lead PCs according to its relationship with the phenotype. For example, if the phenotype is strongly correlated with PC_3 only, the model will happily estimate a large δ for PC_3. In contrast, the LMM expects that the phenotypic variance along each PC is proportional to its eigenvalue. So in this sense, the LMM is less flexible than the PCA model, but it allows us to account for *all* levels of relatedness at once – from major structure to pairs of close relatives, which the PCA, using only lead PCs, cannot.

The PCA model is probably most appropriate if we don't expect any particular parametric relationship between environmental confounders and structure, while the LMM corresponds to our expectations under a polygenic background model.

[788]This approach was first proposed in a now-classic EIGENSTRAT paper by Alkes Price et al (2006).

[789]One practical issue is how many PCs to include. There's not very clear theoretical guidance on how to choose m; for many studies people will use 20 or 30 PCs, with the idea that this is usually more than enough and there's no real cost to using more covariates than necessary in the very large samples that are used for GWAS.

[790]The classic version of family-based tests is Spielman et al's TDT. A practical summary of family-based tests is given in the Plink software manual [Link]. The QFAM test that I describe here was originated by Fulker et al, and Abecasis et al. The early papers on this topic can be hard to follow; the clearest summary of the methods that I know of is by Guan et al 2025. REFS: Spielman et al. Fulker et al (1999, AJHG) and Abecasis et al (2000, AJHG), Guan/Young 2025

[791]Family-based tests can also tease apart complex phenomena including indirect parental effects and assortative mating (which cause bias in conventional GWAS), depending on the sampling design.

[792]I'm abusing the notation for g here by using the subscripts to index family members for a single SNP, instead of the g_l notation above where l indexes SNP. Of course we could expand this out to something like $g_{p,l}$ but with some loss of readability.

[793]The following is based on the Methods section of Guan et al (2025), with minor changes in notation. In Family GWAS (FGWAS) we model the phenotype of a child i as

$$Y_i = \alpha g_i + \alpha^*(g_p + g_m) + \epsilon_i \tag{4.84}$$

where α is the **direct genetic effect** (DGE) of the child's genotype g_i on their phenotype Y_i. Here, g_p and g_m are the paternal and maternal genotypes, and α^* is known as the non-transmitted coefficient (NTC).

The NTC estimates how much of the child's phenotype can be predicted from the alleles that are in the parents *but not transmitted to the child*. In a standard genetic model we would not expect the untransmitted alleles to affect the phenotype of the child, but we can if there is structure confounding: in that case, the untransmitted alleles are also informative about ancestry and hence phenotype. NTCs also capture so-called **indirect genetic effects** from relatives (for example, imagine alleles that cause parents to introduce more books into the house), and **assortative mating**.

The expression above assumes that we have access to parental genotypes, but it's often much easier to collect sibling pairs, especially for adult traits. In this case, we might subtract the expression for Child 2 from the expression for

Child 1, giving us:

$$Y_1 - Y_2 = \alpha(g_1 - g_2) + \epsilon_1 - \epsilon_2 \tag{4.85}$$

which is equivalent to the expression in the main text, after combining the error terms. Guan et al write that "Estimates of " α ... "are free from confounding due to nonrandom mating and most gene-environment correlation, the exception being .... (indirect genetic effects) from siblings". Under random mating, the effect size in conventional GWAS is equal to the sum of the direct and indirect genetic effects: $\alpha + \alpha^*$.

[794]Mostafavi 2019, Kong/Young, others

[795]Arbel paper

[796]Mathieson and McVean 2012, Gusev blog, Wainschtain paper, extended figure 1

[797]Berg, Sohail papers; Barton commentary; more recent better examples, eg Mathieson

[798]Note that BMAL1, aka ARTNTL, is not significant in the version of the summary statistics that we plotted above, but Jones et al report it as significant in the full meta-analysis which is not available.

[799]FBXL3 binds CRY1 and CRY2 to facilitate ubiquitination and degradation. [Link]

[800]RGS16 is a repressor of GPCR signaling and helps to regulate many physiological processes; among these functions it coordinates the circadian cycling of thousands of cellular clocks in the suprachiasmatic nucleus center (SCN) via regulation of PER1 [Link]

[801]https://pubmed.ncbi.nlm.nih.gov/15301991/

[802]Ota et al 2025. REF

[803]GO [Link]. I'm simplifying the details a bit in the main text. Ota's analysis was supplemented with pathways from MSigDB [Link], which includes the heme synthesis pathway.

[804]The key functions are heme metabolism (i.e., genes involved in synthesis of the hemoglobin molecule); hematopoiesis (genes involved in development of blood cells); mitotic cell cycle (these genes affect the speed of maturation of red blood cells); and macromolecule synthesis (here the function name is a bit obtuse, but many of these genes are involved in a process called autophagy that is essential for red blood cell maturation).

[805]S-LDSC analysis of these data correctly pinpoints erythrocyte lineage precursors for hemoglobin.

[806]REFS

[807]Milind et al, and other references

[808]For discussion of gene importance, see Spence et al 2025. There's another important dimension that I have not yet introduced, which is cell type: you can potentially think of a different γ for every cell type, and that's really what we would like to learn.

[809]Stephens/Balding review; Wakefield ABM; FINEMAP, SuSiE, and PAINTOR papers.

[810]This is assuming equal prior probabilities, which is a common starting assumption. Alternatively we may sometimes give different priors according to functional evidence in which case we need to weight this also by the ratio of the priors.

[811]

[812]cite SuSiE again

[813]Suppose that SNP a is causal and SNB b is in LD with strength r^2 to a. Let $\mathrm{E}(z_a^2)$ be the expected value of the GWAS test statistic on SNP a (where z is distributed as a standard normal). Then the expected test statistic on SNP b is $r^2\mathrm{E}(z_A^2)$, assuming $r^2\mathrm{E}(z_A^2) >> 1$ (see e.g., Bulik Sullivan 2014). Then we note also that $-\log_{10}(\text{pvalue}) \approx 0.22 \cdot z^2$ (Spence et al 2025 Supp note Equation 14) from which we see that the $-\log_{10}$(pvalue) of noncausal SNPs scales with r^2 to the causal SNP. REFS: Bulik-Sullivan, Spence.

[814]cite eg Sabeti and Montgomery lab papers

[815]Review by Costanzo et al 2025 [Link]; Key papers include: Engreitz ABC papers; Gazal 2022 [Link]; Weiner 2022 [Link] Weeks 2023 [Link]; Schipper 2025 [Link]

[816]eg Gasperini 2019, Morris 2023

[817]Weeks 2023

[818]Select some suitable references here, eg Marouli Nature, Spence Nature, hs/shet papers.

[819]A gene called LDLR (LDL Receptor) is responsible for extracting LDL from the blood stream, and thus acts as a negative regulator of LDL levels. PCSK9, in turn, inhibits LDLR. Thus, lowering PCSK9 expression increases LDL Receptor activity.

[820]A class of drugs known as statins are widely used for lowering LDL levels but have toxicity issues at high doses. PCSK9 inhibitors are now an important treatment for patients with stubbornly high LDL. REFS: Raedler 2016 [Link]; Hobbs 2024 [Link]

[821]Backman for burden tests.

[822]Adult hemoglobin consists of two α-globin chains (encoded by HBA1 and HBA2) and two β-globin chains (encoded by HBB). Fetal hemoglobin consists of two α-globin chains and two γ-globin chains, encoded by HBG1 and HBG2. After birth, a development switch turns off expression of HBG1 and HBG2 genes in favor of HBB.

[823]Platt 1994 [Link]

[824]Thein 2007 [Link]; Uda 2008 [Link]; Lettre 2008 [Link]

[825]Sankaran 2008 Science; Xu 2011 [Link] Sankaran 2013 [Link]

[826]Esrick 2020; Frangoul 2020, Frangoul 2024; all nejm

[827]link to FDA