4.1 A genome owner's guide to phenotypic variation (the expansion pack)

Here we set the stage for Part 4 of the book, on the genetics of phenotypic variation.

We'll highlight the main themes that run through this section of the book, connecting and distinguishing three main types of genetic traits: monogenic/Mendelian traits, cancer, and complex traits [598]. [a]

This table gives a simplified overview of the different types of traits:

	Monogenic traits	*Cancer*	*Complex traits*
Number of genes	~ 1	oligogenic	polygenic
Mechanisms	coding	coding	regulatory
Types of variants	rare germline	somatic mutations	common SNPs
Gene Discovery	linkage/sequencing	sequencing	GWAS
Selection	purifying	positive	stabilizing

Table 4.1: Simplified overview of three types of traits. *This table emphasizes general patterns (though as we shall see,* ***the details are a bit more complicated****). These terms will become familiar to you as we go along!*

What is a phenotype? The terms **phenotype** and **trait** refer to any feature that we can measure in a human or other organism. The traits that people have studied include measurements at all scales from molecular measures such as gene expression or cholesterol; cellular measures such as red blood cell counts; whole-body traits such as height, weight, or hair color; to diseases such as cystic fibrosis or hemophilia, diabetes, schizophrenia or cancer; and even behavioral traits including tea drinking, smoking, or years-of-education.

Many phenotypes are referred to as **quantitative traits**, meaning that they are drawn from a continuous distribution, including traits such as cholesterol or height. Other traits are **binary**, meaning that we consider only two possible categories – this includes most diseases where we might want to study presence/absence of a particular diagnosis). A handful of traits involve more than two **discrete categories**, such as self-reported eye color or number of children.

For most analyses the mathematical modeling is much easier for quantitative traits so we'll usually start by understanding quantitative traits. The underlying principles are similar for binary traits and diseases but the math is more complicated.

Genotypes and Phenotypes. Given that any two people's genomes differ at millions of positions [b], the next major questions consider how – and whether – this variation affects phenotypes. Specifically, we could ask two related questions:

- *How does genetic variation impact the information encoded in genomes?*

[a] *Rough draft version of chapter.*

[b] *See Chapter 1.3 for a review on types of genetic variation and how these can affect the encoding of genome information.*

- *How does genetic variation affect human traits and diseases?*

And when thinking about genetics, we should also ask:
- *What is the role of environment or other non-genetic effects on phenotype?*

These are among the central questions in human genetics!

As we discussed above, probably less than 10% of the genome encodes specific, sequence-dependent functional information. So most single nucleotide changes have no discernible effect on phenotype at all.

Among those that do affect the encoded information, recall that **genetic variation can affect either protein sequences, or gene regulation**. Since the encoding of proteins is very precise, just a single nucleotide change can alter the amino acid sequence of a protein, or even add a stop mutation with potentially disastrous consequences. Single nucleotide changes can also change gene regulation, but these are usually quantitative changes: dialing expression up a bit or down a bit [c].

Just as we can classify mutations as affecting protein sequences or regulation, we can also classify most inherited traits as being either **monogenic** or **complex**. **Cancer** also has a genetic basis, but instead is primarily due to *somatic* mutations – i.e., mutations that occur during one's lifetime within tissues of the body. We now describe the main features of each:

Monogenic/Mendelian traits. The birth of modern genetics can be traced to Gregor Mendel's work on the inheritance of phenotypes in peas. Mendel made the lucky choice of picking traits that are controlled by single-gene inherited variants; from these, he was able to describe the basic rules of inheritance, including the concepts of recessive and dominant alleles.

Some human traits – mainly disease traits – follow similar patterns of inheritance. Such traits are referred to as **Mendelian**. For example, the pedigree at the right shows inheritance of a severe neurological disease called Huntington's disease within a large family [599]. Individuals with the disease are marked in black.

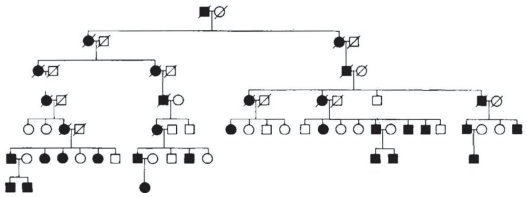

Figure 4.1: ***Inheritance of an STR-based disease.*** *This pedigree is from a classic 1983 paper that ultimately led to the gene and mechanism of Huntington's disease. Affected individuals are in black: transmission patterns reflect dominant inheritance of the expanded STR.* *Modified Fig. 2 of James Gusella et al (1983).* [Link]

Huntington's disease shows a **dominant** transmission pattern, in which individuals with a single copy of a disease allele (i.e., heterozygotes) inherit the disease. According to the rules of Mendelian transmission, an affected (i.e., heterozygous) individual has a 50% chance of passing the allele to each child. As you can infer from the pedigree, a single individual in the first generation carried the disease allele; this allele was then transmitted through each subsequent generation.

Other diseases, such as cystic fibrosis or Tay Sachs disease, show **recessive** transmission, in which heterozygotes are asymptomatic, and are described as *carriers*. Affected individuals are people who have inherited disease alleles in both copies of the relevant gene – i.e., their parents were heterozygous carriers, and they themselves are homozygous.

The hallmark of Mendelian traits is that *they are caused by alleles that encode a major disruption to the function of a particular gene, with highly predictable effects*: namely that in either heterozygous or homozygous form

[c] *If this is unfamiliar to you, I encourage you to revisit Chapters 1.2 and 1.3.*

(depending on whether the disease is dominant or recessive) these alleles are sufficient to lead to a specific disease. We can also use the term **monogenic** to refer to traits for which variants in a single gene are sufficient to cause disease.

While there are now several thousand Mendelian traits that have been characterized (mostly disease-causing, and mostly very rare) it's also becoming clear that *many genetic diseases are monogenic while not showing obvious patterns of Mendelian inheritance*. Many severe developmental disorders in children are caused by major-effect mutations in critical genes [600] [601]. These diseases are often so severe that the affected children are unlikely to reproduce, and so the diseases don't show clear inheritance patterns [602].

Traditionally, Mendelian diseases such as Huntingdon's were studied by looking at transmission of genetic variation within large pedigrees. More recently, modern genome sequencing has opened the door to finding mutations in diseases that are monogenic but without clear Mendelian inheritance.

Complex traits. While monogenic diseases are an important category of disease burden, especially in children, most of the ways that we differ from one another are genetically complex.

Complex traits include phenotypes as diverse as cholesterol or glucose levels; height or weight; personalities, abilities, and other behavioral and educational traits; and most diseases (such as diabetes, heart disease, rheumatoid arthritis, depression, and many others). These traits are influenced by huge numbers – many thousands – of variants across the genome, as well as important contributions from the environment [d].

Prior to modern genomic approaches (starting around 2005), most of what we knew about complex traits came from studying patterns of inheritance within families. Starting in around 2005, a new approach called a **Genome-Wide Association Study (GWAS)** has made it possible to study the genetic basis of complex traits.

For example, GWAS has so far identified more than 200 regions of the genome that affect the probability that someone has schizophrenia, and shows that the genetic contribution comes from more than 10,000 variants overall [603]. These features of schizophrenia are typical of complex traits more broadly: complex traits tend to be hugely **polygenic** – meaning that they are influenced by many variants; individual variants generally make only tiny contributions to overall risk; and environmental or other non-genetic factors also make important contributions.

In sharp contrast to monogenic traits (and also to cancer), most of the genetic effects on complex traits act through **noncoding effects on gene regulation**, instead of by changing protein sequences.

Cancer. Cancer represents a third, distinct, category of genetic trait. Cancer refers to a class of diseases caused by uncontrolled cell growth. While the two disease categories above are caused by inherited variation that

[d] *Most human traits are complex: influenced by thousands of SNPs, each with tiny effects, plus important environmental contributions.*

is (usually) present in the embryo at the time of fertilization, cancer is mainly caused by **somatic mutations**: i.e., mutations that arise within cells of the body, during the course of a patient's lifetime. As a rule, cancer mutations tend to impact genes that control cell proliferation – for example, cancer mutations often increase cell division rates, or eliminate controls on excessive growth. Some cancer mutations act by knocking out genes involved in DNA repair or genome stability, thereby enabling additional mutations that can drive the transition to cancer.

Like monogenic diseases, cancer mutations are most often large-effect mutations that affect protein coding sequences. However, unlike monogenic diseases, full-blown cancers usually require combinations of multiple mutations, together changing the function of a handful of key driver genes (this is referred to as **oligogenic**).

(For some types of cancer, such as breast cancer, inherited genetics also contribute to a patient's overall risk profile, but these still generally involve somatic mutations that drive the transition to cancer [604].)

The role of natural selection in shaping trait genetics. One major theme of this book is to understand how evolutionary pressures shape genetic variation, phenotypic variation, and disease risk.

Aside from a few important exceptions [605], **monogenic disease variants** are subject to strong negative selection that keep them very rare in the population. These processes can be understood with a model called mutation-selection balance, which we will meet in Chapter 4.2.

Complex trait variants are also subject to a form of negative selection known as *stabilizing selection*, but the strength of selection is usually much weaker than for monogenic traits [606]. This makes drift more important, and most of the genetic contribution to complex traits is due to alleles that have drifted up to high frequencies. We'll see more about this in Chapter 4.8.

And lastly, we shall see that **cancer** is driven by **positive selection**. Somatic mutations that increase cell proliferation will tend to increase in numbers within a tissue, as the cells that carry them grow in numbers. This is much like the increase of favored genotypes in Darwinian adaptation, even though in the long run this may lead to death of the organism. We'll discuss this in Chapter 4.3.

Applications of gene discovery. One major goal in human genetics is to determine the genetic basis of different traits and diseases. As we shall see, one fundamental feature of gene mapping is that it can **establish causal links from genes to human phenotypes** – this is something that is virtually impossible by other methods [607].

More specifically, applications of phenotype studies include:
- Improving basic understanding of molecular mechanisms through gene discovery; what are the key mechanisms of disease?
- Identification of genes that can serve as therapeutic targets for drugs or

gene editing.
- Disease diagnosis – especially for developmental disorders;
- Patient stratification and care – especially for targeted cancer treatment, and potentially in other conditions;
- Prediction of who is at risk for any given disease, to improve screening or risk mitigation.

Genetics plays a significant role in who we are, our strengths, limitations, and disease risks – and understanding this is the central theme of this section of the book. But as we discuss next, there are also dangers in *overstating* the degree to which genetics determines our fates.

Genetic determinism, heritability, and environment. The 1997 movie GATTACA imagined a dystopian future in which all of our talents, our limitations, and the diseases we will suffer from, are already written in our genomes at birth. The hero of the movie, played by Ethan Hawke, has been assigned to work on a cleaning crew because of his supposedly poor genetic material. He tries to escape the bonds of genetic discrimination by uploading another man's genome sequence into the computers in place of his own. The movie illustrates an extreme form of **genetic determinism**: the ideology that our strengths and weaknesses are entirely determined by our DNA [608].

Figure 4.2: *Poster for the 1997 movie GATTACA about a genetic determinist dystopia.*
Fair Use. [Link]

But setting aside the exaggerated world of GATTACA, to what extent does genetics determine who we are? This is an important, and sometimes politically-charged, topic. The issues are complicated, but I'll make some brief comments here, before we revisit the topic in Chapter 4.4.

The table below illustrates this for schizophrenia, which is a serious psychiatric condition involving psychosis. The data show the prevalence (i.e., the frequency) of schizophrenia in different types of relatives of an affected individual [609].

Relationship	Prevalence	Recurrence Risk Ratio
MZ twin	44%	52×
DZ twin	12%	14
Sibling	7.3%	8.6
Half-Sib	3.0%	3.5
Cousin	1.5%	1.8
Random	.85%	1

Table 4.2: Schizophrenia rates in relatives of an affected individual. *Prevalence shows the frequency of schizophrenia in different types of relatives. The Recurrence Risk Ratio is the prevalence in a specific type of relative, divided by the population prevalence of* 0.85%.

The table illustrates several important points that are typical of complex traits:
- If someone has an MZ (monozygous, or "identical") twin with schizophrenia, then they themselves have a 44% chance of also suffering from schizophrenia. (MZ twins come from the same fertilized egg, and have virtually identical genomes.) This is about 52-fold higher than the population prevalence of about 0.85%. While a prevalence of 44% is high com-

pared to the general population, the fact that the prevalence is not 100% shows that schizophrenia is not determined by genetics alone (environmental and random factors matter too).
• The rate for DZ (dizygous) twins is much lower than for MZ twins, at about 12%. (Dizygous twins come from different fertilized eggs and have the same genetic relationship as ordinary siblings.) Given that both MZ and DZ twin pairs are born at the same time, we may expect that they are equally likely to share key environmental factors. Hence, the lower recurrence ratio for DZ twins is interpreted as reflecting the lower amount of genetic sharing between DZ twins compared to MZ twins [610].
• Risk decreases steadily with decreasing levels of relatedness, as would be expected for a genetic trait. But we do have to be careful here, because sharing of environmental factors also probably decreases with decreasing levels of relatedness, so this is *consistent* with a role for genetic factors, but we need genetic data to be completely confident. (In fact, we now know that schizophrenia *does* have a strong genetic component [611].)

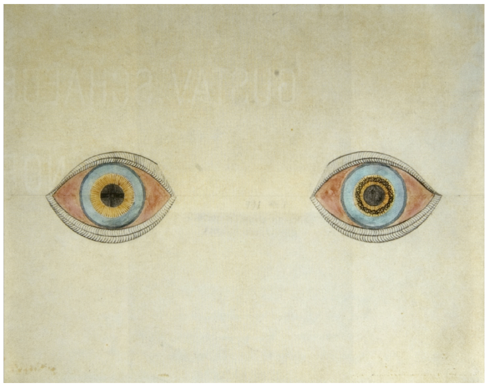

Figure 4.3: ***My Eyes at the Moment of the Apparitions.*** *This haunting image illustrates psychotic episodes experienced by the artist, who suffered from schizophrenia.*

August Natterer, 1913. [Link] Public Domain.

What does it mean for something to be *genetic*? I find the metaphor of a *genome as software* useful when we're thinking about essential biological processes that are exquisitely determined, such as in development: for example, how a single fertilized egg cell develops into a human or a chimpanzee or a dog or a melon. At this scale, genome is destiny.

But when we look at the *differences* between individuals within a species, things become much murkier. Within humans, everyone's genomes are relatively similar, at least compared to the magnitude of differences between us, chimps, dogs, and melons [612]. Phenotypic differences between individuals are indeed influenced by genetics, but they are *more* influenced by environmental factors and random chance.

For example, during the COVID-19 pandemic, our dog Jack sat through an entire class of my online lectures about genetics, and I think he still doesn't understand very much. That was entirely predictable from his genome – he's a dog. At the same time, my human students learned a lot. The fact that they now know more genetics than other Stanford students is because of their environment (i.e., that they have taken a genetics class).

So should we say that ability to learn about genetics is genetically encoded, or environmental? In some sense the answer is both, depending entirely on context.

Ok, that's a silly example. But if you think of all the ways that humans vary – for example in height, weight, disease risk, running speed, and behavior – nearly all of these traits are influenced to some extent by both genetics and environment [613]. While all of these traits have a genetic component, they are also strongly affected by environment and the whims of fortune – for example nutritional intake as a child or as an adult; access to sporting or educational opportunities, household income, privilege or deprivation.

Nature versus Nurture. Can we quantify the extent to which a trait is determined by genetics ("nature") versus environment ("nurture")? Later in the book we'll discuss how the role of genetics within a population is summarized using a measure called **heritability**, which lies between 0 and 1 and measures the fraction of phenotypic variance that is due to genetics [614].

While genetics contributes to most traits, the magnitude of heritability varies widely. For example, studies show that both adult height and educational attainment (measured by years-of-education completed [615]) have a detectable genetic component, but *the importance of genetics is very large for height, and rather modest for educational attainment* [616].

Of course our goal in this book is to understand the role of *genetics* in human phenotypes, but we should be careful not to overstate the role of genetics in behavior, which are strongly shaped by our families and environment. Moreover, it is fundamentally difficult to untangle the precise roles of genetics and environment, and most especially for behavioral traits like education, because family context and experience is inherited alongside genetics. In the next chapters we'll talk about what we do (and don't!) know on these topics.

In the upcoming chapters we'll take a deeper dive into the genetics of human traits.

Notes and References.

[598] Thanks as always to the generosity of people who commented on earlier drafts of this chapter: Hakhamanesh Mostafavi, Molly Przeworski, and Julien Sage. As always, any errors are my own.

[599] Gusella et al 1983 found a genetic marker that localized the disease gene to Chromosome xxx. A later paper, xxx identified the causal gene. REFS

[600] eg DDD reference

[601] This may also happen for rare recessive diseases. Affected individuals may appear sporadically in sibships, but in the absence of inbreeding it's unlikely that there would be other affected relatives

[602] It's also becoming clear that some traits that are traditionally considered monogenic are also influenced by the polygenic background, thus blurring the boundaries between monogenic and complex traits: eg DDD, LDL examples.

[603] Trubetskoy 2022 `https://rdcu.be/d34Eg`

[604] There can also be a significant component of inherited risk, acting either through major-effect (monogenic) mutations (eg retinoblastoma or BRCA1/BRCA2) or as a polygenic (complex) background risk. Prostate?

[605] We encountered balancing selection for sickle cell anemia in Chapter 2.6. Mutations at the cystic fibrosis locus CFTR were probably subject to balancing selection due to protection from cholera.

[606] Selection on complex traits can be a little counter-intuitive. If we think of selection on a disease trait like stroke or schizophrenia, say, it's natural to assume that selection will act against variants that increase risk. But a better way to think about this is that these variants are usually affecting some underlying cellular process, e.g., in endothelial cells or neurons, respectively. The main target of selection is likely to ensure optimal functioning of these cell types, and the disease outcomes are likely a pleiotropic outcome of this. Selection is usually modeled as stabilizing selection.

[607] In principle, clinical trials also establish causality, but these are very expensive and cannot be used for discovery purposes.

[608] For an excellent in-depth consideration of the genetic issues raised by GATTACA see `https://doi.org/10.1093/genetics/iyac142`

[609] Data from Risch 1990. These estimates are similar to more recent estimates, e.g., from Lichtenstein 2006. Note that the original estimates did not report uncertainty, but the available sample sizes for twins are not large, and the estimates must be quite noisy.

[610] Another interesting feature of the data is that ordinary sibs have a lower recurrence risk than DZ sibs. This likely reflects a greater shared environment between twins, including their shared uterine environment

[611] reference genetic studies of scz

[612] The divergence between a human and chimpanzee genome is about 15-fold higher than the divergence between two haploid human genomes. Divergence to these other species is much greater.

[613] Visscher NG paper on heritability

[614] Very briefly, heritability is defined as the fraction of phenotypic variance in a population that is explained by genetic variance. This definition means that heritability depends on what 'population' we are looking at and what other sources of phenotypic variance are present. For example, if we compare a population where all children have adequate nutrition, to one where half the children do not, we can expect that in the latter population there would be greater variance in height, and lower heritability. So while heritability can give us some idea about the importance of genetics in a particular time and place, it's not a fundamental property of a trait, and it should be interpreted carefully.

[615] I have to confess that when I first heard about people doing genome-wide association studies for educational attainment, I thought this was satirical, and many students have a similar reaction. In part, this phenotype is used because it is easy to measure in very large sample sizes. However, GWAS of educational attainment does detect real biological signals, though how much of this is real, and exactly what these measure is still controversial.

[616] Heritability is around 0.6–0.8 for height and probably around 0.2–0.3 for educational attainment. Measurement and interpretation of heritability is especially complicated for education-related traits for technical reasons (assortative mating, indirect genetic effects, and inherited environment).